REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : SIOD 3/M2/2021

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : **Système d'Information Optimisation et Décision (SIOD)**

Bank fraud detection using sequential pattern mining

Par :

BERKANE MABROUKA

Soutenu le 07 juillet 2021 Devant le jury composé de :

…	…	Président
Djeffal Abdelhamid	Professeur	Rapporteur
…	…	Examinateur

Année universitaire 2020-2021

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introuduction

1.1 Introuduction

Toutes les activités commerciales, notamment le secteur bancaire aujourd'hui, collectent et stockent d'énormes quantités de données, qui ne cessent d'augmenter de jour en jour. Cette énorme quantité de données est considérée comme mines d'informations, elles cachent des connaissances importantes face au marchée concurrence, mais ils restent un peu exploités. Pour combler ce besoin d'une nouvelle industrie ou pour répondre à diverses fraudes et complots, le data mining est un ensemble d'algorithmes issus de diverses disciplines scientifiques telles que les statistiques, l'intelligence artificielle et les bases de données afin de construire différents modèles basés sur des données stockées, c'est-à-dire trouver des modèles intéressants selon les critères spécifiés dans Commencer et extraire autant de connaissances de l'entreprise que possible.

La fouille des motifs séquentiels est peut-être le domaine de recherche le plus populaire parmi eux Consiste à trouver des sous-séquences apparaissent fréquemment dans un ensemble de commandes.

Le but de ce travail est de connaître et découvrir l'existence de la fraude bancaire à travers une série de comportements formés par les clients dans un certain laps de temps. Pour cette raison, le mémo commence par le premier chapitre, qui parle de plusieurs concepts liés à la banque la fraude et ses types et moyens qui nous permettent de la détecter. Ainsi que le deuxième chapitre, qui introduit le concept de data mining en donnant quelques définitions de ce terme et les motivations qui ont conduit à l'émergence de ce type de système, Après avoir expliqué toutes les tâches de l'exploration de données, sachant que chacune d'entre elles est représenté une méthode d'extraction de connaissances.

Enfin, nous avons présenté la méthode dont nous avons besoin pour atteindre le résultat attendu, qui s'appelle la fouille des motifs séquentiels celle que nous utiliserons pour créer une application pour détecter la fraude bancaire par le comportement des clients.

Chapitre 2

La fraude bancaire

2.1 Introduction

Les banques sont d'une grande importance pour toute économie, et l'une des composantes les plus importantes de l'État, et elle n'a pas gagné cette importance du vide, mais des rôles et des activités qu'elle exerce car elle permet à cette économie de se développer et de progresser. Mais récemment, la plupart des systèmes bancaires ont été exposés au piratage et au vol, en particulier à l'époque de la Corona, en raison du manque de moyens suffisants pour dissuader les fraudeurs et les voleurs, et dans cette partie, nous nous familiariserons avec la fraude bancaire, son les types et les effets, en particulier les moyens de la combattre et d'essayer de l'éliminer. Alors, qu'est-ce que la fraude bancaire ? Quels sont ses types ? Quels sont les moyens de le détecter et d'en protéger les clients ?

2.2 Définition de la fraude bancaire

La fraude bancaire peut être définie comme un acte contraire à l'éthique et / ou criminel par un individu ou une organisation pour tenter illégalement de posséder ou de recevoir de l'argent d'une banque ou d'une institution financière.[2]

En général, la fraude bancaire peut impliquer toute action délibérée visant à frauder une institution financière. Il peut s'agir d'une action intentionnelle visant à recevoir des actifs, de l'argent, des valeurs mobilières, des crédits ou des biens d'une institution financière en utilisant des informations factices ou fausses. La loi donne une définition assez large de la fraude bancaire, et plusieurs aspects de cette infraction doivent être pris en compte.[32]

La surcharge d'informations est due à sa diversité croissante et continue et à la nécessité de les convertir en données significatives dans plusieurs domaines divers (santé, éducation, commerce, découverte scientifique ...) et c'est ce qui a conduit les chercheurs à s'intéresser au data mining.

2.3 Les Types de la fraude bancaire

La fraude dans les banques revêt de nombreuses formes, elle peut être interne, c'est-à-dire qu'elle est commise par des employés de la banque elle-même ou en dehors de nous,

commise par des clients, des personnes ou des institutions étrangères à la banque. Parmi les plus célèbres de ses types, on trouve les trois formes suivantes :

2.3.1 La Fraude de carte de crédit

Il s'agit d'une tentative par une personne ou une organisation de voler ou d'utiliser une carte de crédit ou de débit sans autorisation appropriée pour un gain financier Illégalement. L'une des formes les plus courantes de fraude par carte de crédit survient après le vol ou la perte d'une carte de débit ou de crédit. Dans ces cas, une partie non autorisée peut accéder aux numéros de carte de crédit ou de débit d'une autre personne (bien que ne pas connaître le code PIN rendrait pratiquement impossible le retrait d'espèces à un guichet automatique).

2.3.2 La fraude par chèque

La fraude par chèque est aujourd'hui l'un des plus grands défis auxquels sont confrontées les entreprises et les institutions financières. Avec les progrès de la technologie informatique, il est devenu de plus en plus facile pour les criminels, indépendamment ou en gangs organisés, de manipuler les chèques de manière à tromper des victimes innocentes qui en attendent pour leur argent.[5]

La fraude par chèque fait référence à l'utilisation illégale d'un chèque pour un gain financier non autorisé et se fait également par le biais de la publication assistée par ordinateur et de la copie pour créer ou copier un document financier réel, ce qui implique de supprimer tout ou partie des informations et de les manipuler au profit du criminel.[5]

Les victimes comprennent les institutions financières, les entreprises qui acceptent et émettent des chèques et le consommateur. Dans la plupart des cas, ces crimes commencent par le vol d'un document financier.[5]

La fraude par chèque peut se produire de plusieurs manières et voici quelques :

- Déposer un chèque sur un compte sans l'autorisation appropriée.
- Modifier un chèque en modifiant les informations bancaires, telles que les numéros de compte.
- Utiliser un chèque pour effectuer un paiement en sachant qu'il n'y a pas suffisamment de fonds sur le compte.

- Modification du montant du paiement sur un chèque.
- Utilisation de chèques pour les fausses factures. [2]

2.3.3 Le vol d'identité

Le vol d'identité consiste à obtenir des données personnelles sensibles d'une personne sans son consentement. En termes plus simples, lorsqu'une personne vole vos informations personnelles (par exemple, votre numéro de sécurité sociale, votre numéro d'identification, les données de votre carte bancaire ou votre date de naissance), cela est considéré comme un vol identification.

En règle générale, un voleur utilise ces informations personnelles pour effectuer des achats, obtenir un crédit ou faire autre chose en votre nom.[1]

Nous mentionnons certains des endroits où les criminels recherchent des informations pour voler l'identité de quelqu'un :

- Paniers ou bacs à litière
- l'Internet
- Boîtes aux lettres (ouverture du courrier)
- Téléphones et télécopieurs. [4]

Pour éviter le vol d'identité, nous vous proposons quelques conseils pour vous aider à protéger vos informations.

- Protégez vos informations personnelles.
- Ne partagez pas d'informations personnelles par téléphone, e-mail ou Internet à moins que vous n'ayez déjà commencé à appeler et que vous sachiez à qui vous avez affaire.
- Ne partagez jamais votre NIP ou vos mots de passe avec qui que ce soit.
- Tous les documents contenant des informations personnelles doivent être éliminés de manière sûre, par exemple par déchiquetage.
- Créez des mots de passe forts et mettez-les à jour fréquemment.
- Vérifiez votre rapport de solvabilité au moins une fois par an. [4]

2.3.4 La fraude hypothécaire

Lorsqu'un client a besoin d'argent ou d'un prêt, la banque doit lui faire signer un document de crédit contenant une garantie ou une hypothèque pour garantir le remboursement du crédit en cas de difficultés à rembourser le client.

De nombreux clients accordent à la banque des prêts hypothécaires frauduleux ou gonflés qui ne lui permettent pas de recouvrer son crédit. Ce type de fraude donne aux banques une grande partie de leurs pertes. Les indicateurs utilisés pour détecter ce type de fraude sont les informations personnelles et professionnelles des clients en plus de l'hypothèque consentie.[15]

2.3.5 La fraude à l'assurance

Il y a fraude à l'assurance lorsqu'une personne fournit de fausses informations à une compagnie d'assurance afin d'obtenir quelque chose de valeur qu'elle n'aurait pas obtenu si la vérité avait été dite [3] .C'est également un acte illégal de la part de l'acheteur ou du vendeur de contracter une assurance.

La fraude à l'assurance de l'émetteur comprend les politiques de vente de sociétés inexistantes, le défaut de fournir des primes et des politiques contradictoires pour générer plus de commissions. Pendant ce temps, la fraude à l'acheteur peut consister en des réclamations exagérées, de faux antécédents médicaux, des politiques différées, une fraude légale, de faux décès ou enlèvement et un meurtre. [15]

2.3.5.1 Exemple de fraude à l'assurance

Le propriétaire d'un véhicule peut tenter de réduire les coûts des primes d'assurance en utilisant une fausse immatriculation. Si le propriétaire du véhicule habite dans une région où les primes sont élevées en raison d'un vol de voiture récurrent dans le quartier ou pour d'autres raisons, le propriétaire peut essayer d'immatriculer le véhicule dans une autre région pour réduire ses primes.

