FSBM
Faculté des sciences ben m'sick
Université Hassan II de Casalanca

Data Science & Big Data
Master DSBD

MATHÉMATIQUES & INFORMATIQUE

Projet de Fin d'Année

Utilisation du Process Mining pour l'Analyse des Sinistres d'Assurance et Détection des Violations & Fraudes

Master Data Science & Big data

Réalisé Par :

Abdelhakim El Ghayoubi abdo.elghayoubi@gmail.com

Examiné par :

Pr. Mohammed Ait Daoud **Encadrant**

Année Universitaire : 2023/2024

Présenté le : XX/XX/2024

Dédicace

A nos très chers parents En témoignage de notre amour et de notre gratitude pour les sacrifices consentis à notre égard A nos frères et sœurs Pour leur soutien moral et leurs conseils précieux et leur encouragement à nos chères ami(e)s Pour leurs aides et supports dans les moments difficiles. A toute ma famille Pour avoir attendu avec patience les fruits de leur bonne éducation A tous ceux qui de près ou de loin nous sont chers Nous dédions ce mémoire.

Remerciement

Je tiens à exprimer ma profonde gratitude à Monsieur Mohammed Ait Daoud pour son rôle d'encadrant tout au long de ce projet. Son soutien constant, ses conseils avisés et son expertise ont été essentiels pour mener à bien ce travail.

En tant qu'étudiant dans le programme de Master en Data Science & Big Data, j'ai particulièrement apprécié l'accompagnement précieux qu'il a fourni, me permettant de surmonter les défis rencontrés et d'atteindre les objectifs fixés. Sa confiance en mes capacités a été une source d'inspiration et de motivation tout au long du projet.

Je le remercie sincèrement pour son engagement et son dévouement, qui ont grandement contribué à la réussite de ce projet.

Résume

L'objectif principal de ce travail de **PFA** est de mettre en œuvre un modèle qui prédit les anomalies en s'appuyant sur l'analyse basée sur les techniques de **process mining** pour l'analyse des sinistres d'assurance et la **détection des fraudes et des violations**. Ce projet propose une approche innovante pour améliorer l'efficacité opérationnelle et réduire les pertes financières dans le secteur de l'assurance. Grâce à **l'exploitation des journaux d'événements** et à l'application d'algorithmes avancés, le process mining permet de modéliser et d'analyser les processus de réclamation en temps réel, ce qui fournit des indicateurs clés pour identifier les anomalies et détecter ainsi les comportements potentiellement frauduleux.

Sommaire

Liste des Figures :

Liste des Tableaux :

Introduction Générale

Le présent rapport décrit le travail réalisé dans le cadre du projet de fin d'année à la Faculté des Sciences Ben M'sick, en Licence des Sciences Mathématiques et Informatique, parcours bases de données. Ce projet se concentre sur l'application du process mining pour l'analyse des sinistres d'assurance, avec pour objectif principal la détection des anomalies et des fraudes potentielles.

Dans le secteur de l'assurance, la fraude représente un défi majeur qui entraîne des pertes financières considérables et affaiblit la confiance des assurés. Grâce à l'augmentation du volume de données numériques et à l'évolution des technologies d'analyse, le process mining s'affirme comme une solution prometteuse. Cette technologie permet d'extraire des informations pertinentes à partir des journaux d'événements, facilitant ainsi l'identification des inefficacités et des comportements suspects dans les processus métiers.

L'étude commence par la préparation des données relatives aux réclamations d'assurance. Ces données sont ensuite utilisées pour modéliser les processus de réclamation, identifiant les déviations par rapport aux procédures standard. Une transformation et l'étiquetage des données est ensuite effectuée pour extraire des caractéristiques significatives qui permettent d'améliorer la précision des analyses.

Finalement, des techniques d'apprentissage automatique sont appliquées pour prédire les anomalies et détecter les réclamations potentiellement frauduleuses. Cette approche montre le potentiel du process mining en tant qu'outil stratégique pour renforcer l'efficacité opérationnelle dans le secteur de l'assurance. Le projet ouvre également des perspectives pour intégrer des systèmes de détection de fraude plus avancés, contribuant ainsi à une meilleure gestion des sinistres et à une réduction des pertes financières.

Chapitre 1 : Présentation & Préparation de Projet

Introduction

Ce chapitre vise à situer notre projet dans son contexte global en présentant son périmètre. Tout d'abord, on commence par une présentation du contexte du projet, en mettant en évidence la problématique générale à résoudre et les solutions de développement envisagées. Ensuite, nous détaillerons les objectifs du projet, en veillant à mentionner son périmètre d'action.

1. Cadre d'accueil :

1.1. Présentation de la faculté des sciences Ben M'sick:

La Faculté des Sciences Ben M'sick, en continuelle mutation dans la Ville de Casablanca, est rattachée à l'Université Hassan II qui regroupe 18 établissements universitaires. Depuis sa création en 1984, elle n'a cessé de contribuer à la diversification des spécialités et des circuits de formation et à la consolidation de la recherche scientifique au profit de ses étudiants, ainsi qu'au développement socioéconomique de Casablanca.

Dès son ouverture, la Faculté des Sciences Ben M'sik a accordé un intérêt particulier au développement de la recherche scientifique parallèlement à sa mission d'enseignement et de formation.

En effet, elle englobe actuellement 6 parcours de licences fondamentales et 18

Masters. La recherche scientifique connait aujourd'hui une progression importante. 23 structures de recherche (2 centres de recherche, un observatoire « ordipu », une plateforme PINTECH, 19 laboratoires) ont vu le jour dans divers domaines tels que : les sciences et techniques de l'ingénieur, les matériaux, la biotechnologie, la géoscience

Depuis 2003, la faculté des sciences Ben M'sik dispense une formation modulaire et semestrielle dans le cadre de la réforme pédagogique de l'enseignement supérieur conformément au système LMD. Dans le cadre de la structuration de la recherche, que l'Université Hassan II – Casablanca, la Faculté des Sciences Ben M'sick a procédé à une nouvelle organisation et restructuration de ses équipes et laboratoires de recherche. C'est ainsi que la recherche à la faculté des sciences s'est organisée en 23 laboratoires et 2 équipes de recherches.

En 2008 suite à la réorganisation du cycle doctorat. La Faculté des Sciences Ben M'sick a mis en place le Centre d'Etude Doctoral (CED) : « Sciences et applications ». Ce centre est adossé à l'ensemble des structures de recherches accréditées par l'université.

La filière SMI donne une information de base à prédominance informatique avec un enseignement des modules de mathématiques et de physique, Les deux premiers semestres constituent un tronc commun avec la filière "SMA". Durant la deuxième et troisième année, on commence à introduire les notions de base (l'algorithmique, programmations, structure des données, base de données, réseau informatique, système d'exploitation …), sanctionné par un diplôme "Licence" permettant aux étudiants l'insertion dans la vie professionnelle ou la poursuite des études supérieures en Master.

Afin de renforcer la coopération, la Faculté des Sciences Ben M'sick a noué des relations de coopération internationale à travers la signature de conventions et accords avec des universités, des institutions universitaires, des centres et des laboratoires de recherche au Maroc et à l'étranger. Elle est aussi engagée dans divers projets de coopération universitaire internationale.

1.2. Organigramme :

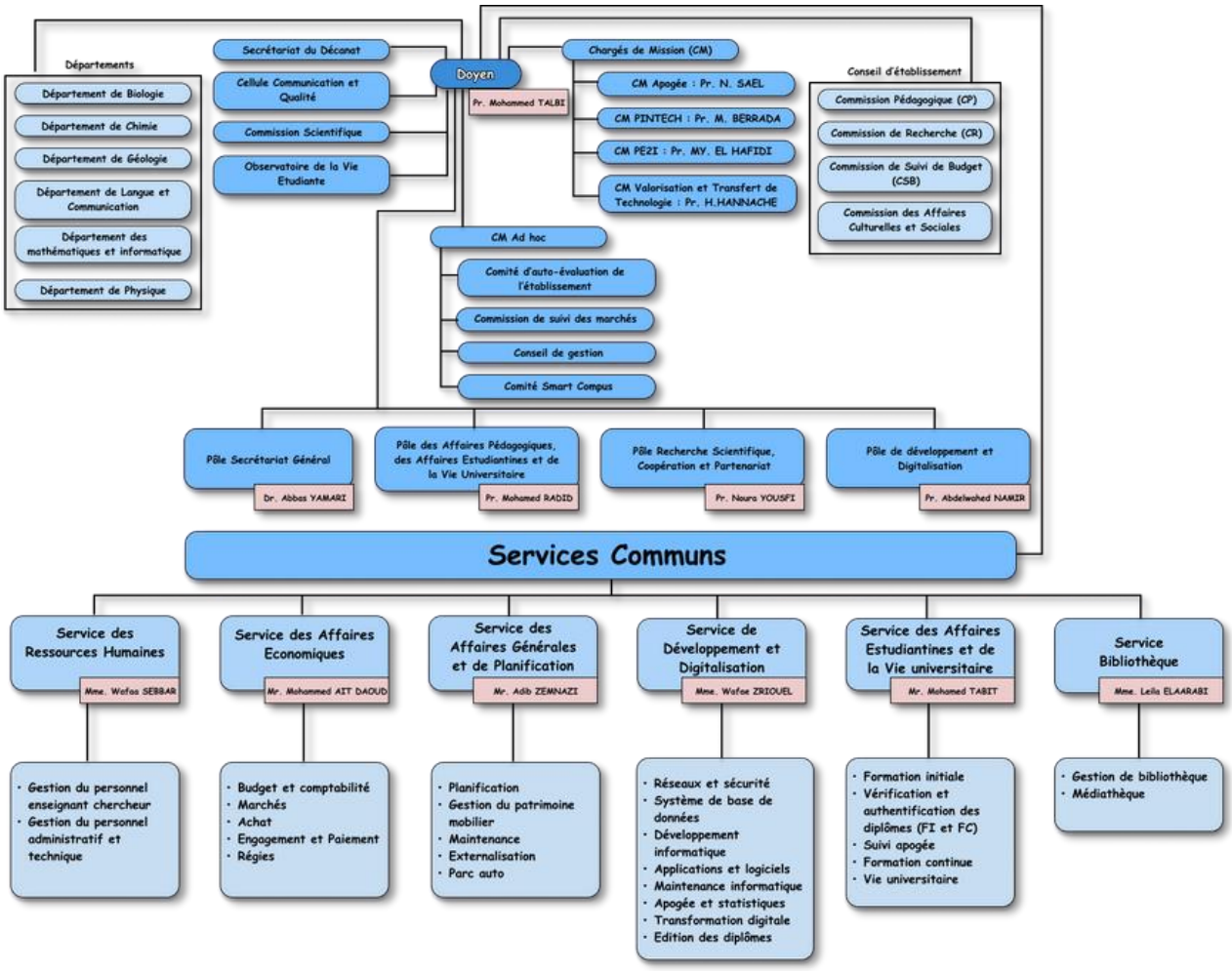


Figure 1: Organigramme de la faculté des sciences Ben M'Sik

2. Cadre de projet :

2.1. Contexte :

Dans le secteur de l'assurance, la gestion des réclamations est souvent complexe et sujette à des inefficacités, ainsi qu'à des tentatives de fraude. Les méthodes traditionnelles de détection de la fraude, telles que les audits manuels et les systèmes basés sur des règles, sont limitées par leur capacité à s'adapter aux schémas de fraude évolutifs. De plus, elles ne permettent pas toujours de détecter des comportements frauduleux subtils.

2.2. Problématique :

La détection des fraudes dans les réclamations d'assurance est un défi complexe en raison de la diversité des comportements frauduleux et de l'évolution continue des techniques de fraude. Les méthodes actuelles basées sur des règles fixes ou des audits manuels ne permettent pas toujours de détecter des anomalies subtiles ou émergentes. Il est donc crucial

de développer une approche plus dynamique et adaptable pour analyser les données des réclamations et identifier les comportements suspects.

2.3. Solution et Développement :

La solution proposée repose sur l'application du process mining pour analyser les réclamations d'assurance, Le développement comprend plusieurs étapes clés :

- **Collecte et Préparation des Données** : Rassemblement des journaux d'événements relatifs aux réclamations d'assurance et préparation des données pour l'analyse.
- **Analyse des Processus** : Utilisation des outils de process mining pour modéliser les processus de réclamation et identifier les déviations par rapport aux procédures standards.
- **Transformation des Données** : Extraction des caractéristiques pertinentes pour améliorer la précision des modèles d'analyse.
- **Prédiction des Anomalies** : Application de techniques d'apprentissage automatique pour détecter les anomalies et les comportements potentiellement frauduleux.