Les travaux de réparation sur un véhicule peuvent également devenir une source de fraude à l'assurance. Par exemple, un atelier de réparation qui attend le paiement de l'assureur peut facturer des travaux importants, mais utiliser ensuite des produits de remplacement bon marché ou même de faux. Ils peuvent également surcharger l'assureur

en surestimant l'étendue des réparations nécessaires.[4]

2.4 L'impact mondial de la fraude bancaire

L'impact de la fraude dans le secteur bancaire étranger est ressenti par tout le monde, sinon en tant que client, alors, en tant que citoyen

La fraude a de nombreux effets négatifs sur la société, car cette industrie constitue une position vitale dans notre société et une partie importante de l'économie, en particulier le secteur bancaire qui est imprégné de fraude et Son succès ou son échec est une raison très importante pour déterminer le succès de la société.

La fraude est une cause majeure de faillite bancaire. En effet, le nombre de fraudes qui se produisent dans les banques étrangères est en constante augmentation et cela a provoqué une vague d'inquiétude sévère car cela affecte totalement la mauvaise performance des banques. Les sommes prélevées sur les coffres bancaires sont considérées comme une perte pour lui car elles ne génèrent aucun revenu pour la banque. Il en résulte plutôt un problème difficile pour la banque, qui est le manque de liquidités.

Les points suivants sont considérés parmi les effets les plus importants de la fraude sur les banques :

- L'impact de la fraude sur des entités telles que les banques, et le coût économique de la fraude peuvent être énormes en termes de potentiel d'agitation et de confiance dans le système bancaire et peuvent nuire à l'intégrité et à la stabilité de l'économie.
- Cela pourrait conduire à l'effondrement des banques, miner le rôle de supervision de la banque centrale et même créer des troubles sociaux, du mécontentement et des troubles politiques.

L'exposition des banques à la fraude s'est accrue dans les technologies récentes.[46]

2.5 Méthodes de détection

S'il n'est pas possible d'éliminer complètement la fraude bancaire. Il existe des moyens de l'empêcher et de réduire la probabilité que cela se produise .Il existe deux méthodes plus efficaces qui sont les suivantes :

2.5.1 Les méthodes manuelles

2.5.1.1 Empreinte digitale

L'empreinte digitale représente une technique de détection de fraude importante, en particulier dans le domaine bancaire, car elle contient une caractéristique qui est le fait que l'empreinte digitale diffère d'une personne à l'autre, ce qui fournit un moyen d'identification unique pour chaque personne dans le monde. Elle est utilisée dans le domaine des services bancaires en ligne et en raison des difficultés que rencontrent les banques pour identifier les appareils sources en fonction de l'adresse IP uniquement parce qu'elle peut évoluer dans le temps.

Et pour cela, une solution appropriée a été proposée à savoir que le dispositif d'accès est déterminé par un composant qui doit être téléchargé et installé dans l'appareil du client. Ce composant crée une empreinte digitale du dispositif d'accès et l'envoie au site Web de la banque dans le cadre des données de chaque transaction.

Ensuite, l'empreinte digitale est calculée en appliquant une fonction de cryptage aux informations matérielles et logicielles, telles que le processeur, les numéros de série du système d'exploitation, l'adresse MAC et certains détails de configuration.

Pour implémenter le composant, nous avons besoin de trois exigences de base, qui sont les suivantes :

- Génère une empreinte digitale différente pour chaque périphérique d'accès différent.
- Fournit un certain caractère aléatoire tout en générant des empreintes digitales en raison de la difficulté d'usurpation d'identité par d'autres appareils ;
- Il notifie la nouvelle empreinte digitale chaque fois que la configuration de l'appareil change.

En fait, le système proposé est basé sur le composant qui est réellement utilisé par le système bancaire en ligne actuel.[45]

2.5.1.2 La signature

La vérification de la signature est la méthode la plus courante utilisée par les banques et les institutions financières ainsi que leurs clients pour authentifier l'identité d'une personne mais c'est une méthode fatigante qui demande de la pratique et de la diligence et qui nécessite également une signature valide pour la comparaison.

Pour mener à bien ce processus, les back-offices des institutions financières sont les meilleurs endroits pour vérifier les signatures.

Cependant, ils sont confrontés à de nombreux défis, et nous mentionnons les deux défis les plus importants :

- Le volume de chèques présentés au paiement est trop important pour que la banque envisage de vérifier la validité des signatures sur chaque article proposé au paiement.
- La prolifération de périphériques de numérisation bon marché fait de la copie de signatures non autorisée un moyen très simple de créer de faux éléments

Présenter des cartes d'identité comme les permis de conduire, les cartes de sécurité sociale, etc. n'est pas un moyen efficace d'authentifier un client. Il existe des sites Web sur Internet qui fournissent de fausses cartes d'identité à tout étudiant sans exiger aucune preuve de la véritable identité de cette personne.
La présentation de l'identité physique, bien qu'elle soit encore utilisée, est le moyen le moins efficace d'authentifier un individu.[23]

2.5.2 Les Méthodes automatiques

2.5.2.1 Méthodes basées sur des propriétés statiques

Une analyse expérimentale a été menée sur un ensemble de données de transaction du monde réel afin de révéler et d'accéder que la plupart des fraudes ont certaines caractéristiques comportementales, qui sont les suivantes :

- Un grand nombre de comptes différents auxquels un seul fraudeur accède.
- Transactions impliquant de petites valeurs dans de nombreux comptes.
- Plus de transactions de paiement que d'habitude sur un seul compte.
- Augmentation du nombre d'échecs de mot de passe avant que la fraude ne se produise.

Alors que les deux derniers traits peuvent être détectés par une analyse différentielle à l'aide de fonctionnalités locales, les deux premiers traits nécessitent des informations sur des attaques similaires dans d'autres comptes. [45]

2.5.2.2 Les méthodes d'apprentissage

1.Méthodes basées sur les caractéristiques des clients

- **1.K voisins les plus proches (KNN) :**

  La méthode du k-plus proche voisin (K-Nearest Neighbor (KNN)) est une méthode supervisée et c'est un algorithme simple qui stocke toutes les instances disponibles ; puis il classe toutes les nouvelles instances en fonction d'une mesure de similarité. L'algorithme KNN est un exemple d'apprenant basé sur une instance. Dans la méthode de classification du voisin le plus proche, chaque nouvelle instance est comparée aux instances existantes en utilisant une métrique de distance, et l'instance existante la plus proche est utilisée pour affecter la classe à la nouvelle [34]. Parfois, plus d'un voisin le plus proche est utilisé et la classe majoritaire des K voisins les plus proches est affectés à la nouvelle instance.

  1. **Les avantages :**
     - La qualité de la méthode s'améliore en introduisant de nouvelles données sans nécessiter la reconstruction d'un modèle. Ce qui représente une différence majeure avec des méthodes telles que les arbres de décision et les réseaux de neurones.
     - Facile à mettre en œuvre.
     - La clarté des résultats : la classe attribuée à un objet peut être expliquée en exhibant les plus proches voisins qui ont amené à ce choix.
     - La méthode peut s'appliquer à tout type de données même les données complexes tels que des informations géographiques, des textes, des images et du son. C'est parfois un critère de choix de la méthode PPV car les autres méthodes traitent difficilement les données complexes. Nous pouvons noter également, que la méthode est robuste au bruit. [13] [27]
  2. **Les inconvénients**
     - temps de classification : la méthode ne nécessite pas d'apprentissage ce qui implique que tous les calculs sont effectués lors de la classification. Contrairement aux autres méthodes qui nécessite un apprentissage (éventuellement long) mais qui sont rapides en classification.
     - méthode donnera de mauvais résultats Si le nombre d'attributs pertinents

est faible relativement au nombre total d'attributs, car la proximité sur les attributs pertinents sera noyée par les distances sur les attributs non pertinents.

- Les performances de la méthode dépendent du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins. [13] [27]

2.Support Vector Machine (SVM) :

La méthode SVM est une méthode introduite par Vladimir Fabnik au début des années 1990. Il s'agit d'un ensemble d'algorithmes d'apprentissage automatique qui résolvent les problèmes de classification et visent à trouver le meilleur séparateur pour séparer les objets.

A partir de chaque catégorie [43] en utilisant la méthode linéaire pour que la distance entre les groupes de classes différentes représente le maximum entre eux.

Il est appliqué dans de nombreux domaines, y compris la détection de visage, la classification de texte et d'hypertexte, la classification d'images et la bio-informatique. [26]

Il est également connu pour ses solides garanties théoriques, sa grande flexibilité et sa facilité d'utilisation.[14]

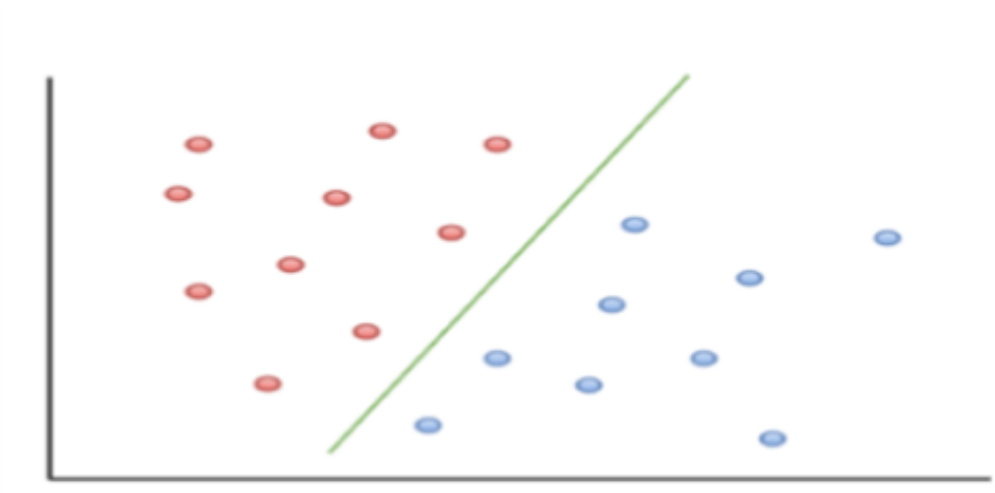

FIGURE 2.1: Machine à vecteurs de support

1. **Les avantages de la méthode SVM :**
   - Utilisation des données qualitatives et quantitatives ;

- Grande précision de prédiction ;
- Meilleur fonctionnement sur les bases de données de taille réduite.[26]

2. **Les inconvénients**
   - Temps d'entraînement long avec les bases volumineuses ;
   - Moins efficace sur les jeux de données contenant de bruits.[26]

2.Les Méthodes basées sur le comportement

- **Les Motifs séquentiels (SPM) :**

  L'exploration de modèles consiste à découvrir des modèles intéressants, utiles et inattendus dans des bases de données. Différents types de modèles peuvent être découverts dans des bases de données, tels que des ensembles d'éléments fréquents, des associations, des sous-graphiques, des règles séquentielles et des modèles périodiques.

  La tâche d'exploration de modèles séquentiels est une tâche d'exploration de données spécialisée pour analyser des données séquentielles, pour découvrir des modèles séquentiels. Plus précisément, il consiste à découvrir des sous-séquences intéressantes dans un ensemble de séquences, où l'intérêt d'une sous-séquence peut être mesuré en fonction de divers critères tels que sa fréquence d'occurrence, sa longueur et son bénéfice. L'exploration séquentielle de modèles a de nombreuses applications réelles en raison du fait que les données sont naturellement codées sous forme de séquences de symboles dans de nombreux domaines tels que la bio-informatique, l'apprentissage en ligne, la détection de fraude bancaire, l'analyse du panier de marché, les textes et l'analyse des flux de clics de pages Web.[44]

2.6 Conclusion

dans ce chapitre, nous avons beaucoup appris sur la fraude bancaire et comment nous pouvons la détecter, la combattre en utilisant les techniques d'exploration de données, l'association, le clustering, la prévision et la classification pour analyser les données clients afin d'identifier les modèles qui peuvent conduire à des fraudes.

Dans le prochain chapitre, nous verrons l'une des méthodes les plus utiles pour la détection de la fraude bancaire.

Chapitre 3

La fouille de motifs séquentiels

3.1 Introduction

Après les cas de vol qui se produisent pour beaucoup de gens, en particulier dans les banques, la fraude bancaire est devenue un gros problème dans nos vies et les scientifiques tentent de créer des méthodes de détection. De nombreuses méthodes ont été développées pour détecter les fraudes telles que l'exploration séquentielle de modèles (SPM), Support Vector Machine (SVM) et K-Nearest Neighbor (KNN), car elles ont été utilisées pour détecter des activités anormales et pour la détection de fraude dans de nombreux domaines, tels que comme le blanchiment d'argent, la fraude par carte de crédit, la fraude par chèque.

Dans ce chapitre, je parlerai de la façon dont nous utilisons la fouille de données basée sur le comportement séquentiel des clients pour détecter la fraude et la combattre.

3.2 La fouille de données

3.2.1 Définition

Les définitions de la fouille de données ne font parfois pas la différence entre la fouille de données qui est des données et le KDD ou la découverte de connaissances à partir de données qui peuvent être traduites par l'extraction de connaissances à partir de données. Nous prenons les deux définitions suivantes :

- **Fayyad :** "l'extraction de connaissances à partir de données est un processus non trivial pour identifier des inconnus, valides et potentiellement utilisables dans des bases de données".
- **Frawley :** "extraction non triviale d'informations implicites, auparavant inconnues et potentiellement utiles à partir des données".

Selon les deux définitions précédentes, le domaine est ouvert aux techniques et applications de DM. . Nous citons la classification, la régression, le clustering, les règles d'association, etc.[19]

3.2.2 Evolution de la fouille de données

La fouille de données est une évolution naturelle de l'exploitation des données par des humains utilisant des ordinateurs. Cette évolution peut être résumée dans les points suivants :

a) **Début de l'informatique (années 1940) :** utilisation de l'informatique pour les besoins de calcul, Traitement statistique et analyse de données : début de DM.

b) **Fin des années 80 :** utilisation du contenu des bases de données pour rechercher association de règles : utilisation du terme exploration de base de données.

c) **1989 :** premier atelier sur la découverte de connaissances - proposition du terme découverte de connaissances par Gregory Piatetsky-Shapiro.

d) **1995 :** première conférence sur l'exploration de données. De plus, le DM a été influencé par L'explosion du volume de données produites et stockées et La maturité des outils de reportent de données et l'évolution des besoins des utilisateurs (de la gestion des données à la prise de décision).

e) L'évolution de la relation client : vers le profilage client et la production orientée client.[19]

3.2.3 Les types de données qui sont appliqués par la fouille de données

La fouille de données peut s'appliquer à tous les types de données. On rappelle quelques exemples de types de données auxquels peut s'appliquer la fouille de données sont [17]

3.2.3.1 "Flat file" les fichiers plats :

Sont actuellement la source de donnée le plus commune pour les algorithmes du fouille de donnée, ils sont des fichiers en format texte ou binaire, contenant un enregistrement par ligne, avec des champs séparés par des délimiteurs, tels que les virgules ou les tabulations. Dans ce type de fichiers, les données peuvent être des transactions, des séries temporelles etc.

3.2.3.2 Base de données relationnelle

Est une base de données consistant dans des tableaux séparés, avec des liaisons explicitement dénies et dont les éléments peuvent être combinés sélectivement comme des résultats à des interrogations. Chaque tableau contient des colonnes (correspondantes à des tuples) et des lignes (correspondantes à des attributs), La fouille de donnée peut profiter du SQL pour la sélection, la transformation et la consolidation.

3.2.3.3 Les entrepôts de données (Data Warehouse)

Est un support de données dans laquelle est centralisé un volume important de données consolidées à partir des différentes sources de données (souvent hétérogènes), par exemple Si le directeur de l'entreprise veut accéder aux données de tous les magasins pour prendre des décisions stratégiques, il serait plus approprié si toutes les données étaient stockées dans un seul emplacement avec une structure homogène qui permet l'analyse interactive des données. Autrement dit, les données de différents magasins peuvent être chargées, nettoyées, transformées et intégrées ensemble. Pour faciliter la prise de décisions et les vues multidimensionnelles.

3.2.3.4 Base de données transactionnelle

Est un ensemble d'enregistrements représentant des transactions, Une transaction contient un identifient unique (transaction ID) et une liste d'items composant la transaction.[28]

3.2.3.5 Bases de données orientées objet et relationnelle objet

Il s'agit d'un type spécial de base de données (ou base de données relationnelle) où les données sont des objets.[37]

3.2.3.6 Les bases de données multimédia

Comportent des documents sonores, des vidéos, des images et des médias en textes et audio. Elles peuvent être stockées sur des bases de données orientées objets ou objets relationnelles ou simplement sur un fichier système. Le multimédia est caractérisé par sa haute dimension ce qui rend le datamining sur ce type de données très difficile. [20]

3.2.4 Les tâches de fouille de données

Tâches d'exploration de données De nombreux problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en fonction des six tâches suivantes :

3.2.4.1 La classification

La classification consiste à examiner les caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini.

- Exemples :
  - attribuer ou non un prêt à un client, établir un diagnostic,
  - accepter ou refuser un retrait dans un distributeur,
  - attribuer un sujet principal à un article de presse. [24]

3.2.4.2 L'estimation

contrairement à la classification, le résultat d'une estimation fournit une variable continue. Ceci est obtenu par une ou plusieurs fonctions combinant les données d'entrée. Le résultat d'une estimation permet de faire des classifications à l'aide d'une échelle.

- Exemples :

  Noter un candidat à un prêt : peut être utilisée pour attribuer un prêt (classification), par exemple, en fixant un seuil d'attribution, estimer les revenus d'un client.[24]

3.2.4.3 La prédiction

La prédiction est comme la classification et l'estimation mais sur une échelle de temps différente. Comme les tâches précédentes, il s'inspire du passé et du présent, mais son résultat se situe dans un futur généralement spécifié.