Cette approche permet non seulement d'améliorer la gestion des réclamations mais aussi de détecter des schémas de fraude plus complexes, offrant ainsi une solution robuste et adaptable.

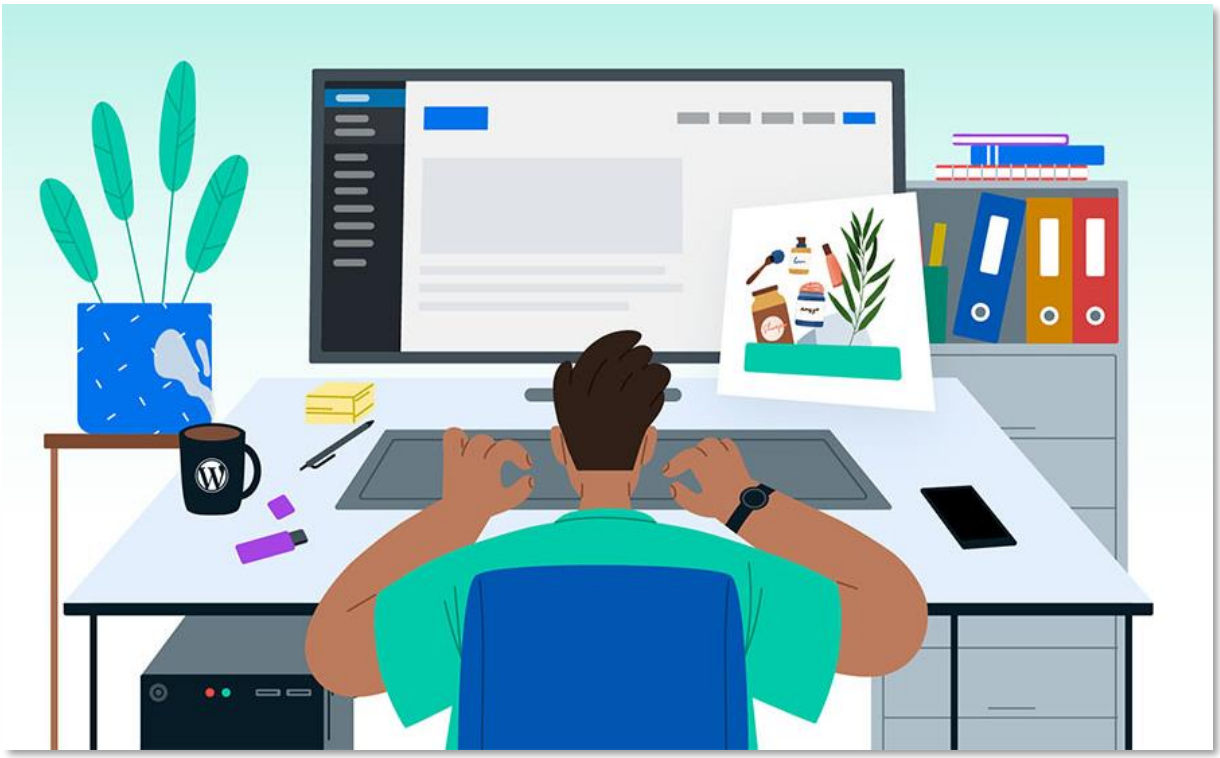
2.4. Périmètre du projet :

Le périmètre du projet se concentre sur l'analyse des réclamations d'assurance à partir des journaux d'événements fournis. Il comprend la collecte et la préparation des données, l'application des techniques de process mining pour modéliser les processus de réclamation, et l'utilisation de méthodes d'apprentissage automatique pour détecter les anomalies et les fraudes. Le projet ne couvre pas les aspects de la gestion des sinistres post-détection ou l'intégration de la solution dans des systèmes existants d'assurance, mais se limite à l'analyse et à la prédiction basées sur les données disponibles.

Conclusion

Ce chapitre a défini le contexte du projet, en présentant les spécifications et les objectifs à atteindre, tout en alignant ces éléments avec le plan général. La solution développée illustre comment le process mining peut améliorer l'analyse des réclamations d'assurance et détecter les anomalies de manière efficace

Chapitre 2 : Étude & Analyse du Projet



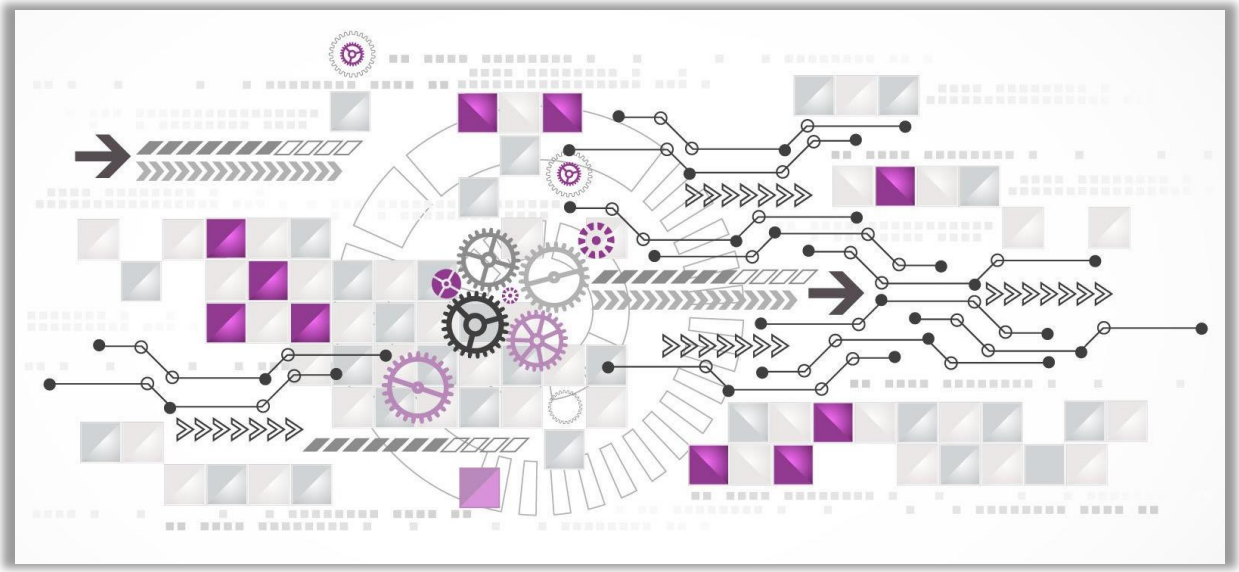
Introduction

Dans ce chapitre, l'étude se concentre sur une analyse approfondie du projet, mettant en avant l'application des techniques de process mining pour l'analyse des sinistres d'assurance. L'objectif principal est de comprendre les aspects techniques et conceptuels associés à l'utilisation du process mining pour détecter les anomalies et les fraudes dans les réclamations. Cette étude examine les différentes approches méthodologiques et technologiques, en mettant l'accent sur l'application des outils de modélisation des processus et des techniques d'apprentissage automatique. L'implémentation de ces solutions jouera un rôle crucial dans l'amélioration de l'efficacité de la gestion des réclamations et dans la détection précoce des comportements frauduleux.

1. Généralités sur Process Mining

Nous commencerons par introduire les concepts fondamentaux de l'IoT (Internet des objets), après nous explorerons l'histoire de l'IoT. Enfin nous discuterons également des objectifs& architecture des systèmes IoT.

1.1 Introduction :



Le process mining est une technique d'analyse des processus qui extrait des connaissances à partir des journaux d'événements générés par les systèmes informatiques. En offrant une vue détaillée des processus opérationnels réels, le process mining permet d'identifier les inefficacités, les écarts et les anomalies dans les flux de travail. Cette méthode repose sur l'analyse des données d'événements pour visualiser, comprendre et améliorer les processus métiers en temps réel.

1.2 L'histoire :

Le concept de process mining a émergé au début des années 2000 avec le développement des outils de gestion des processus et l'augmentation des capacités de traitement des données. Initialement axé sur la modélisation des processus basés sur des événements, le process mining a évolué pour intégrer des techniques avancées d'analyse et d'apprentissage automatique, offrant ainsi des solutions plus sophistiquées pour l'optimisation des processus et la détection des anomalies.

1.3 L'objectif :

Le process mining vise à améliorer la compréhension des processus métiers en fournissant des représentations visuelles et analytiques des flux de travail. Les objectifs principaux incluent la détection des inefficacités, l'optimisation des processus, et l'identification des anomalies ou des fraudes. En facilitant une analyse approfondie des données d'événements, le process mining permet aux organisations d'améliorer leur efficacité opérationnelle et de prendre des décisions basées sur des données précises.

1.4 L’Architecture :

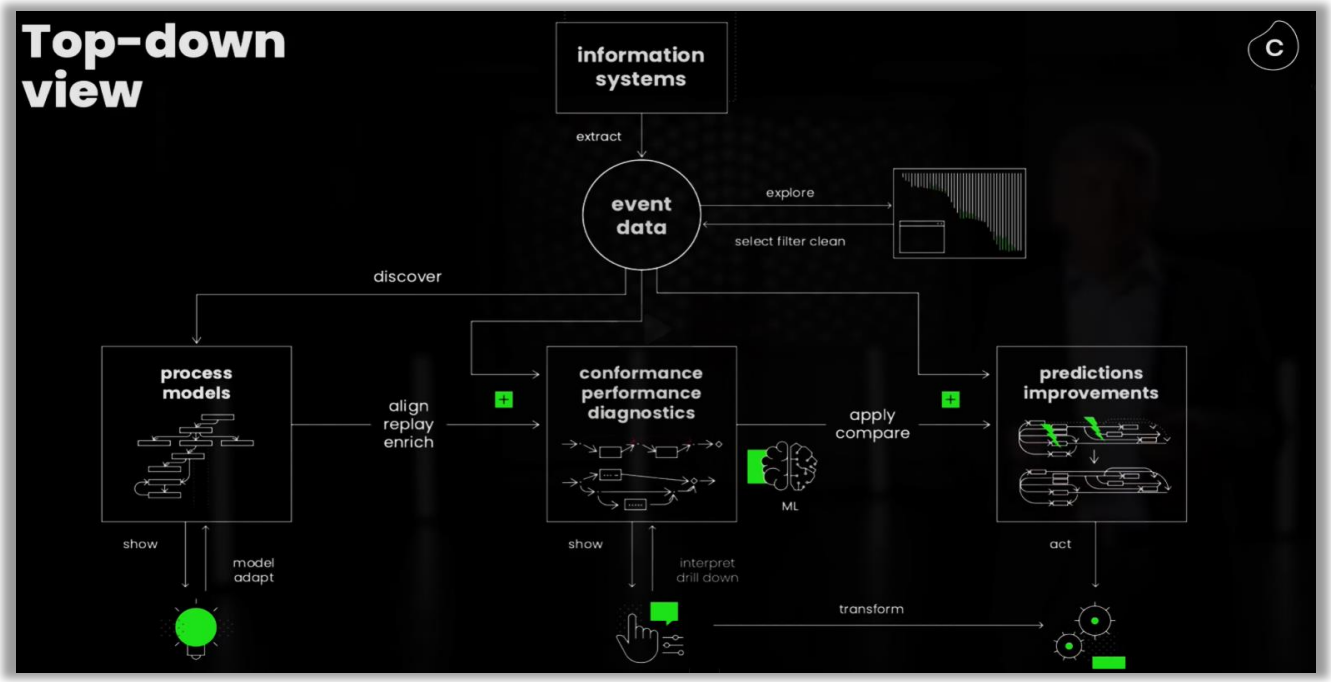


Figure 2 : architecture générale de process mining

Le process mining est une méthode utilisée pour découvrir, surveiller et améliorer les processus réels (et non les processus supposés) en extrayant des connaissances à partir des journaux d'événements disponibles dans les systèmes d'information modernes. Les principaux types de process mining sont :

- **Découverte** : Création d'un modèle de processus à partir des journaux d'événements sans aucune information préalable.
- **Conformité** : Vérification de la conformité de la réalité, telle qu'enregistrée dans le journal, avec le modèle et vice versa.
- **Amélioration** : Amélioration ou extension d'un modèle de processus existant en utilisant les informations sur le processus réel enregistré dans le journal.