- Exemples :

  prédire les valeurs futures d'actions, prédire au vu de leurs actions passées les départs de clients. [24]

3.2.4.4 La segmentation

La segmentation consiste à segmenter une population hétérogène en sous-populations homogènes. Contrairement à la classification, les sous-populations ne sont pas préétablies.

3.2.4.5 La Description

C'est souvent l'une des premières tâches requises d'un outil d'exploration de données. Il est invité à décrire les données d'une base de données complexe. Cela conduit souvent à une exploitation supplémentaire afin de fournir des explications.

3.2.4.6 L'optimisation

Pour résoudre de nombreux problèmes, il est courant que chaque solution potentielle intègre une fonction d'évaluation. Le but de l'optimisation est de maximiser ou de minimiser cette fonction. Certains spécialistes considèrent que ce type de problème ne concerne pas la fouille de données.[18]

3.2.5 Les étapes du processus de la fouille de données

Le KDD (ou ECD en français) est un "processus non trivial de définition de structures Inconnu, valide et exploitable dans les bases de données **(Fayyad*, 1996)**. Ce processus est divisé en sept étapes suivantes :

3.2.5.1 Collecte de données

combinaison de plusieurs sources de données, souvent hétérogènes, dans une base de données .[48] [39]

3.2.5.2 Nettoyage des données

Normalisation des données, élimination du bruit (attributs avec des valeurs invalides et attributs sans valeurs).[48] [39]

3.2.5.3 Sélection des données

Sélectionnez les attributs utiles de la base de données pour une tâche particulière d'exploration de données .[22]

3.2.5.4 Transformation de données

Processus de transformation des structures d'attributs pour qu'elles soient adéquates pour la procédure d'extraction d'informations .[41]

3.2.5.5 Extraction d'informations (Data mining)

L'application de certains algorithmes de la fouille de données sur les données produites par l'étape précédente (Knowledge Discovery in Databases, ou KDD).[39] [22]

3.2.5.6 Visualisation des données

l'utilisation de techniques de visualisation (histogramme, camembert, arbre, visualisation 3D) pour l'exploration interactive des données (découverte de modèles de données).[41] [39]

3.2.5.7 Evaluation des modèles

l'identification de modèles strictement intéressants en se basant sur des mesures données).[48]

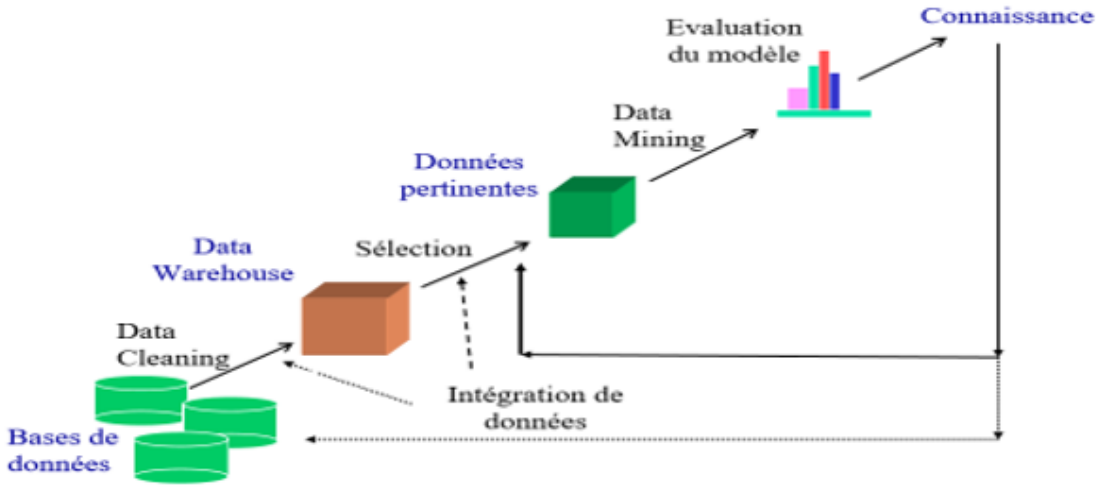

FIGURE 3.1: les Techniques analyse des sentiments [10]

3.2.6 Domaines d'application

L'objectif principal de la fouille de données reste d'extraire des connaissances ciblées qui peuvent être appliquées. C'est pourquoi de nombreux chercheurs et intéressés cherchent à l'appliquer dans tous les domaines de la vie.

Dans ce qui suit, nous mentionnerons certains des principaux domaines dans lesquels la fouille de données a été utilisée :

3.2.6.1 Domaine des assurances

L'une des priorités les plus importantes des compagnies d'assurance est de fidéliser leurs clients tout en augmentant leurs profits et c'est la principale raison de leur intérêt pour l'utilisation des méthodes de la fouille de données. Et c'est à travers :

a) Analyse des risques pour les clients (caractérisation des clients à risque, etc.).

b) Comment traiter les demandes (diagnostiquer les dommages et déterminer automatiquement le montant de l'indemnisation).[20]

3.2.6.2 Le secteur bancaire

L'industrie bancaire est l'une des industries les plus importantes sur lesquelles nous comptons dans nos vies, et pour maintenir la crédibilité des banques, elles utilisent l'exploration de données dans plusieurs services spécifiques, dont le plus important est de maintenir les clients et leurs intérêts.
Les applications les plus importantes dans lesquelles j'ai utilisé l'exploration de données dans ce secteur sont les suivantes :

a) Attribution de prêt automatisée et aide à la décision de crédit.

b) Détection de fraude, telle que la détection de faux passeport par identification personnelle, détection de fraude par carte de crédit et détection de vol scientifique.

c) Déterminer qui peut changer de banque grâce au comportement des clients.

d) Identification des clients en retard, notamment dans le cas de prêts à long terme.[20]

3.2.6.3 La médecine et la pharmacie

Les utilisations de la fouille de données en médecine sont très bénéfiques pour le développement de méthodes de traitement et le diagnostic précoce de la maladie, et cela est souligné dans les points suivants :

a) Prédiction la présence de maladies et/ou complications (aide au diagnostic).

b) Choisissez un antibiotique pour l'infection.

c) Choisissez une technique spécifique (suture, suture) dans la chirurgie.

Comme en la pharmacie, la classification des médicaments permet de fournir les médicaments les plus courants qui ont le même effet pour différentes maladies présentant des symptômes similaires. [20]

3.3 Etude de La fouille des motifs séquentiels

3.3.1 Concepts généraux

Les motifs séquentiels sont des cas particulaires de motifs non-fréquents et pour faire l'étude de la fouille des motifs séquentiels. On besoins d'expliquer plusieurs termes comme suite :

3.3.1.1 Base de données formelles

Une base de données formelle est définie par un triplet (O, P, R) où :

- O est un ensemble fini d'objets.
- P est un ensemble fini de propriétés. auparavant inconnues et potentiellement
- R est une relation sur O –> P qui permet d'indiquer si un objet x a une propriété p (noté xRp) ou non. [6]

Exemple : Analyse du panier dans un supermarché, O = x1, x2, x3, x4, x5, x6 est l'ensemble des transactions d'achat, P = a, b, c, d, e est l'ensemble d'articles et R est la relation indiquant si un article a est acheté dans la transaction t. [6]

Tr	a	b	c	d	e
X1	1	0	1	1	0
X2	0	1	1	0	1
X3	1	1	1	0	1
X4	0	1	0	0	1
X4	1	1	1	0	1
X4	0	1	1	0	1

TABLE 3.1: La base de données formelle [6]

3.3.1.2 Motif

Un motif d'une base de données formelle (O, P, R) est un sous-ensemble de P. L'ensemble de tous les motifs d'une base est donc l'ensemble des parties de P, noté 2^p. [6]

Exemple : Dans la base formelle précédente, x1 possède les motifs : a, c, d, ac, ad, cd et acd.[6]

3.3.1.3 Support d'un motif

Soit m $\in 2^p$, un motif. Le support de m est la proportion d'objets dans O qui possèdent le motif :

Support :

$2^p - - > [0, 1]$

m —> Support(m) = |f(m)| / |O|

- **Exemple :**

  Ce sont des exemples dans Tableau On a :

  — Support(a) = 3/6

  — Support(b) = 5/6

  — Support (ab) = 2/6

3.3.1.4 Motif fréquent

C'est un ensemble d'éléments présents dans un nombre « suffisamment grand » de lignes d'une base de données.

Soit $\sigma \in [0, 1]$. Un motif m est fréquent (sous-entendu, relativement au seuil σ) si Support(m) $\geq \sigma$. Sinon, il est dit non fréquent.[6]

3.3.1.5 Item

On appelle un item est un champ et chaque instance de bd est un item un ensemble fini d'éléments distinct I = {i1, i2, . . . , in}[42]

- **Exemple :** Dans une application de vents les articles des vents sont des items.[42]

3.3.1.6 ItemSet

Soit I = {x1, . . . , xn} être un ensemble d'éléments, chacun étant peut-être associé à un ensemble d'attributs, tels que la valeur, le prix, le profit, la distance d'appel, la période, etc. La valeur d'un attribut A de l'élément x est notée x.A. Un ensemble d'éléments est un sous-ensemble d'éléments non vides, et un ensemble d'éléments avec k-items est appelé [35] **k-ensemble d'éléments.**

- **Exemple :** Itemset {A, B, C, D} est un 4 itemset représentant les articles : **Café**, **Sucre** , **pain** et **lait**. Ces quater articles ont pu ou non être achetés ensemble lors d'une même transaction. [42]

3.3.1.7 Algorithme Apriori

L'algorithme Apriori a été le premier algorithme proposé pour l'extraction d'éléments fréquents. Il a ensuite été amélioré par R Agarwal et R Srikant et est devenu connu sous le nom d'Apriori. Cet algorithme utilise deux étapes « joindre » et « élaguer » pour réduire l'espace de recherche.