1.5 Les Étapes :

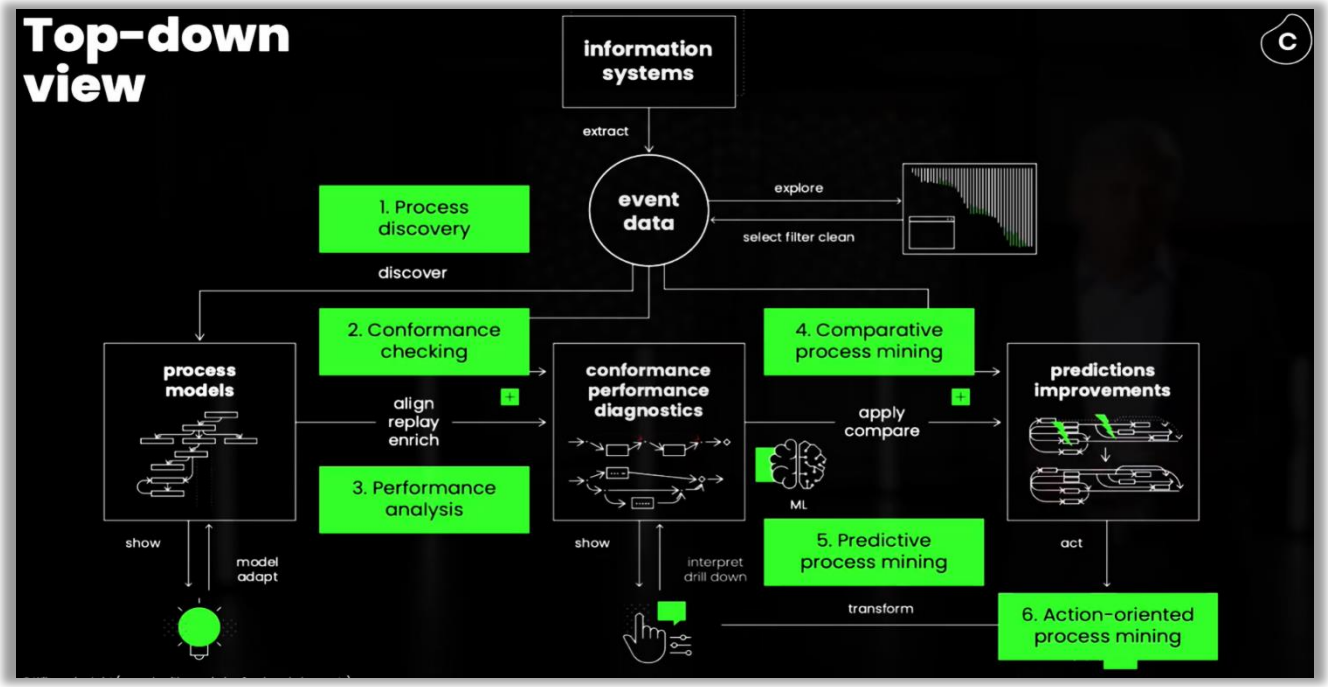


Figure 3 : les étapes d'un projet process mining

- **Découverte des Processus (Process Discovery)** : Cette étape consiste à utiliser les données d'événements pour créer un modèle de processus. En appliquant des techniques de process mining, un modèle visuel des processus réels est généré à partir des données collectées. Ce modèle permet de visualiser les flux de travail et d'identifier les variantes de processus.
- **Vérification de la Conformité (Conformance Checking)** : Une fois le modèle créé, il est comparé au comportement observé dans les données réelles. Cette phase permet de vérifier si le modèle correspond aux processus réels et d'identifier les écarts ou les non-conformités. Cela aide à comprendre les divergences entre les processus théoriques et les processus réels.
- **Analyse de la Performance (Performance Analysis)** : Cette étape vise à identifier les goulets d'étranglement et les points de défaillance dans le processus. En examinant les performances des processus, cette phase permet de comprendre les causes sous-jacentes des inefficacités et de proposer des améliorations pour optimiser les processus.
- **Mining Comparatif des Processus (Comparative Process Mining)** : Le mining comparatif consiste à comparer différents processus ou différentes instances du même processus. Cette comparaison permet de trouver des domaines d'amélioration en identifiant les variations de performance et les meilleures pratiques à adopter.
- **Mining Prédictif des Processus (Predictive Process Mining)** : Cette étape utilise les données historiques pour prédire les problèmes futurs et les risques potentiels. En appliquant des techniques prédictives, il est possible d'anticiper les anomalies et les défaillances avant qu'elles ne se produisent, permettant ainsi une action précoce pour les éviter.

- **Mining Orienté Action (Action-Oriented Process Mining)** : La dernière étape transforme les insights obtenus en actions concrètes. En se basant sur les analyses et les prévisions, des recommandations sont élaborées pour améliorer les processus. Cette phase met en œuvre des mesures pour optimiser les opérations et adresser les problèmes identifiés.

1.6 Approche et Concepts :

- **Approche Principale** : Le process mining est devenu l'approche principale basée sur les données pour la gestion des processus métier (BPM).
- **Intersection des Sciences** : Le process mining est le point de rencontre entre *la Data Science* (qui ne considère pas les processus) et *la Science des Processus* (qui ne considère pas les données).

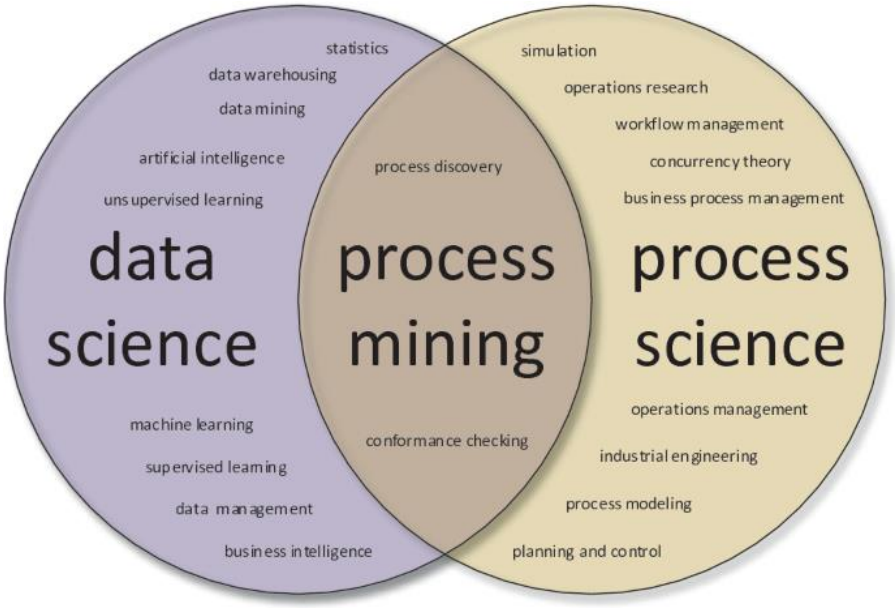


Figure 4 : intersection entre data & process science

- **Objectif** : Utiliser les données d'événements pour améliorer les processus, la performance et la conformité, afin de suivre et d'optimiser les processus de manière correcte.
- **Données d'Entrée** : Données d'événements / journaux d'événements.
- **Composants des Événements** : Les événements peuvent avoir plusieurs cas et se composent de :
  - Cas
  - Activité
  - Horodatage

- Ressources (optionnelle)

Figure 5 : exemple d'un jeu des données - events log

- **Concept de Trace** : Pour les grands ensembles de données, si les données d'événements contiennent des événements qui se produisent plusieurs fois mais dans des cas ou horodatages différents, la séquence la plus fréquente est appelée trace, représentant un processus normal/correct.

Figure 6 : regroupement des cas par fréquence

- **Avantages** : Aide à identifier les problèmes de performance et de conformité, à comprendre ce qui se passe et comment les processus sont exécutés.

- **Perspectives sur les Processus** : Suivre les cas individuels, identifier les goulets d'étranglement, les écarts par rapport aux chemins normaux, et mettre en évidence les processus les plus courants et les écarts.

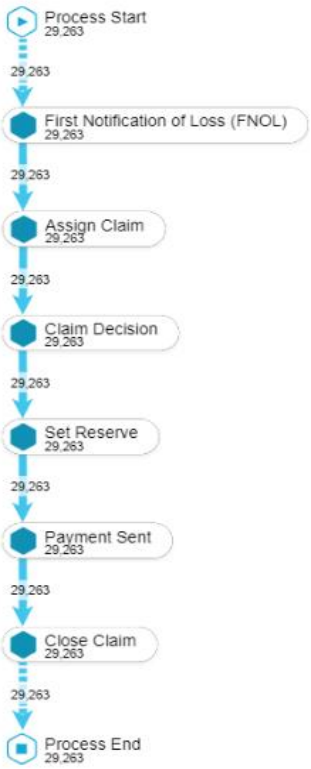


Figure 7 : exemple de processus utilisant DFG

- **Modèle de Processus Spaghetti** : Les processus initiaux peuvent être complexes (modèles spaghetti) et nécessitent un filtrage pour éliminer les chemins non souhaités ou non optimaux.

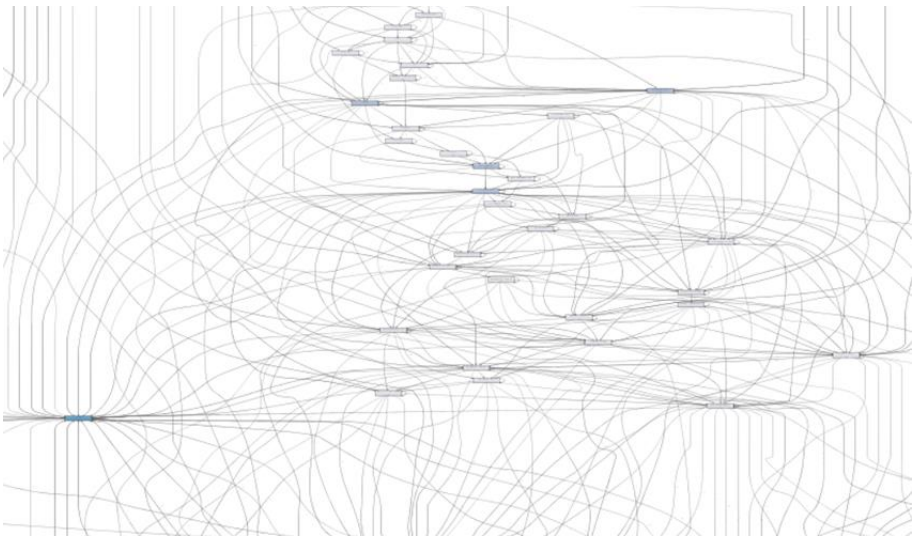


Figure 8 : exemple du modèle de processus spaghetti

- **Complexité** : Les processus réels ont souvent de nombreuses variations en raison des passations, des problèmes de réseau, des duplications, d'une communication inefficace, etc.
- **Diagnostic et Amélioration** : Le process mining aide les organisations à identifier les problèmes de performance et de conformité, à comprendre leurs causes, et à suggérer des actions correctives.
- **Capacité Prédictive** : Prédit si certains processus rencontreront des problèmes à l'avenir (probabilité d'écart).
- **Intégration de l'Apprentissage Automatique** : Les techniques d'apprentissage automatique classiques ont du mal avec les données d'événements, mais les modèles de processus alignés avec les données d'événements peuvent tirer parti des techniques d'apprentissage automatique standard.
- **Objectif** : Améliorer les processus grâce aux insights pour modifier les processus ou déclencher automatiquement des flux de travail pour résoudre les problèmes identifiés ou prédits.