C'est une approche itérative pour découvrir les itemsets les plus fréquents.[11] L'algorithme Apriori est donné par la suite d'instructions suivantes :

a) **Génération de candidats de taille 1 :**

C1 = {a, b, c, d, e }

Algorithme 1 Apriori

ENTRÉES: Base de données de transactions D, Seuil de support minimum σ

SORTIES: Ensemble des items fréquents

$i \leftarrow 1$

$C_1 \leftarrow$ ensemble des motifs de taille 1 (un seul item)

tantque $C_i \neq \phi$ **faire**

Calculer le Support de chaque motif $m \in C_i$ dans la base

$F_i \leftarrow \{m \in C_i | support(m) \geq \sigma\}$

$C_{i+1} \leftarrow$ toutes les combinaisons possibles des motifs de F_i de taille $i+1$

$i \leftarrow i+1$

fin tantque

retourner $\cup_{(i \geq 1)} F_i$

FIGURE 3.2: L'algorithme Apriori [6]

Supports : {4, 5, 4, 4,2}.

Donc F1 = {a, b, c, d} (aucun motif fréquent ne contiendra e).

b) **Génération de candidats de taille 2 :**

Combiner 2 à 2 les candidats de taille 1 de F1 :

C2 = {ab, ac, ad, bc, bd, cd}

Donc F2 ={ab, ac,bc,bd}.

c) **Génération de candidats de taille 3 :**

Combiner 2 à 2 les candidats de taille 2 de F2 (et ne considérer que ceux qui donnent des motifs de taille 3) :

C3 = {abc, abd, acd, bcd}

Supports : {3, 2,1, 2}

Donc F3 ={abc }

d) **Génération de candidats de taille 4 :**

C4= {0} . Donc F4 = ϑ . L'algorithme retourne alors l'ensemble des motifs fréquents : F1 $\bigcup$ F2 $\bigcup$ F3

Tr	a	b	c	d	e	ItemSet
T1	1	1	1	0	0	a,b,c
T2	0	1	1	1	0	b,c,d
T3	0	0	0	1	1	d,e
T4	1	1	0	1	0	a,b,d
T5	1	1	1	0	1	a,b,c,e
T6	1	1	1	1	0	a,b,c,d

TABLE 3.2: Base de données à six transactions

3.3.2 La fouille des motifs séquentiels

3.3.2.1 Définition

La fouille de motifs séquentiels est l'extraction d'événements ou de sous-séquences ordonnés qui se produisent fréquemment en tant que modèles. Étant donné une base de données de séquences, toute séquence satisfaisant au minimum de prise en charge est fréquente et est appelée motif séquentiel.

- **Exemple :** "Les clients qui achètent un appareil photo numérique Canon sont susceptibles d'acheter une imprimante couleur HP dans un délai d'un mois." Les algorithmes pour la fouille de motifs séquentiels incluent GSP, SPADE et PréfixSpan, ainsi que CloSpan (qui extrait des modèles séquentiels fermés). [6]

3.3.2.2 Notions fondamentaux

1.Transaction : Nous appelons une transaction pour un client C'est un triple (CID, Date, Itemset) formé par l'identifiant unique du client la valeur de l'identifiant temporel de cette transaction et tous les articles de la transaction, représentant tous les articles achetés par C à la même date.[42]

2.Base de données temporelle : Nous appelons une base de données temporelle D est un ensemble de transactions (CID, Date, Itemset) Avec D = {T : (TID, Date, Itemset)

tel que Itemset = {a ∈ I}}.

CID	Date	ItemSet
C1	01/01/2008	{B,F}
	02/01/2008	{B}
	04/04/2008	{C}
	18/01/2008	{H,I}
C2	11/01/2008	{A}
	12/01/2008	{C}
	29/01/2008	{D,F,G}
C3	05/01/2008	{C,E,G}
	12/02/2008	{A,B}
	18/02/2008	{I}
C4	06/02/2008	{B,C}
	07/02/2008	{D,G}

TABLE 3.3: Base de données à six transactions [42]

Cette table représente une base de données temporelle ordonnée en fonction de l'identifiant unique du client CID et de l'identifiant temporel Date de cette base de données. La transaction effectuée par le client C1 le 18/01/2008 est considérée comme un triplet (C1 18/01/2008, H, I). Les deux articles H et I ont été achetés au cours de cette même transaction. [42]

3.Séquence de données : Une Séquence est une structure de données qui permet d'organiser un ensemble d'éléments grâce à une relation d'ordre entre ces éléments [50], C'est une suite de transactions chronologiquement ordonnées et se rapportant à un même sujet.

Une séquence utilise le principe de précédence c'est à dire chaque élément de la liste est précédé des éléments qui l'ont précédé dans les transactions d'un client donné .[25]

- **Exemple :** Revenons au DB du tableau La liste ordonnée des trois transactions effectuées par le client C2 est donnée par la séquence < {A} {C} {D, F, G}> avec : S1 = {A}, S2 = {C} et S3 = {D, F, G}.

  Cette séquence se lit comme suit :

  Le client C2 a acheté l'article A, puis l'article C, puis simultanément les trois articles D, F et G.[25]

4.Longueur d'une séquence : La longueur d'une séquence S est le nombre d'éléments dans cette séquence. Une séquence de longueur k est une k séquence.

- **Exemple :** La séquence < {A} {C} {D} {C, E}> est une séquence de 5, même si cette dernière ne contient que 4 itemsets. L'article C fait partie de deux transactions et est donc compté deux fois.[36]

  Remarque : La longueur d'une séquence S dépend du nombre d'articles contenus dans la séquence et non du nombre d'ensembles d'articles ou de transactions. [36]

5.Fréquence d'une séquence : Une séquence est considérée fréquente, si le support de cette séquence est supérieur ou égal au support minimum. Celui –ci est introduit par le client afin de mesurer la pertinence d'une séquence. **6.Support d'une séquence :** Support d'une séquence est le pourcentage de clients qui support cette séquence. [31]

3.4 Fonction de la fouille de motifs séquentiels

Le problème de la recherche de modèles séquentiels est la découverte de l'ensemble complet des modèles séquentiels, et pour cette raison, les scientifiques et les chercheurs dans ce domaine cherchent à développer des méthodes, et les algorithmes les plus importants utilisés pour résoudre ce problème sont représentés dans le suivant : [35]

3.4.1 Les algorithmes de la fouille de motifs séquentiels

La fouille des motifs séquentiels peut être divisée en quatre parties :

a) Algorithme basé sur Apriori

b) Stratégie basée sur la largeur d'abord (BFS)

c) Stratégie basée sur la profondeur d'abord (DFS)

d) Motif fermé séquentiel (closed pattern) [35]

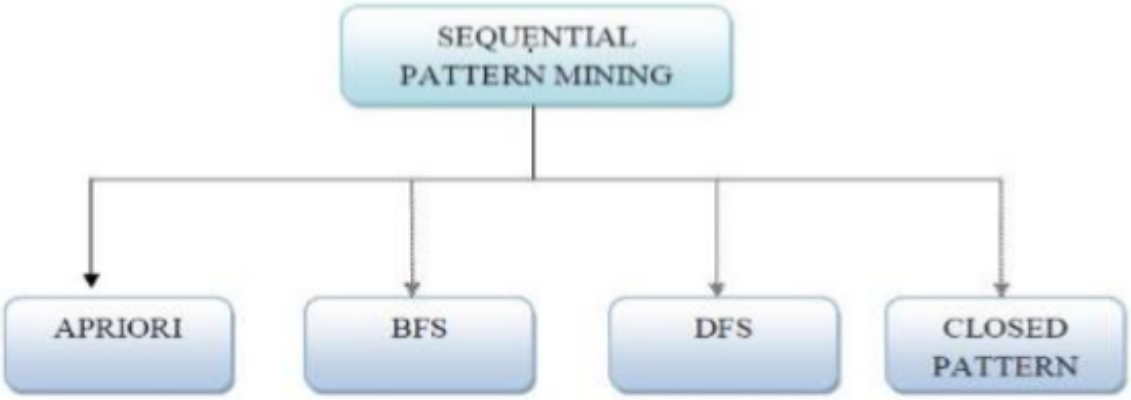

FIGURE 3.3: Classification des algorithmes de la fouille des motifs séquentiels [7]

3.4.1.1 Algorithme basé sur Apriori

La première introduction d'algorithmes la fouille des motifs séquentiels classiques basés sur Apriori remonte à 1995. Dans les algorithmes Apriori, le support minimum est spécifié par les utilisateurs sur la base d'hypothèses. Il est impossible que les utilisateurs fournissent un support minimum approprié pour une base de données à exploiter si les utilisateurs n'ont aucune connaissance des bases de données. Pour cette raison, une stratégie

de minage floue avec une prise en charge minimale indépendante de la base de données peut être utilisée, ce qui fournit une bonne interface machine qui permet aux utilisateurs de spécifier le seuil de prise en charge minimale sans aucune connaissance concernant leurs bases de données à exploiter. Apriori est utilisé dans la base de données de transactions qui comprend les séquences client. Cette base de données contient trois attributs (ID client, heure de transaction et article acheté). [35]

Apriori all : L'algorithme apriori all est une adaptation de l'algorithme de base Apriori pour les séquences, la génération de candidats et le calcul de support sont effectués d'une manière différente.