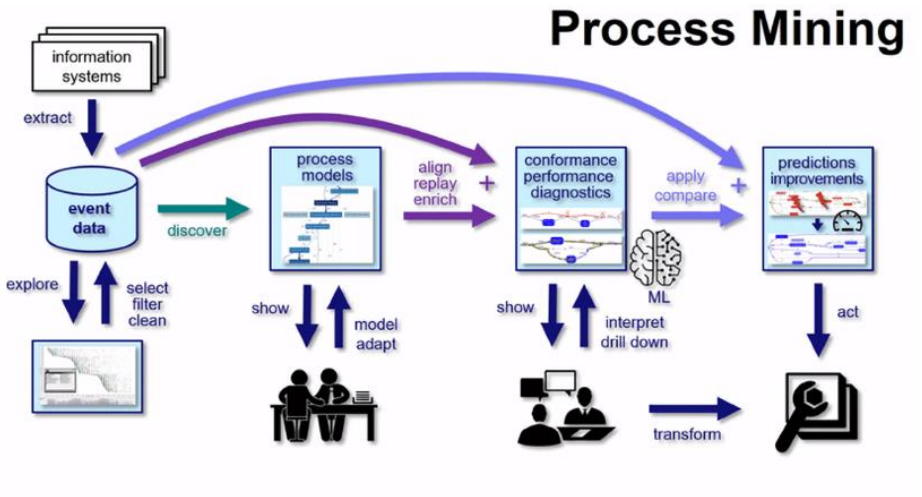


Figure 9 : cycle de vie d'un projet process mining

2. Données de Journal des Sinistres d'Assurance

2.1 Description :

L'analyse des données du journal des réclamations d'assurance permet de comprendre et d'optimiser le processus de traitement des réclamations. Les données fournies incluent des informations sur chaque événement lié à une réclamation, depuis la notification initiale jusqu'à la clôture finale. Cette étude vise à explorer la structure des données, à analyser les séquences d'activités, et à identifier les points d'amélioration potentiels.

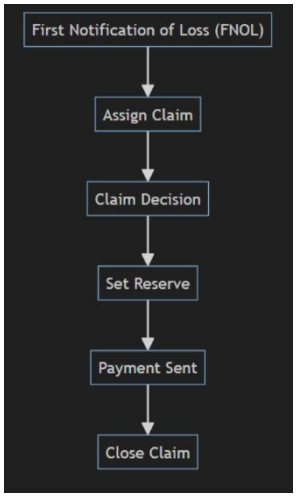
2.2 Structure des Données :

Les données se présentent sous forme de tableau avec les colonnes suivantes :

- `case_id` : Identifiant unique pour chaque dossier de réclamation.
- `activity_name` : Nom de l'activité ou étape du processus de traitement de la réclamation.
- `timestamp` : Date et heure de l'événement.
- `claimant_name` : Nom du demandeur de l'assurance.
- `agent_name` : Nom de l'agent chargé du dossier.
- `adjuster_name` : Nom de l'expert en évaluation.
- `claim_amount` : Montant total de la réclamation.
- `claimant_age` : Âge du demandeur.
- `type_of_policy` : Type de police d'assurance (par exemple, complète, collision).
- `car_make` : Marque de la voiture concernée.
- `car_model` : Modèle de la voiture.
- `car_year` : Année de fabrication de la voiture.
- `type_of_accident` : Type d'accident (par exemple, frontal, arrière).
- `user_type` : Type d'utilisateur ayant traité la réclamation (par exemple, humain, RPA).

2.3 Les Activités :

Les principales activités identifiées dans les données sont :



- **First Notification of Loss (FNOL)** : La notification initiale de la perte est la première étape où le demandeur informe l'assurance de l'événement.
- **Assign Claim** : Attribution de la réclamation à un agent ou un expert.
- **Claim Decision** : Décision sur la réclamation, déterminant si elle est acceptée ou rejetée.
- **Set Reserve** : Fixation de la réserve financière pour la réclamation.
- **Payment Sent** : Envoi du paiement au demandeur.
- **Close Claim** : Clôture du dossier après le règlement complet.

Chaque dossier suit cette séquence d'activités, et les données montrent les timestamps précis pour chaque étape.

2.4 Exemples de Cas :

Prenons le dossier avec `l'case_id 000112d5-9d04-450f-820f-3edfc0626cf9` :

- **Chronologie des Activités** :
  - **FNOL** : 19 avril 2022
  - **Assign Claim** : 1er mai 2022
  - **Claim Decision** : 3 mai 2022
  - **Set Reserve** : 8 mai 2022
  - **Payment Sent** : 15 mai 2022
  - **Close Claim** : 20 mai 2022
- **Montant de la Réclamation** : 9266,19 USD
- **Détails de la Voiture** : Hyundai Elantra 2021

- **Type d'Accident** : Collision frontale
- **Type d'Utilisateur** : Humain

2.5 Exploration de données :

Le jeu de données contient les informations suivantes :

- **cas :** 30 000
- **clients :** 25 778
- **d'agents :** 25 862
- **d'ajusteurs :** 25 797
- **types de polices :** 3
- **marques de voiture :** 7
- **types d'accidents :** 4
- **activités :** 6
- **Durée moyenne de traitement des réclamations (jours) :** 35
- **d'états :** 3

3. Architecture générale

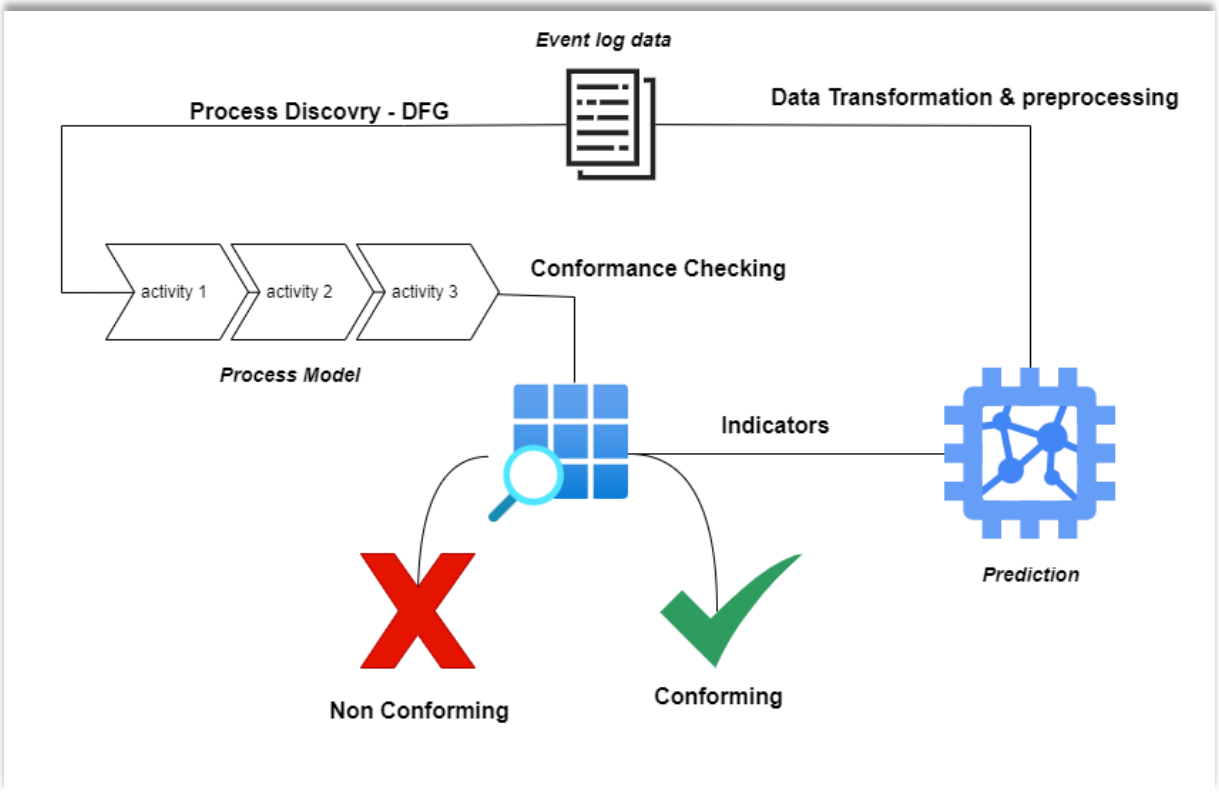


Figure 10 : l'architecture générale du projet

- **Commencer avec les Données de Logs d'Événements :** Commencez par collecter et intégrer les données de logs d'événements, qui incluent des informations détaillées sur chaque activité dans le processus de réclamation.

- **Découverte de Processus - DFG :** Utilisez des techniques de découverte de processus comme les Graphes Suivi-Direct (DFG) pour visualiser le flux réel des activités, en créant un modèle de processus à partir des logs d'événements.
- **Contrôle de Conformité :** Comparez le modèle de processus découvert avec le modèle de processus idéal pour identifier les cas conformes et non conformes.
- **Analyse des Indicateurs :** Analysez les indicateurs clés pour mieux comprendre la performance du processus et détecter les problèmes potentiels.
- **Transformation & Prétraitement des Données :** Transformez et prétraitez les données pour les préparer à l'analyse, en veillant à ce qu'elles soient propres et standardisées.
- **Prédiction :** Utilisez des modèles d'apprentissage automatique pour prédire le statut de nouveaux cas basé sur l'analyse des cas passés et les indicateurs identifiés.

4. Modèle de processus

4.1 Flux de processus :

D'après l'application de la technique Directly Follows Graphs (DFG), on trouve que 98 % des cas ont une variante qui consiste en un flux de processus :

First Notification of Loss (FNOL) → Assign Claim → Claim Decision → Set Reserve → Payment Sent → Close Claim

On dit que ce processus est standard

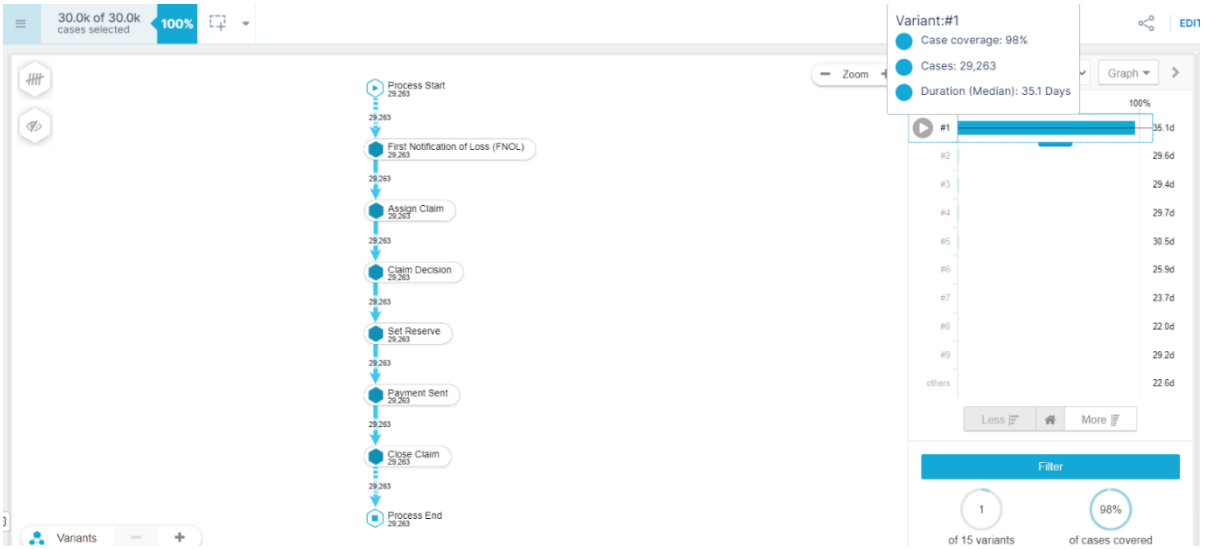


Figure 11 : DFG du sinistres d'assurance avec une variante

4.2 Flux des processus avec Déviation

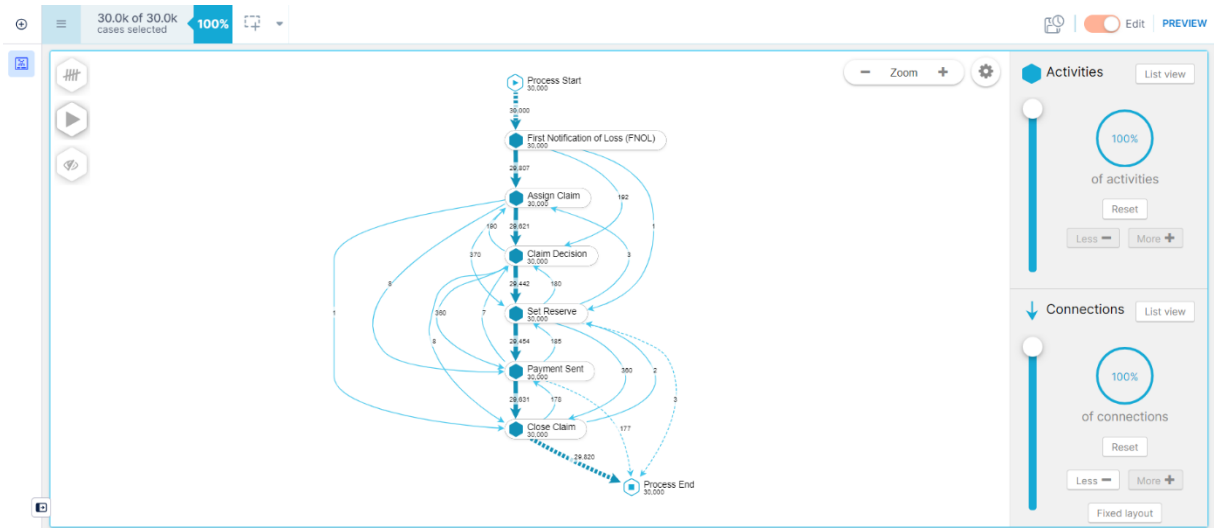


Figure 12 : DFG du sinistres d'assurance avec toutes les variantes

L'analyse des flux des processus avec déviation peut révéler des incohérences ou des écarts par rapport aux procédures standardisées, exemple des déviations dans l'ordre des activités :

- **Non-Conformité avec FNOL** : Le processus ne commence pas toujours par la "First Notification of Loss (FNOL)".
- **Suivi Inhabituel après FNOL** : Dans certains cas, la "First Notification of Loss (FNOL)" est suivie directement par d'autres activités telles que "Set Reserve", "Payment Sent", ou "Claim Decision", ce qui peut indiquer des anomalies.
- **Incohérences dans la Séquence** d'Assignation : L'activité "Assign Claim" est parfois suivie directement par "Set Reserve" ou "Payment Sent", ce qui ne correspond pas au flux attendu.
- **Discontinuité après Payment Sent** : L'activité "Payment Sent" est suivie par "Set Reserve" dans certains cas, ce qui est atypique.

5. Contrôle de conformité

5.1 Processus standardisé comparant les données :

L'analyse des flux de processus met en lumière plusieurs aspects clés du fonctionnement des réclamations, ainsi que les déviations par rapport au modèle attendu.

Le Processus Standard est :

- *First Notification of Loss (FNOL) → Assign Claim → Claim Decision → Set Reserve → Payment Sent → Close Claim*
- Ce flux décrit la séquence typique des activités de traitement des réclamations d'assurance, assurant un passage systématique par chaque étape essentielle.

Modèle de Processus Métier (BPM) :

- **Business Process Model (BPM) observé :** Après l'application des techniques de contrôle de conformité, le processus observé peut différer du modèle standard à cause de divers facteurs opérationnels.
- **Visualisation du BPM :** La visualisation montre comment les réclamations sont effectivement traitées, en comparant les flux standards avec les chemins réellement suivis dans les cas réels.

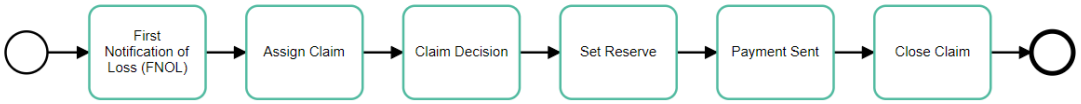


Figure 13 : Représentation du processus BMP

Conformité avec le Modèle :

- **98% de conformité :** 98% des cas suivent strictement le modèle de processus standard, ce qui indique une forte adhésion aux procédures établies.
- **2% de non-conformité :** 2% des cas ne respectent pas le modèle de processus, révélant des déviations potentiellement dues à des erreurs, des cas exceptionnels ou des inefficacités dans le système.

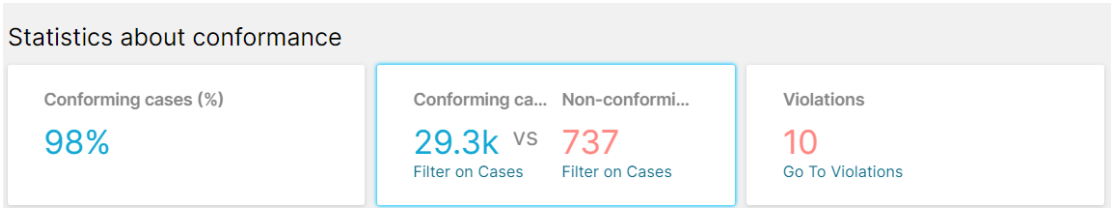


Figure 14 : résumé du contrôle de conformité

Violations du Processus :

- **10 Violations Identifiées :** Ces violations représentent des séquences d'activités qui ne respectent pas le modèle de processus standard.
- **Impact des Violations :** Les violations peuvent indiquer des problèmes structurels dans le flux de travail, nécessitant une enquête approfondie pour déterminer leur cause et leur impact sur l'efficacité globale du traitement des réclamations.
- **Non-Conformant Cases :** Les violations uniques identifiées correspondent à des cas non conformes où l'ordre des activités a été modifié, perturbant ainsi le processus fluide attendu.

5.2 Exemple des Déviations :

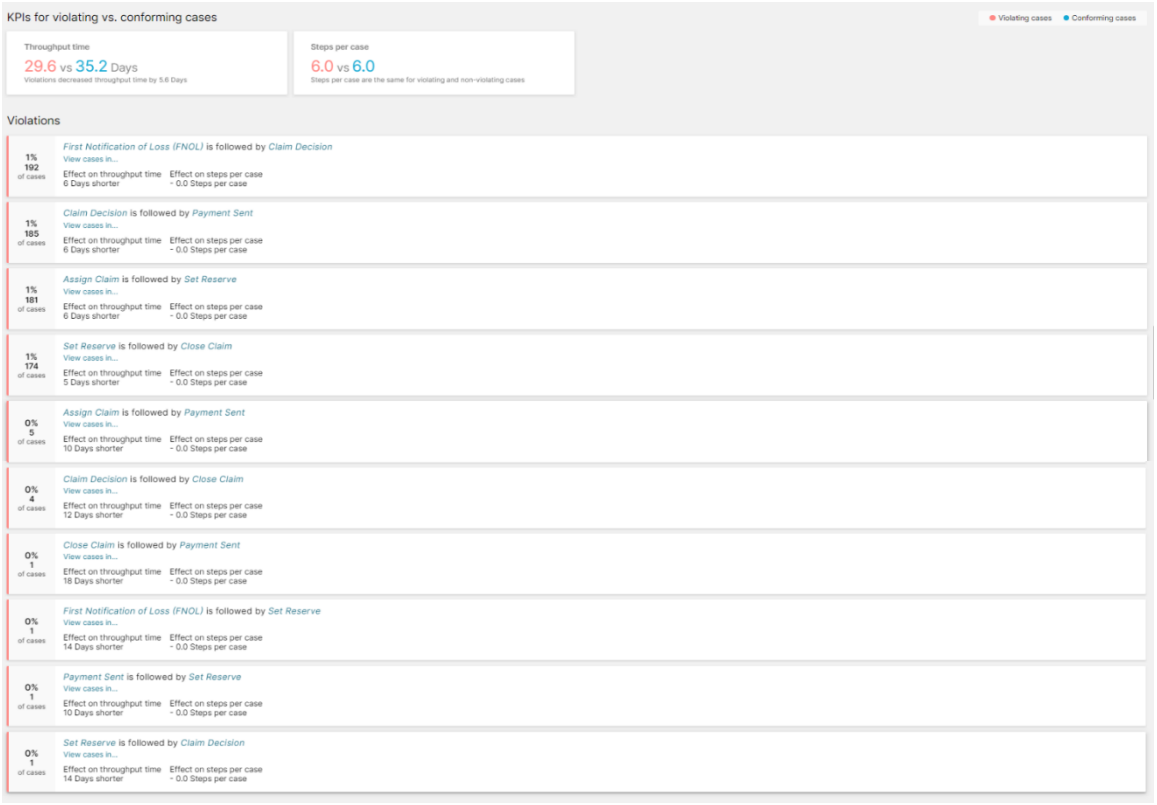


Figure 15 : liste des déviations et de leurs causes

L'analyse des flux de processus de réclamations d'assurance montre que certaines séquences d'activités peuvent indiquer des cas frauduleux ou des violations, selon l'ordre dans lequel les étapes sont exécutées

- **Not Started with FNOL (Fraude potentielle) :** Un processus de réclamation qui ne commence pas par la **Première Notification de Perte (FNOL)** peut indiquer une tentative de dissimulation d'une réclamation non légitime, où des étapes critiques sont omises pour éviter un examen minutieux.
- **First Notification of Loss (FNOL) is Followed by Set Reserve (Fraude potentielle) :** Passer directement à la **Mise en Réserve** sans passer par l'**Attribution de la Réclamation** ou la **Décision de Réclamation** peut être un signe que les réserves sont manipulées prématurément pour affecter les finances de la compagnie.
- **First Notification of Loss (FNOL) is Followed by Payment Sent (Fraude potentielle)** Envoyer le paiement directement après la FNOL sans évaluation ou décision peut indiquer une tentative d'accélération frauduleuse du processus de paiement, potentiellement pour des réclamations non légitimes.
- **First Notification of Loss (FNOL) is Followed by Claim Decision (Fraude potentielle) :** La prise de décision immédiate sans une évaluation appropriée (comme l'**Attribution de la Réclamation**) peut indiquer un traitement bâclé ou manipulé pour approuver des réclamations non fondées.
- **Assign Claim is Followed by Set Reserve (Violation du Processus)** Si la **Mise en Réserve** suit immédiatement l'**Attribution de la Réclamation**, cela peut montrer une violation

du flux normal du processus, suggérant une évaluation prématurée qui pourrait ne pas tenir compte de toutes les données.

- **Assign Claim is Followed by Payment Sent (Fraude potentielle) :** Le paiement envoyé sans une décision appropriée ou mise en réserve peut signifier que les fonds sont distribués sans une justification suffisante, ce qui est une indication de fraude potentielle.
- **Payment Sent is Followed by Set Reserve (Violation du Processus) :** La mise en réserve après l'envoi du paiement est généralement anormale et suggère que le contrôle financier est perdu après la dispersion des fonds, ce qui est un signal d'alerte pour une possible manipulation de données.

6. Analyse prédictive

L'analyse prédictive dans le contexte des réclamations d'assurance vise à utiliser les données historiques pour prévoir l'état futur d'une réclamation, qu'elle soit conforme, une violation, ou potentiellement frauduleuse. En s'appuyant sur des indicateurs clés et des techniques de **machine learning supervisé**, nous pouvons former des modèles pour identifier et prédire ces états avec précision.

6.1 Indicateurs et Leur Rôle dans l'Analyse Prédictive

- **Nombre d'Activités Inférieur à 6 :** Un nombre d'activités inférieur à 6 peut indiquer qu'une réclamation n'a pas suivi le processus standard ou qu'elle a sauté certaines étapes critiques. Cela est souvent un signe d'anomalie et peut être un indicateur de fraude ou de violation. En formant un modèle supervisé, ces cas peuvent être étiquetés comme non conformes, et le modèle peut apprendre à les identifier comme potentiellement problématiques.
- **Temps de Traitement Inférieur à 29 Jours (Moyenne de 35 Jours) :** Un délai de traitement plus court que la moyenne peut suggérer que certaines étapes ont été omises ou exécutées trop rapidement, ce qui pourrait signaler une fraude. Le modèle d'apprentissage supervisé peut utiliser cette information pour étiqueter ces cas comme des anomalies. En analysant des données historiques, le modèle peut apprendre que les cas plus rapides que la moyenne ont une probabilité accrue d'être non conformes.
- **Processus Déviant Indiquant une Fraude :** Les séquences d'activités inhabituelles, telles que l'absence de la **Première Notification de Perte (FNOL)** ou un enchaînement anormal des étapes, peuvent servir d'indicateurs clés de fraude. Ces séquences sont cruciales pour entraîner un modèle à reconnaître les schémas qui divergent du processus standard et à prédire les cas potentiellement frauduleux ou suspects.
- **Violations Identifiées :** Les violations du processus, comme des étapes exécutées dans le mauvais ordre, permettent d'étiqueter des cas comme potentiellement non conformes dans un modèle supervisé. Les données de ces violations fournissent un ensemble riche pour entraîner le modèle à identifier ces schémas de manière automatique.