L'algorithme apriori all comporte cinq étapes : la phase de tri, la phase de transformation de phase des ensembles d'éléments, la phase de séquence et la phase maximale .[16] [51]

Dans la phase de tri, la base de données est triée avec l'ID client comme clé principale et le temps de transaction comme clé mineure, puis la base de données est convertie en base de données de séquence.

Dans la phase des ensembles d'éléments, la base de données est analysée pour obtenir des 1-séquences et aussi grand 1itemset. L'ensemble des ensembles d'éléments est mappé à un ensemble d'entiers contigus.

Dans la phase de transformation, la base de données de séquences est transformée en un ensemble de bases de données de séquences par ces grands ensembles d'éléments. Désormais, si la base de données de séquences ne contient pas d'ensembles d'éléments volumineux, cette séquence n'est pas reportée dans la séquence transformée.

Dans la phase de séquence, des passes multiples sont effectuées sur la base de données pour trouver le motif séquentiel.[51]

Dans l'algorithme apriori all, les séquences 1 fréquentes sont identifiées à partir de la base de données transformée et cette séquence 1 fréquente devient l'ensemble de départ pour trouver les séquences 2 fréquentes et le processus se déroulera jusqu'au point où aucune génération de séquences plus fréquentes ne sera possible. Après la phase de séquence, de nombreuses séquences fréquentes sont générées.[35]

Il utilise une base transactionnelle sur laquelle il effectue les étapes suivantes :

a) **Étape de tri :** consiste à trier la base de données en fonction de l'identifiant du client et de la date de la transaction. Le but est de mapper la base transactionnelle

sur une base séquentielle.

b) **Étape ItemSet :** consiste à trouver toutes les séquences fréquentes de longueur 1 à l'aide de l'algorithme Apriori. Veuillez noter que le support est le nombre de clients et non le nombre de transactions.

c) **Étape de transformation :** consiste à mapper chaque transaction sur tous les modèles fréquents séquentiels contenus dans la transaction puis à mapper chaque modèle de fréquence séquentielle à un entier.

d) **Étape de séquence :** consiste à trouver toutes les séquences fréquentes à l'aide d'un algorithme ressemblant à l'algorithme Apriori. [21]

3.4.1.2 Algorithmes basés sur BFS

Les algorithmes de recherche par respiration d'abord (au niveau du niveau) décrivent les algorithmes basés sur Apriori car toutes les séquences k sont construits ensemble à chaque k éme itération de l'algorithme lorsqu'ils traversent l'espace de recherche. Plusieurs algorithmes sont développés en utilisant le principe des algorithmes BFS. Certains d'entre eux sont illustrés dans les sections suivantes .[7]

Algorithme GSP : L'algorithme GSP proposé dans [40], est le même que l'algorithme apriori all, mais il ne nécessite pas de trouver tout l'élément est défini en premier. Cet algorithme permet :

a) Placer des limites sur la séparation temporelle entre les éléments adjacents dans un motif,

b) Permettre aux éléments inclus dans l'élément de modèle de couvrir un ensemble de transactions dans une fenêtre de temps spécifiée par l'utilisateur,

c) Permettre la découverte de modèles à différents niveaux d'une taxonomie définie par l'utilisateur. De plus, GSP est conçu pour découvrir des modèles séquentiels généralisés.

L'algorithme GSP effectue plusieurs passages sur la base de données de séquences comme suit :

- à la première passe, il trouve les séquences fréquentes qui ont le support minimum
- A chaque passage, chaque séquence de données est examinée afin de mettre à jour le numéro d'occurrence des candidats contenus dans cette séquence.[7]

3.4.1.3 Algorithmes basés sur DFS

Les algorithmes adoptant cette caractéristique ne montrent qu'une méthode d'élagage inefficace et engendrent un grand nombre de séquences candidates, ce qui nécessite de consommer beaucoup de mémoire au début du minage. Plusieurs algorithmes sont développés en utilisant le principe des algorithmes DFS.

Certains d'entre eux sont mentionnés dans les sections suivantes :

1.Algorithme SPADE Cet algorithme est proposé dans [49] et il inclut les caractéristiques d'un partitionnement de l'espace de recherche où l'espace de recherche comprend la disposition verticale de la base de données. L'espace de recherche dans SPADE est représenté comme une structure en treillis et utilise la notion de classes d'équivalence pour le partitionner. Il décompose le réseau d'origine en sous-réseaux plus légers, de sorte que chaque sous-réseau peut être entièrement traité à l'aide d'une méthode de recherche en largeur d'abord ou en profondeur d'abord (SPADE est également une méthode basée sur DFS).

Le comptage de support SPADE de la méthode de séquence candidate comprend des opérations au niveau du bit ou logiques. L'avantage de SPADE est, il utilise une méthode de comptage de support plus efficace basée sur la structure et SPADE montre une linéaire évolutivité par rapport au nombre de séquences.

2.Algorithme FreeSpan FreeSpan est un algorithme proposé par Pei et al en 2001[[49] dans le but de réduire la génération de sous-séquences candidate. Il utilise des bases de données projetées pour générer des annotations de base de données afin de guider le processus d'exploration pour trouver modèles fréquents. L'idée générale de FreeSpan est d'utiliser des éléments fréquents pour projeter des bases de données de séquences dans un ensemble des bases de données projetées plus petites de manière récursive en utilisant les ensembles fréquents actuellement extraits et des fragments de sous-séquences dans chaque base de données projetée sont générées, respectivement.

Deux alternatives de projections de base de données peuvent être utilisées Niveau par niveau projection ou projection de niveau alternatif.

La méthode utilisée par FreeSpan divise les données et l'ensemble des modèles fréquents à tester, et limite chaque test en cours à la base de données projetée correspondante plus petite. FreeSpan analyse la base de données d'origine que trois fois, quelle que soit la lon-

gueur maximale de la séquence. Les résultats expérimentaux montrent que FreeSpan est efficace et exploite l'ensemble complet des modèles et il est considérablement plus rapide que l'algorithme GSP. Le coût principal de FreeSpan est de traiter les bases de données projetées. [7]

3.Algorithme PrefixSpan Cet algorithme proposé dans [29], cet algorithme utilise un algorithme basé sur la projection. L'idée générale est de ne vérifier que les sous-séquences de préfixe et seules leurs sous-séquences de suffixe correspondantes sont projetées dans des bases de données projetées, plutôt que la projection de la base de données de séquences. PrefixSpan utilise une application directe de la propriété apriori afin de réduire séquences candidates aux côtés des bases de données projetées.

De plus, PrefixSpan est efficace car il extrait l'intégralité d'ensemble de modèles et a une exécution beaucoup plus rapide que l'algorithme GSP et FreeSpan. Le coût majeur de PrefixSpan, de la même manière que FreeSpan, est la construction de bases de données projetées. Pour chaque base de données séquentielle, PrefixSpan doit construire une base de données projetée. Une fois la projection de la base de données effectuée, l'utilisation de la projection à deux niveaux représentés dans FreeSpan et PrefixSpan par la S-Matrix [49] [29] est un moyen plus rapide d'exploiter. L'idée principale de PrefixSpan L'algorithme consiste à utiliser des préfixes fréquents pour diviser l'espace de recherche et pour projeter des bases de données de séquences. Son objectif est de rechercher les séquences pertinentes. [7]

3.4.1.4 Algorithmes séquentiels fermés basés sur des modèles

Les algorithmes d'exploration de motifs séquentiels présentés précédemment exploitent l'ensemble complet des sous-séquences fréquentes satisfaisant un seuil minimal de prise en charge. Néanmoins, comme une longue séquence fréquente contient un nombre combiné de séquences fréquentes sous-séquences, le processus d'extraction générera un grand nombre de sous-séquences fréquentes pour les motifs longs, ce qui est cher en temps et en espace. L'exploration de motifs fréquente (ensembles d'objets et séquences) n'a pas besoin d'être minée tous les modèles fermés car cela conduit à une meilleure efficacité, ce qui peut vraiment réduire le nombre de sous-séquences fréquentes [47]. Dans la section suivante, deux algorithmes CloSpan et BIDE [30] : sont décrits :

1.Algorithme CloSpan Cet algorithme proposé par [47] dans le but de réduire le coût en

temps et en espace lors de la génération de nombres explosifs des modèles de séquences fréquents. CloSpan n'exploite que des sous-séquences fermées fréquentes, c'est-à-dire les séquences ne contenant pas de super séquence avec le même support, au lieu d'exploiter l'ensemble complet des sous-séquences fréquentes. Le procédé d'extraction utilisé par CloSpan est divisé en deux étapes :

a) Un ensemble candidat est généré dans la première étape qui est plus grand que le final ensemble de séquences fermées. Cet ensemble est appelé ensemble de séquences fermées suspectes.

b) Une méthode d'élagage est appelée dans la deuxième étape pour éliminer les séquences non fermées. La principale différence entre CloSpan et PrefixSpan est que CloSpan évite la traversée inutile de l'espace de recherche. L'utilisation de méthodes de sous-modèle en arrière et de super-modèle en arrière, certaines les motifs seront absorbés ou fusionnés, ce qui réduira la croissance de l'espace de recherche.