6.2 Indicateurs & Étiqueter des Données pour le Modèle Supervisé :

Pour créer un modèle d'apprentissage supervisé efficace, il est essentiel de disposer de données correctement étiquetées qui représentent différents états possibles des réclamations

Étiquetage des Données :

- En analysant les indicateurs comme le nombre d'activités, le temps de traitement, et les séquences d'activités, nous pouvons attribuer des labels aux cas dans le jeu de données, par exemple :
  - **Conforme** : Si le processus suit exactement le chemin standard sans déviations.
  - **Violation** : Si des anomalies mineures sont détectées mais ne sont pas nécessairement frauduleuses.
  - **Fraude** : Si les schémas détectés correspondent à des indicateurs clairs de comportement suspect ou de tentative de tromperie.
- Ce type d'étiquetage est crucial pour entraîner le modèle à distinguer entre les différents états de réclamations avec précision.

Apprentissage des Schémas :

- En fournissant ces données étiquetées, le modèle peut apprendre à reconnaître les schémas associés à chaque catégorie. Par exemple, des modèles comme les arbres de décision, les forêts aléatoires, k-voisins les plus proches, ou les réseaux neuronaux peuvent être utilisés pour capturer ces schémas complexes et prévoir l'état d'une réclamation future.

7. Étude des solutions

Dans le cadre de l'analyse des réclamations d'assurance et de la détection des violations et des fraudes, plusieurs outils et technologies ont été utilisés pour mettre en œuvre une solution robuste et efficace

7.1 Celonis :

Figure 16 : celonis logo

- **Description :** Celonis est une plateforme de process mining et d'analytique qui permet de visualiser et d'optimiser les processus métiers en analysant les logs d'événements. Elle offre des capacités avancées pour découvrir des modèles de processus, vérifier la conformité, et identifier les inefficacités et anomalies dans les processus.
- **Fonctionnalités Clés :**
  - **Process Mining :** Identification automatique des processus réels à partir des données.
  - **Conformance Checking :** Comparaison des processus réels avec le modèle de processus idéal pour détecter les déviations.
  - **Visualisation Interactive :** Visualisation graphique des flux de processus pour une meilleure compréhension.
  - **Analytique Temporelle :** Analyse des délais entre les activités pour identifier les goulets d'étranglement.
- **Utilisation dans le Projet :**
  - Celonis a été utilisé pour découvrir les modèles de processus à partir des logs d'événements et pour effectuer un contrôle de conformité sur les réclamations d'assurance. Cela nous a permis de visualiser le flux de processus et d'identifier les écarts par rapport au modèle standard.
- **Avantages :**
  - Interface intuitive et riche en fonctionnalités.
  - Capacités d'analyse en temps réel.
  - Intégration facile avec d'autres systèmes.
- **Limites :**
  - Peut être coûteux pour les petites entreprises.
  - Courbe d'apprentissage pour les nouvelles utilisateurs.

7.2 Python & Jupyter Notebook :

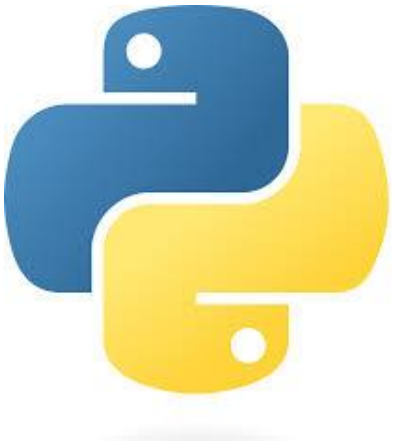


Figure 17 : python logo

Figure 18 : jupyter nootebook logo

- **Description :** Python est un langage de programmation polyvalent et puissant, largement utilisé pour l'analyse de données et le développement d'applications d'apprentissage automatique. Jupyter Notebook est un environnement interactif qui permet d'exécuter du code Python et de documenter des analyses dans un format de notebook.
- **Fonctionnalités Clés :**
  - **Flexibilité :** Large choix de bibliothèques pour la manipulation des données et l'analyse statistique.
  - **Intégration avec des Outils de Science des Données :** Bibliothèques comme Pandas, NumPy, Matplotlib, etc.
  - **Support pour l'Apprentissage Automatique :** Intégration avec des bibliothèques de machine learning comme Scikit-learn et TensorFlow.
- **Utilisation dans le Projet :**
  - Python a été utilisé pour transformer et prétraiter les données des logs d'événements, effectuer des analyses exploratoires et implémenter des modèles d'apprentissage automatique pour la prédiction de conformité des cas.
  - Jupyter Notebook a facilité la documentation de l'analyse et le partage de résultats de manière interactive.
- **Avantages :**
  - Open-source et gratuit.
  - Écosystème riche et en croissance rapide.
  - Grande communauté et abondance de ressources.
- **Limites :**
  - Peut être lent pour des opérations intensives sur de grands volumes de données sans optimisation.
  - Nécessite des compétences en programmation.

7.3 Scikit-learn :

Figure 19 : sklearn logo

- **Description :** Scikit-learn est une bibliothèque Python spécialisée dans l'apprentissage automatique. Elle propose des outils efficaces et simples pour le data mining et l'analyse de données.
- **Fonctionnalités Clés :**
  - **Algorithmes d'Apprentissage Supervisé :** Inclut des algorithmes de classification, régression et clustering.
  - **Sélection de Modèles :** Outils pour la validation croisée et la recherche d'hyperparamètres.
  - **Prétraitement des Données :** Capacités pour normaliser, transformer et sélectionner des caractéristiques.
- **Utilisation dans le Projet :**
  - Scikit-learn a été utilisé pour développer et entraîner des modèles de classification afin de prédire la conformité des réclamations d'assurance. Cela inclut la sélection des caractéristiques, le tuning des hyperparamètres, et l'évaluation des modèles.
- **Avantages :**
  - Facilité d'utilisation et documentation détaillée.
  - Performance efficace pour des jeux de données de taille moyenne.
  - Intégration fluide avec d'autres bibliothèques Python.
- **Limites :**
  - Limitée pour les très grands jeux de données et les applications complexes d'apprentissage profond.
  - Manque de support intégré pour les GPU.

7.4 Comparaison et Intégration :

- **Celonis** est idéal pour la découverte de processus et l'analyse des logs d'événements à grande échelle, offrant des insights immédiats sur les processus métiers et les inefficacités.
- **Python & Jupyter Notebook** fournissent une flexibilité inégalée pour les analyses exploratoires et la construction de modèles d'apprentissage machine, permettant une personnalisation et un développement rapide.

- **Scikit-learn** complète ces outils en fournissant une boîte à outils robuste pour l'apprentissage automatique, essentielle pour prédire et classer les cas en conformité ou non.

Conclusion

Ce chapitre nous a permis d'approfondir notre compréhension du projet en explorant les différents aspects liés au Process Mining, aux technologies associées, et à l'étude des différents aspects du projet, depuis la collecte des données jusqu'à la prédiction. Ces connaissances constituent une base solide pour la mise en œuvre du modèle de prédiction, en nous permettant de mieux appréhender les flux de travail, d'identifier les écarts et d'optimiser les processus grâce aux techniques avancées d'analyse et d'apprentissage automatique.

Chapitre 3 : Implémentation

Introduction

Ce chapitre se concentre sur l'implémentation des processus de Process Mining et des techniques de modélisation prédictive. Nous explorons le chargement et l'analyse des données, ainsi que leur transformation et étiquetage pour une meilleure préparation des modèles. Nous détaillons également la comparaison des modèles de classification, avec un accent particulier sur le réglage des hyperparamètres et l'évaluation des performances du modèle K-Nearest Neighbors (KNN), soulignant les méthodes utilisées pour optimiser les résultats.

1. Process Mining :

1.1. Chargement des données :

Celonis utilise une approche structurée pour le chargement et l'analyse des données, impliquant deux composants principaux : **Data Pools** (Pépinières de Données) et **Data Models** (Modèles de Données)

- *Celonis data intergration*

Data Pool (Pépinière de Données)

- Définition : Une Pépinière de Données est un dépôt dans Celonis où les données provenant de diverses sources sont collectées et stockées. Elle sert de zone de staging pour les données brutes avant toute analyse ou traitement.
- Processus :
  - Intégration des Données : Les données sont importées dans la Pépinière de Données depuis différentes sources telles que les bases de données, les systèmes ERP ou les fichiers plats. Cela peut se faire via les connecteurs Celonis ou par des intégrations personnalisées.
  - Stockage des Données : Une fois importées, les données sont stockées sous leur forme brute. Celonis prend en charge divers formats de données et peut gérer de grands volumes de données.
  - Transformation des Données : Des transformations et nettoyages de base peuvent être effectués au sein de la Pépinière de Données. Cela peut inclure des filtrages, des agrégations ou des jointures de jeux de données.

Data Model (Modèle de Données)

- Définition : Un Modèle de Données est construit sur la Pépinière de Données. Il définit la structure et les relations des données utilisées pour l'analyse. Le Modèle de Données fournit un cadre pour l'organisation, le lien et l'utilisation des données à des fins analytiques.
- Processus :
  - Définition du Schéma des Données : Les utilisateurs définissent le schéma du Modèle de Données, y compris les tables, les champs et les relations entre les différentes tables. Cela implique la création d'entités telles que les faits (ex. : transactions) et les dimensions (ex. : détails clients).
  - Transformation et Enrichissement des Données : Dans le Modèle de Données, les données peuvent être davantage transformées et enrichies. Cela peut inclure la création de champs calculés, l'agrégation des données ou l'application de règles métiers pour obtenir des insights.
  - Intégration des Données : Les données de la Pépinière de Données sont intégrées dans le Modèle de Données selon le schéma défini. Cette intégration garantit que les données sont structurées et prêtes pour l'analyse.
  - Exploration et Analyse des Données : Une fois le Modèle de Données établi, les utilisateurs peuvent effectuer des analyses avancées, créer des tableaux de bord et générer des rapports. Le Modèle de Données permet des requêtes sophistiquées et une visualisation approfondie des données.

1.2. Analyse des données :

Celonis Studio est l'environnement où les utilisateurs peuvent effectuer des analyses avancées et créer des visualisations pour explorer et interpréter les données.

- *Celonis analysis studio*

Accès à Celonis Studio :

- **Interface :** Celonis Studio est accessible via l'interface web de Celonis. Les utilisateurs naviguent dans l'environnement à partir du tableau de bord principal pour accéder à des outils d'analyse et de visualisation.

Analyse des Données :

- **Exploration des Données :** Les utilisateurs peuvent explorer les données en utilisant des filtres et des outils de recherche pour examiner des sous-ensembles spécifiques de données.
- **Visualisation :** Celonis Studio permet de créer diverses visualisations de données, telles que des graphiques, des tableaux de bord interactifs, et des cartes, pour faciliter l'analyse des tendances et des patterns.

Création de Rapports :

- **Tableaux de Bord :** Les utilisateurs peuvent concevoir des tableaux de bord personnalisés qui affichent des visualisations de données pertinentes. Ces tableaux de bord peuvent être partagés avec d'autres utilisateurs ou intégrés dans des rapports plus larges.
- **Rapports Dynamiques :** La création de rapports dynamiques permet de générer des documents détaillés basés sur les analyses effectuées. Ces rapports peuvent inclure des graphiques, des tableaux et des commentaires pour fournir un aperçu complet des données.

Analyse Avancée :

- **Process Mining :** Celonis Studio offre des outils de Process Mining pour analyser les processus métiers et identifier les inefficacités, les goulets d'étranglement et les opportunités d'amélioration. Les utilisateurs peuvent créer des modèles de processus et visualiser les flux de travail pour détecter des anomalies.
- **Data Modeling :** Les utilisateurs peuvent créer et affiner des modèles de données, définir des calculs personnalisés, et appliquer des transformations pour adapter les données aux besoins d'analyse spécifiques.

Collaboration et Partage :

- **Collaboration :** Les utilisateurs peuvent collaborer en temps réel sur des projets d'analyse. Ils peuvent ajouter des commentaires, partager des insights, et discuter des découvertes directement au sein de l'environnement de Celonis Studio.
- **Partage de Résultats :** Les résultats des analyses et les tableaux de bord peuvent être partagés avec d'autres membres de l'organisation ou intégrés dans des outils de reporting externes.