3.5 Domaines d'application

Les données séquentielles considérées sont des séquences ordonnées de symboles (lettres, signaux, états, événements ...) et sont au cœur de domaines aussi divers que :

a) Examiner les séquences d'ADN.

b) Activation de l'appareil de surveillance.

c) Etudier le comportement dans le temps des acheteurs ou des utilisateurs.

d) Etude de carrière.

e) Aide à la décision.

f) Analyse des réseaux de communication etc. [12]

g) Marketing.

h) Séquences d'achat client

i) Modèles d'appels téléphoniques, flux de clics Weblog.[20]

3.6 Conclusion

Compte tenu des problèmes auxquels le spécialiste est confronté notamment dans la détection de la fraude bancaire et les moyens de la combattre, nous avons abordé l'application permettant la détection de la fraude bancaire en fonction du comportement séquentiel des clients. L'objectif de ce chapitre est d'étudier les techniques de la fouille des motifs séquentiels tels que l'algorithme d'apriori all dans le problème de la fraude bancaire et comment nous le résolvons.

Chapitre 4

Conception

4.1 Introduction

Après avoir défini les concepts théoriques liés à la fraude bancaire et à ses méthodes de détection, nous expliquerons dans ce chapitre la finalité, la structure et la conception générale du projet. Nous détaillerons chaque étape en citant les principaux algorithmes et techniques utilisés dans chacune des étapes.

4.2 L'ojectif

L'objectif de cette étude est de créer une application qui permet aux banques d'éviter les pertes résultant du vol et de la fraude bancaire et est basée sur le principe d'une base de données en série, qui est une séquence de comportements clients dans une période de temps spécifique et nous utilisons deux méthodes d'apprentissage représentées dans knn et spmf.

4.3 Architecture globale

Notre système classe les séquences d'une certaine manière pour faire la detection de fraude, et le système comme entrée, une base de données séquentiels pour les clients et leurs propriétés et être utilisée dans la phase d'apprentissage.

Cet étape est appelée prétraitement, elle contient des opérations spécifiques à la base de données où notre base de données besoin d'être nettoyée et filtrée. Après avoir obtenu une nouveau base de données traitée qui est appelée une base de données séquentiels, nous la divisons en deux parties, une partie pour les entraînements et une partie pour les tests.

Le module d'apprentissage utilise l'algorithme d'apprentissage comme KNN et SPMF pour obtenir un modèle qui est appliqué à la règle de test.

Une fois la performance de notre modèle est jugée satisfaisante, c'est-à-dire qu'il a pu atteindre un taux de reconnaissance acceptable (un certain pourcentage). alors ce modèle peut être appliquée.

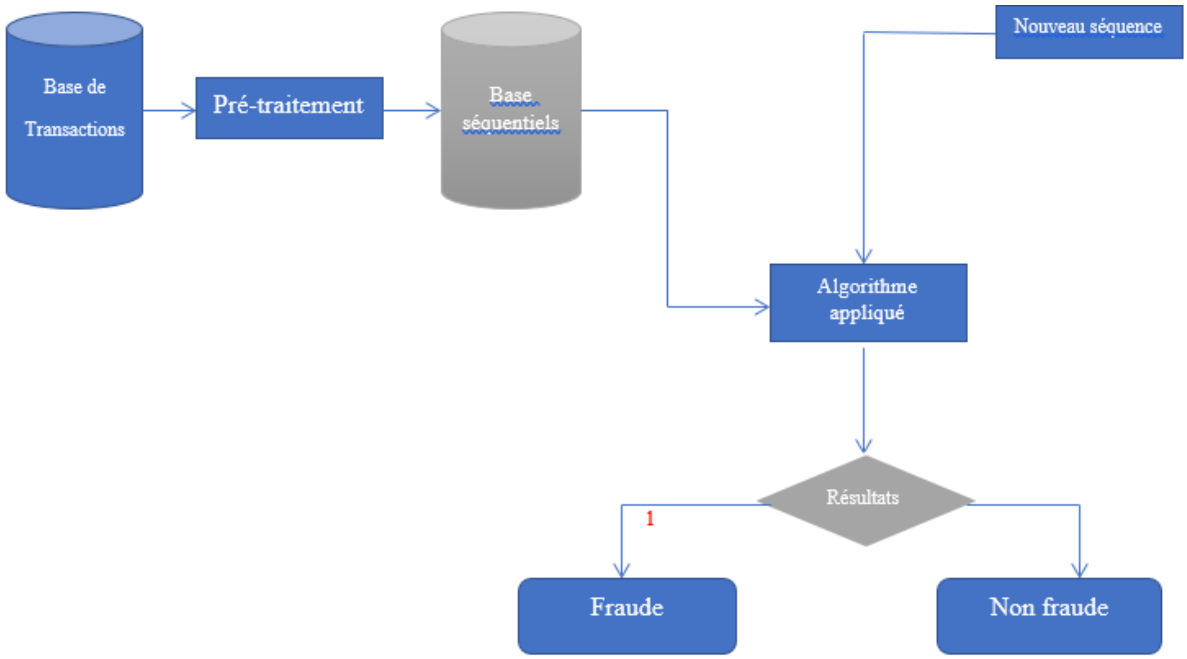

FIGURE 4.1: Architecture globale

4.4 Architecture détaillée

Ci-dessous, nous détaillons chacune des étapes que nous avons traversées dans notre système. 4.1 Description de la base de données :

4.4.1 Description de la base de données

Nous avons exécuté BankSim pendant 180 étapes (environ six mois), plusieurs fois et calibré les paramètres afin d'obtenir une distribution suffisamment proche pour être fiable pour les tests. Nous avons collecté plusieurs fichiers journaux et sélectionné les plus précis. Nous avons injecté des voleurs qui visent à voler en moyenne trois cartes par étape et à effectuer environ deux transactions frauduleuses par jour. Nous avons produit 594643 enregistrements au total. Où 587443 sont des paiements normaux et 7200 transactions frauduleuses. Puisqu'il s'agit d'une simulation aléatoire, les valeurs ne sont bien sûr pas identiques aux données d'origine.[33]

	step	customer	age	gender	zipcodeOri	merchant	zipMerchant	category	amount	fraud
0	30	'C1000148617'	'5'	'M'	'28007'	'M1888755466'	'28007'	'es_otherservices'	143.87	0
1	38	'C1000148617'	'5'	'M'	'28007'	'M1741626453'	'28007'	'es_sportsandtoys'	16.69	0
2	42	'C1000148617'	'5'	'M'	'28007'	'M1888755466'	'28007'	'es_otherservices'	56.18	0
3	43	'C1000148617'	'5'	'M'	'28007'	'M840466850'	'28007'	'es_tech'	14.74	0
4	44	'C1000148617'	'5'	'M'	'28007'	'M1823072687'	'28007'	'es_transportation'	47.42	0
...	...	...	...	...	...	...	...	...	...	...
594638	174	'C999723254'	'2'	'M'	'28007'	'M1823072687'	'28007'	'es_transportation'	31.94	0
594639	176	'C999723254'	'2'	'M'	'28007'	'M1823072687'	'28007'	'es_transportation'	1.92	0
594640	177	'C999723254'	'2'	'M'	'28007'	'M85975013'	'28007'	'es_food'	62.55	0
594641	178	'C999723254'	'2'	'M'	'28007'	'M1823072687'	'28007'	'es_transportation'	25.96	0
594642	179	'C999723254'	'2'	'M'	'28007'	'M1823072687'	'28007'	'es_transportation'	32.96	0

FIGURE 4.2: une base de données transactionnelle

4.4.2 Pré-traitement des données

L'étape de Pré-traitement des données est basée sur la récupération de l'ensemble des données brutes, les nettoyer et les agréger pour pouvoir les comprendre et les exploiter.

Notre objectif est obtenir une base correcte qui est basée sur les séquences des clients de la banque et ce sont des données textuelles représentant les clients, l'état de la fraude, les séquences où il est traité.

4.4.2.1 Base de données séquentiel

On calcule les séquences de chaque client et faire l'état de fraud :

4.4.3 Apprentissage

Il combine deux modules, formation et validation, chacun utilisant une partie.
A partir de la base de caractéristiques elle est divisée en deux parties, une base d'entraînement et une base.
Examen. Le module de formation utilise la règle de formation pour fournir un modèle la décision alors que l'unité de validation utilise la règle de test pour mesurer Soumettez un modèle de performance.