1.3. Tableau de bord :

CLAIMANT_NAME	CAR_MAKE	TYPE_OF_ACCIDENT	TYPE_OF_POLICY	CLAIM_AMOUNT	AGENT_NAME	USER_TYPE
Pamela Underwood	Hyundai	Side-impact	Collision	10k	Amber Moyer	Human
Marie Williams	Toyota	Rollover	Comprehensive	10k	Jessica Ruiz	Human
Michele Pineda	Ford	Side-impact	Collision	10k	Michael Howell	RPA
Carly Daniels	Jeep	Rear-end	Comprehensive	10k	Sharon Flores	Human
Morgan Walker	Jeep	Side-impact	Liability	10k	Leslie Travis	Human
Christopher Jacobs	Ford	Rollover	Liability	10k	Sandy Bailey	Human
Richard Miller	Ford	Rollover	Collision	10k	Samantha Olsen	Human
Tricia Levine	Honda	Rear-end	Liability	10k	Matthew Bender	RPA
Christine Sanchez	Nissan	Rear-end	Liability	10k	Jessica Goodwin	RPA
Jose Thompson	Toyota	Head-on	Comprehensive	10k	Stephen Howell	Human
Kelly Vazquez	Hyundai	Rollover	Liability	10k	Nathan Romero	Human
Allison Taylor	Ford	Side-impact	Collision	10k	Christopher Davis	Human
Megan Diaz	Chevrolet	Rear-end	Liability	10k	Ryan Marshall	Human
Robert Molina	Nissan	Head-on	Comprehensive	10k	Tony Thompson	RPA
David Wright	Hyundai	Rear-end	Collision	10k	Sarah Miller	Human

Figure 20 : tableau de bord de l'analyse

La création de tableaux de bord est une étape clé pour visualiser les résultats de l'analyse. Ces tableaux de bord permettent de présenter les insights de manière claire et interactive, facilitant la prise de décision basée sur les données.

2. Transformation et étiquetage des Données :

2.1. Réduire la dimension des données :

case_id	activity_name	timestamp	claimant_name	…
a626bc58-a7b2-4eff-aca9-2a622cd0c492	First Notification of Loss (FNOL)	2020-09-24 14:08:42.423530	Karen Lopez	…
a626bc58-a7b2-4eff-aca9-2a622cd0c492	Assign Claim	2020-10-13 03:51:28.627923	Karen Lopez	…
a626bc58-a7b2-4eff-aca9-2a622cd0c492	Claim Decision	2020-10-14 22:30:14.948970	Karen Lopez	…
a626bc58-a7b2-4eff-aca9-2a622cd0c492	Set Reserve	2020-10-24 11:49:39.053890	Karen Lopez	…
a626bc58-a7b2-4eff-aca9-2a622cd0c492	Payment Sent	2020-10-27 05:46:04.378718	Karen Lopez	…
….	…	…	…	…

Tableau 1 : échantillon de l'ensemble des données

Vous avez d'abord converti la colonne de timestamp en format datetime pour faciliter les calculs chronologiques. Ensuite, les données ont été triées par identifiant de cas et timestamp pour organiser les activités dans l'ordre chronologique. Vous avez compté le nombre total d'activités par cas et calculé les timestamps de début et de fin pour chaque cas. La durée totale de traitement a été calculée en soustrayant le timestamp de début du timestamp de fin, puis convertie en jours et arrondie.

Les colonnes inutiles, comme les timestamps de début et de fin, ainsi que les durées de traitement, ont été supprimées pour alléger le jeu de données. Vous avez ensuite créé une colonne pour l'ordre des activités, puis utilisé un tableau croisé dynamique pour restructurer les données en fonction de l'ordre des activités pour chaque cas. Les colonnes du tableau croisé ont été réorganisées selon un ordre prédéfini, et les colonnes manquantes ont été ajoutées avec des valeurs par défaut.

Enfin, les données originales ont été fusionnées avec le tableau croisé mis à jour, les colonnes non pertinentes ont été supprimées, et les doublons basés sur l'identifiant de cas ont été éliminés pour conserver uniquement les enregistrements uniques.

case_id	…	activities	duration	FNOL	AC	CD	SR	PS	CC
000112d5-9d04-450f-820f-3edfc0626cf9	…	6	31	1	2	3	4	5	6
0001c62c-696c-4251-a604-8d319fc73fac	…	6	30	1	2	3	4	5	6
00048c02-65b5-423b-bf38-139a099a5624	…	6	34	1	2	3	4	5	6
0005b367-5a31-4bcd-82f3-05103f43b462	…	6	36	1	2	3	4	5	6
…	…	…	…	…	…	…	…	…	…

Tableau 2 : jeu de données après transformation

2.2. Etiquetage d'après indicateurs :

Vérification du nombre d'activités :

- Condition : Le nombre d'activités doit être inférieur à 6.
- Action : Si le nombre d'activités est inférieur à 6, des vérifications supplémentaires sont effectuées pour identifier des violations spécifiques du flux de processus ou une fraude potentielle.

Vérification du temps de traitement :

- Condition : Le temps de traitement (en jours) doit être inférieur à 29.
- Action : Les réclamations avec un temps de traitement inférieur à 29 jours sont signalées pour une validation plus approfondie.

Violations du flux de processus, Violation si :

- Non débuté avec FNOL : Le processus ne commence pas par "First Notification of Loss (FNOL)".
- FNOL suivi par Set Reserve : Le flux de processus a incorrectement "First Notification of Loss (FNOL)" suivi directement par "Set Reserve".
- FNOL suivi par Payment Sent : Le flux de processus a incorrectement "First Notification of Loss (FNOL)" suivi directement par "Payment Sent".
- FNOL suivi par Claim Decision : Le flux de processus a incorrectement "First Notification of Loss (FNOL)" suivi directement par "Claim Decision".
- Assign Claim suivi par Set Reserve : Le flux de processus a incorrectement "Assign Claim" suivi directement par "Set Reserve".
- Assign Claim suivi par Payment Sent : Le flux de processus a incorrectement "Assign Claim" suivi directement par "Payment Sent".
- Payment Sent suivi par Set Reserve : Le flux de processus a incorrectement "Payment Sent" suivi directement par "Set Reserve".

3. Model de Prédiction :

3.1. Elimination des attribues :

Éliminer des attributs comme `case_id`, `client`, `agent`, et `adjuster` dans le cadre de la prédiction de fraude ou de violations est souvent justifié pour se concentrer sur les données les plus pertinentes et éviter les problèmes potentiels liés à la confidentialité, la complexité et la performance du modèle.

- **Pertinence de l'Attribut** : les attributs n'apportent pas d'information utile pour la détection de fraude ou de violations, ils peuvent être éliminés. Par exemple, ces attributs peuvent être spécifiques aux individus impliqués et ne pas aider à identifier des anomalies dans les processus.
- **Confidentialité et Sécurité** : Les informations sur les clients, les agents et les ajusteurs peuvent être sensibles ou privées. Les supprimer peut aider à garantir la confidentialité des données, surtout si le modèle est utilisé dans un environnement où la protection des données est cruciale.
- **Réduction du Bruit** : Les attributs non pertinents peuvent introduire du bruit dans le modèle. Si ces attributs ne sont pas directement liés aux violations ou à la fraude, les inclure pourrait diluer l'impact des caractéristiques réellement importantes pour la détection.
- **Simplicité du Modèle** : Un modèle plus simple avec moins de caractéristiques est souvent plus facile à interpréter et à évaluer. Éliminer des attributs superflus peut aider à éviter l'overfitting (surapprentissage) et améliorer la généralisation du modèle.

- **Amélioration de la Performance** : Certains attributs peuvent ne pas contribuer positivement à la performance du modèle. En les supprimant, vous pouvez améliorer la vitesse d'entraînement et la précision du modèle.
- **Évitement de la Colinéarité** : Si les attributs sont fortement corrélés avec d'autres variables ou entre eux, leur présence peut entraîner des problèmes de colinéarité. Éliminer ces attributs peut simplifier le modèle et éviter ces problèmes.

3.2. Encodage des données :

- **Objectif :** Convertir les variables catégorielles en valeurs numériques pour la compatibilité avec les algorithmes de machine learning.
- **Méthode :** Utilisation d'un encodeur de labels pour transformer chaque catégorie en un nombre unique.
- **Variables Encodées :**
  - **`type_of_policy` :** Type de police d'assurance (par exemple, Collision, Complet, etc.)
  - **`car_brand` :** Marque de la voiture (par exemple, Honda, Toyota, etc.)
  - **`car_model` :** Modèle de la voiture (par exemple, Civic, Camry, etc.)
  - **`type_of_accident` :** Type d'accident (par exemple, Collision, Vol, etc.)
  - **`user_type` :** Type d'utilisateur (par exemple, Régulier, Premium, etc.

claim_amount	client_age	type_of_policy	car_brand	car_model	car_year	type_of_accident	user_type
9266	60	1	3	3	2021	0	0
4636	21	1	5	0	2012	0	0
4897	44	1	3	3	2011	1	1
9596	81	0	4	6	2017	3	1
7181	61	0	3	3	2017	2	0

Tableau 3 : jeu de données aprés l'encodage

3.3. Équilibrage des données:

La Distribution Initiale des Classes : Avant toute opération de rééquilibrage, vous avez vérifié la distribution initiale des classes dans votre DataFrame :

- `0` **(valid) :** 29 265 échantillons
- `1` **(fraud) :** 559 échantillons
- `2` **(violation) :** 176 échantillons

Cette distribution montre une forte imbalance entre les classes, avec une prépondérance des classes `valid` par rapport aux classes `fraud` et `violation`.

Sur-échantillonnage des Classes Minoritaires avec SMOTE : Vous avez utilisé **SMOTE (Synthetic Minority Over-sampling Technique)** pour augmenter le nombre des classes minoritaires (`fraud` et `violation`). SMOTE crée des échantillons synthétiques pour les classes minoritaires afin d'équilibrer la distribution des classes :

- `0` **(valid) :** 29 265 échantillons (inchangé)

- `1` **(fraud) :** 8 000 échantillons
- `2` **(violation) :** 8 000 échantillons

Cela a permis d'augmenter les échantillons des classes `fraud` et `violation` à un niveau égal, bien que la classe `valid` soit restée inchangée.

Utilisation de SMOTETomek pour Nettoyer et Rééquilibrer : Vous avez ensuite appliqué **SMOTETomek**, une technique combinée de sur-échantillonnage et de sous-échantillonnage. SMOTETomek combine SMOTE avec Tomek Links (une technique de nettoyage des données) pour corriger les frontières de décision et réduire les bruits dans les données :

- `0` **(valid) :** 29 134 échantillons
- `1` **(fraud) :** 29 194 échantillons
- `2` **(violation) :** 29 201 échantillons

Cette étape a permis d'ajuster la distribution des classes en éliminant certains échantillons pour réduire les erreurs de classification potentielles.

Sous-échantillonnage de la Classe Majoritaire avec RandomUnderSampler : Pour finaliser le rééquilibrage, vous avez utilisé **RandomUnderSampler** pour réduire le nombre d'échantillons de la classe majoritaire (`valid`). Cela permet d'obtenir une distribution plus équilibrée entre les classes :

- `0` **(valid) :** 12 000 échantillons
- `1` **(fraud) :** 8 000 échantillons
- `2` **(violation) :** 8 000 échantillons

Le sous-échantillonnage a réduit le nombre d'échantillons de la classe `valid` pour atteindre un équilibre entre les trois classes.

Cette démarche garantit que le modèle de machine learning sera formé sur un ensemble de données où les classes sont équilibrées, ce qui peut améliorer les performances du modèle et éviter les biais envers la classe majoritaire.

3.4. Comparaison des modèles de classification :

Objectif : Le but de cette analyse est de comparer les performances de différents modèles de classification sur un ensemble de données équilibré d'assurances. Les modèles évalués sont :

Régression Logistique :

- `Logistic Regression`
- `Support Vector Machine (SVM)`
- `Decision Tree Classifier`
- `Random Forest Classifier`
- `K-Nearest Neighbors (KNN)`
- `Naive Bayes`
- `XGBoost`

Méthodologie :

- **Préparation des Données :**

- Les données ont été chargées depuis un fichier CSV et séparées en caractéristiques (X) et cible (y).
- Les données ont été divisées en ensembles d'entraînement et de test (80%/20%) en utilisant une stratification pour maintenir la distribution des classes.

- **Entraînement et Évaluation des Modèles :**
  - Chaque modèle a été entraîné sur l'ensemble d'entraînement.
  - Les prédictions ont été faites sur l'ensemble de test.
  - Les métriques de performance suivantes ont été calculées : Précision, Rappel, Score F1 et Exactitude.

Résultats Les résultats obtenus pour chaque modèle sont résumés dans le tableau ci-dessous

Modèle	Exactitude	Précision	Rappel	Score F1
Régression Logistique	0.775536	0.780581	0.773056	0.774775
SVM	0.428571	0.142857	0.333333	0.200000
Arbre de Décision	1.000000	1.000000	1.000000	1.000000
Forêt Aléatoire	1.000000	1.000000	1.000000	1.000000
KNN	0.819643	0.821463	0.840972	0.821679
Naive Bayes	0.996250	0.995654	0.995625	0.995625
XGBoost	1.000000	1.000000	1.000000	1.000000

Tableau 4 : performance des models de classification

Interprétation des Résultats :

- Arbre de Décision et Forêt Aléatoire : Ces modèles ont obtenu une performance parfaite (exactitude, précision, rappel et score F1 de 1.000). Ils semblent bien adaptés à ce jeu de données équilibré.
- XGBoost : A également montré des performances parfaites, indiquant une grande efficacité dans la classification.
- Naive Bayes : A obtenu des résultats très élevés, mais légèrement inférieurs à ceux de XGBoost et de la Forêt Aléatoire.
- KNN : A également montré de bonnes performances avec un score F1 de 0.821679, ce qui est assez bon, bien que pas aussi élevé que ceux des modèles précédents.
- Régression Logistique : A montré une performance correcte mais pas aussi élevée que les autres modèles.

- SVM : A eu des performances faibles avec une précision et un rappel bien inférieur à ceux des autres modèles, indiquant qu'il pourrait ne pas être bien adapté à ce problème spécifique.

Problèmes de Fitting (over - under)

- Decision Tree : Les arbres de décision peuvent être sensibles à des ensembles de données très déséquilibrés. Ils peuvent se surajuster aux données d'entraînement et ne pas bien généraliser aux nouvelles données, surtout si les données sont très déséquilibrées ou si les arbres sont trop profonds.
- Random Forest : Bien que les forêts aléatoires soient généralement robustes, elles peuvent rencontrer des problèmes de fitting si le nombre d'arbres est insuffisant ou si les arbres sont trop profonds. De plus, elles peuvent ne pas bien gérer des déséquilibres extrêmes dans les classes.
- Naive Bayes : Le modèle Naive Bayes repose sur des hypothèses d'indépendance entre les caractéristiques. Si cette hypothèse est violée ou si les données contiennent beaucoup de bruit, le modèle peut ne pas bien se comporter. Il peut aussi être moins performant avec des caractéristiques très corrélées ou peu représentatives.
- XGBoost : Bien que XGBoost soit souvent performant, il peut souffrir de problèmes de fitting si les hyperparamètres ne sont pas correctement réglés ou si le modèle est surajusté aux données d'entraînement. Le réglage des hyperparamètres et la gestion du surajustement sont cruciaux pour obtenir de bonnes performances.

Modèle Optimal

- K-Nearest Neighbors (KNN) : Le modèle KNN semble être le plus optimal pour votre cas d'utilisation. Il offre de bonnes performances avec une précision et un score F1 élevés. KNN est un modèle simple mais efficace, surtout si les données sont bien équilibrées et si les caractéristiques sont représentatives.

3.5. KNN hyperparamètres Tuning:

Pourquoi avons-nous utilisé l'optimisation des hyperparamètres ?

L'optimisation des hyperparamètres permet de trouver la meilleure combinaison de paramètres pour le modèle, ce qui peut considérablement améliorer sa performance. Pour le modèle KNN, les principaux hyperparamètres à optimiser sont :

- `n_neighbors` : Le nombre de voisins à considérer pour faire une prédiction. Un nombre trop faible peut entraîner un modèle surajusté (overfitting), tandis qu'un nombre trop élevé peut rendre le modèle sous-ajusté (underfitting).
- `weights` : La fonction de pondération des voisins. Les options sont 'uniform' (tous les voisins ont le même poids) et 'distance' (les voisins plus proches ont un poids plus élevé).
- `algorithm` : L'algorithme utilisé pour calculer les voisins. Les options incluent 'auto', 'ball_tree', 'kd_tree', et 'brute'.
- `p` : Le paramètre pour la distance de Minkowski. Les valeurs possibles sont 1 (distance de Manhattan) et 2 (distance Euclidienne).

Processus de tuning :

- Définition de la grille des paramètres : Une grille de paramètres a été définie pour explorer différentes combinaisons possibles des hyperparamètres. Cela inclut une gamme de valeurs pour `n_neighbors`, des options pour `weights`, `algorithm`, et `p`.
- Grid Search avec Cross-Validation : Le processus de Grid Search avec une validation croisée (cross-validation) a été utilisé pour évaluer chaque combinaison de paramètres. Cela implique la division des données d'entraînement en plusieurs sous-ensembles pour tester le modèle sur des portions différentes à chaque itération, afin de réduire le risque de surajustement.
- Évaluation des modèles : Chaque modèle a été évalué en termes de score de validation croisée, et les meilleures performances ont été retenues.

Paramètres optimaux trouvés :

- `algorithm` : 'auto'
- `n_neighbors` : 3
- `p` : 1
- `weights` : 'distance'

3.6. KNN Performance :

Impact de l'optimisation des hyperparamètres : L'optimisation des hyperparamètres a conduit à la sélection des meilleurs paramètres pour le modèle KNN.

Illustré par les résultats suivants :

- Score de validation croisée optimal : 0.8890
- Précision sur le jeu de test : 0.9048

Métriques de performance du meilleur modèle :

- Matériel de confusion:

Classe	Précision	Rappel	Score F1	Support
0	0.98	0.80	0.88	2400
1	0.85	0.98	0.91	1600
2	0.88	0.99	0.93	1600
Accuracy	0.90	-	-	5600
Macro Average	0.90	0.92	0.91	5600

Tableau 5 : performance de KNN

- Rapport de classification:

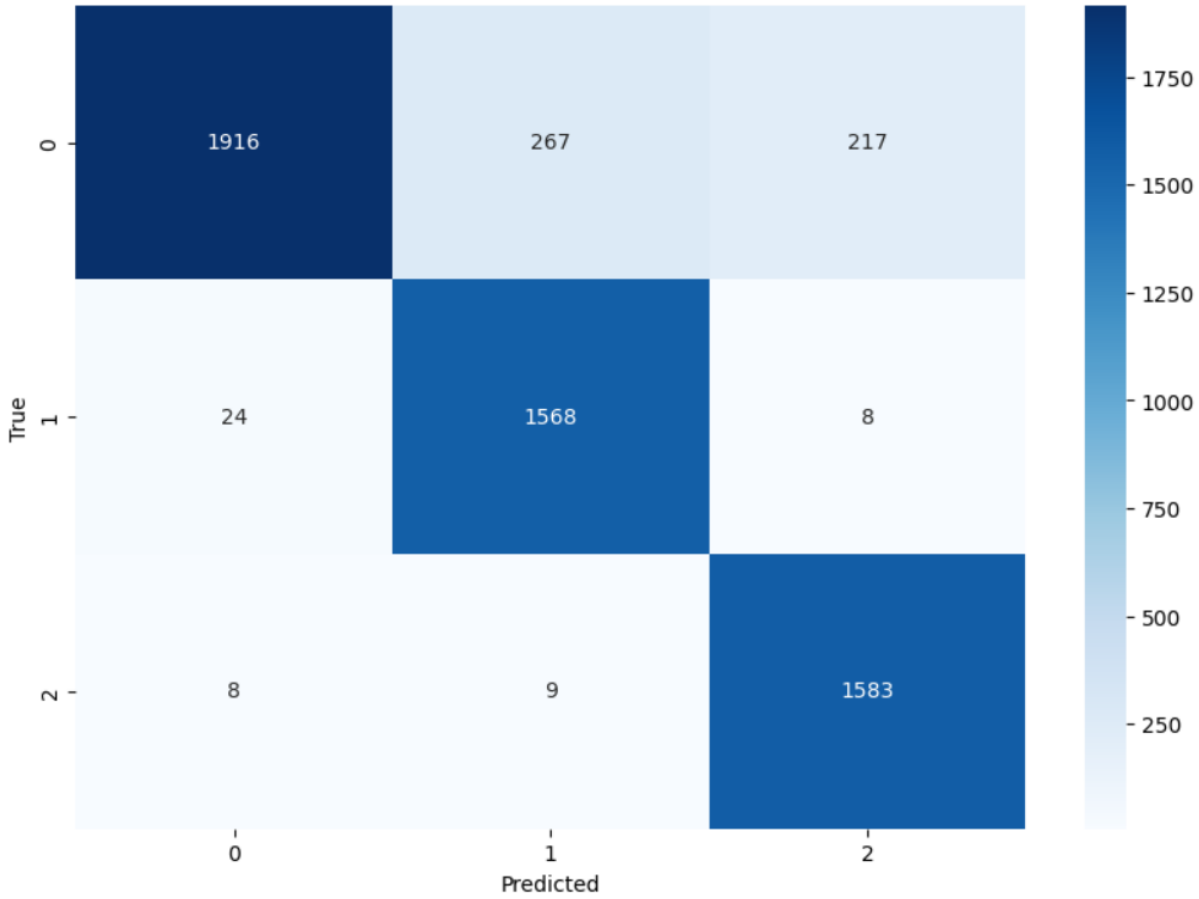


Figure 21 : modèle KNN rapport de performance

Analyse des résultats :

- Précision globale élevée : Le modèle atteint une précision de 90,48 % sur les données de test, ce qui indique une bonne capacité à faire des prédictions correctes.
- Amélioration des métriques pour chaque classe : Les métriques de précision, de rappel et de F1-score montrent des améliorations significatives, particulièrement pour la classe 1 (fraud), avec une précision et un rappel élevé.

L'optimisation des hyperparamètres a permis d'améliorer la performance du modèle KNN en choisissant les paramètres les plus appropriés. Cette approche a permis de maximiser la précision et la fiabilité du modèle, ce qui est crucial pour les applications pratiques telles que la détection de fraude dans les réclamations d'assurance.

Conclusion

Ce Chapitre décrit les étapes essentielles de l'implémentation du Process Mining, allant du chargement et de l'analyse des données à la préparation et à l'étiquetage nécessaires pour la modélisation. Il met en lumière comment l'ajustement des hyperparamètres du modèle KNN améliore considérablement ses performances, illustrant ainsi l'importance de l'optimisation pour obtenir des résultats analytiques précis et fiables.

Conclusion et perspectives

Ce rapport a présenté une approche complète pour l'implémentation de techniques de Process Mining et de modélisation prédictive afin d'analyser les sinistres d'assurance. Nous avons exploré le processus de chargement et d'analyse des données, essentiel pour comprendre les anomalies et les processus internes. La transformation des données, y compris la réduction de la dimension et l'étiquetage basé sur des indicateurs, a permis de préparer des jeux de données robustes pour l'analyse.

La comparaison des modèles de classification a révélé que le réglage des hyperparamètres du KNN a significativement amélioré la performance du modèle, avec une précision notable et une meilleure capacité à détecter les anomalies. Les résultats montrent que des techniques telles que le tuning des hyperparamètres sont cruciales pour optimiser les performances des modèles de machine learning.

Pour l'avenir, il serait avantageux d'explorer l'intégration de techniques de machine learning plus avancées, telles que les réseaux neuronaux profonds, pour capturer des patterns plus complexes dans les données. L'expansion des ensembles de données et l'inclusion de nouvelles variables pourraient également enrichir les analyses et fournir des insights encore plus précis. Enfin, la mise en œuvre de solutions d'analyse avancées et la continuité de l'amélioration des modèles prédictifs contribueront à une gestion plus efficace des sinistres et à une prise de décision plus éclairée.

Webographie

Tutoriel Process Mining :

edx RWTHx: A Hands-On Introduction to Process Mining by *Wil van der Aalst*

https://fluxicon.com/book/read/

Livre :

Process Mining Data Science in Action Second Edition

Celonis Documentation :

https://docs.celonis.com/en/getting-started.html

PyCelonis

Ressource :

https://processmining.org/home.html

Projet :

https://github.com/Abdelhakim-gh/PFA-Process-Mining-Fraud-detection