	CUSTUMER	SEQUENCE	FRAUD
2	C1000148617	JKJLMMMMDMMMMMMMFMMMMMMMMMMMMMMMMMMMMM	0
3	C100045114	MMMMAOECNEEOOFFGHMMMMKMOOOOAMMMMMOOOMMMM	0
4	C1000699316	MDMMMMMMMMMMMMDMMMMMMMMMMMMMDMDMMMMDM	0
5	C1001065306	EEEEOEKOOEEECOOFKEEEEOEEELKEKK	0
6	C1002658784	KJHMAMMMMMMMMMCMMEMMMMMMMEEKHHMMMMMMMMM	0
7	C1002759277	MMMMOMMMMMMMMMMMMMHMAMMMMMKMMMMMMMM	0
8	C1004109477	OOCHMMMMMMMBMMMMMMCMMMMMMMMFMMMMMMMM	0
9	C1004300450	MMMMMMMMMMMMMMMMMMMMMMMMMMMMAMMMMA	0
10	C1004532392	OAMMMMMMOMMMMMMMMMMMMMMMMMMMMMAMMMMMM	0
11	C1005126300	MMMMCMMMMMCMMOMMMMMMDMMMMMMMMDEDMMMMM	0
12	C1005495267	MOOMMMMMMMMMMMMMMFMMMMMMMMMMDMMMMMMMM	0
13	C1005806982	EEEEEOOOOOKEEACEEEEEEC	0
14	C1006176917	MMMMMMMMMMMMMMMMMMMMMMMMAMMMMMMMMMMM	0
15	C1007572087	JHMEMMMMMMMMMMMHMMMMMMMMMMMMMMMMMNMMMM	0
16	C1007790716	MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM	0
17	C1008918174	GOKFCOOOCLMMMDMMMEMMMMMMMMMMMMMMMMMDM	0
18	C1009080922	AMMMMMMMMMMMMAMMMMMMMMMMMMMMMMMMMHM	0
19	C100992504	LMMMMMMMMDMMFMMMMMMMMMMMMMMMFFMMMMMMM	0

FIGURE 4.3: une base de données séquentielle

4.4.3.1 Entraînement

Pour entraîner notre modèle, nous avons choisi l'algorithme knn avec une base séquentiel qui nous l'avons déjà montré dans le chapitre précédent pour classer les états de fraudes.

Nous avons utilisé beaucoup de fichier d'entrée de données de base et l'importation seuil base de données spécifiée par les séquences.

Input :

Le fichier d'entrée c'est un fichier texte où chaque ligne représente une séquence à partir d'une base de données de séquence. Chaque élément d'une séquence est un transaction et les éléments du même ensemble d'éléments dans une séquence sont séparés par un espace unique.

Le fichier sorti : C'est un résultat de la prédiction de la nouvelle séquence avec la méthode est utilisée.

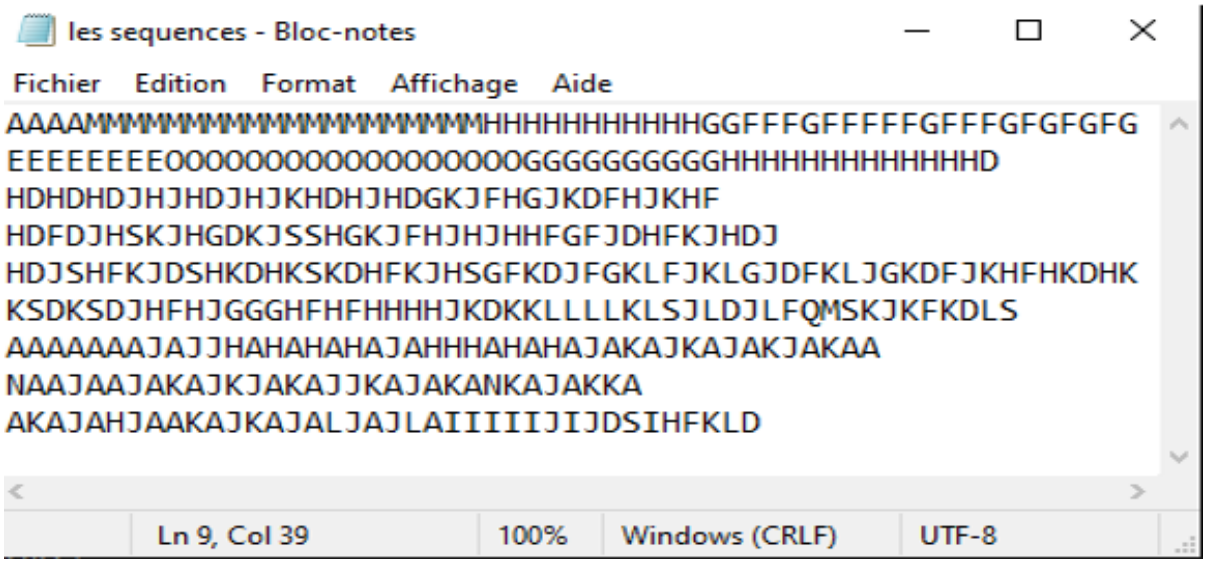

FIGURE 4.4: Présenter le ficher d'entrée de base des données

4.4.4 Utilisation

C'est l'étape le plus importante dans notre application. Après être arrivé au meilleur taux de reconnaissance, avec la meilleur modèle dans la phase précédente, nous devons l'utiliser sur des séquences de clients informations non étiquetées, et le modèle nous permet de prédire la classe de la nouveau séquence si elle est fraude ou non fraude comme suite :

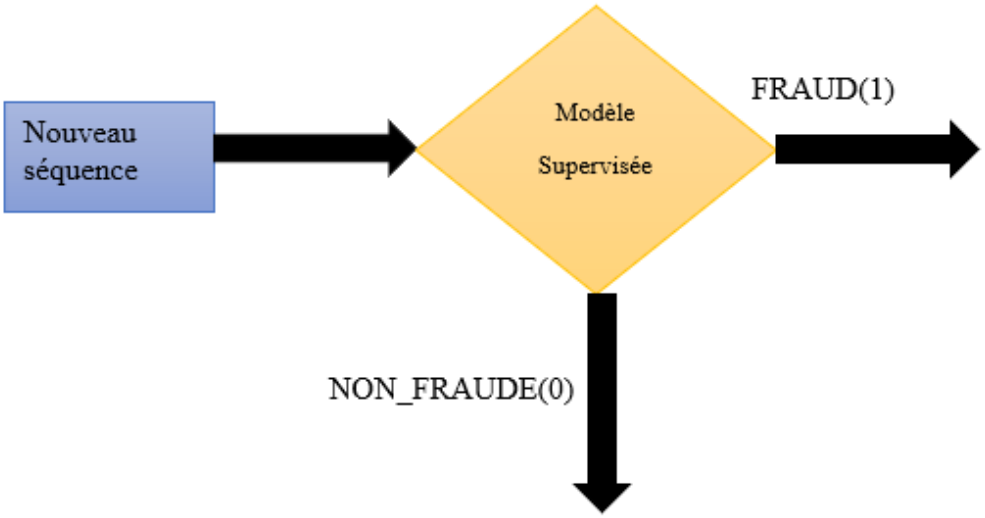

FIGURE 4.5: Utilisation

4.5 Conclusion

La conception est l'étape la plus importante dans notre application.

Dans lequel nous faisons toutes les étapes nécessaires pour avoir un modèle efficace, dans le chapitre suivant, nous implémentons notre système.

Chapitre 5

Implémentation

5.1 Introduction

Dans ce chapitre, Nous allons implémenter notre solution pour detection la fraude dans les séquences bancaire des clients. Cette implémentation consiste à :

- Constituer la base de données séquentiel à partir la base de données transactionnelle de la banque dans 6 moins.
- Prétraiter les données pour améliorer leur qualité et l'adapter au traitement.
- Configurer et tester les séquences de dans la prédiction.

5.2 Environnement d'exécution

L'implémentation de cette application est réalisée sur un ordinateur ayant les caractéristiques suivantes :

5.3 Outils et langages de développement

L'implémentation de notre application nécessite un langage de développement et un éditeur de texte destiné à la programmation. Ce choix a été fait sur le langage Python et Spyder (anaconda3) notre problématique.

5.3.1 Python

Python est un langage de programmation interprété multi-paradigme qui prend en charge la programmation orienté objet, impérative structurelle et fonctionnelle. Il est considère comme un langage de haut niveau grâce à ses fonctionnalités avancées tels que la gestion automatique de la m´mémoire (garbage collection).[8]

FIGURE 5.1: Logo Python

En plus de sa simplicité et facilité d'utilisation, le langage Python fonctionne sur la plupart des plates-formes informatiques, des smartphones aux ordinateurs centraux, de Windows a Unix avec notamment GNU/Linux en passant par MacOs, ou encore Android, iOS. Son avantage majeur est sa modularité, la définition de langage est très compacte autour de son noyau, il existe des nombreuses librairies et modules qui ont été développées et qui optimisent la productivité des programmeurs et facilitent la tâche de programmation. Afin de créer un modèle d'apprentissage, plusieurs librairies peuvent ˆêtre utilisé, on note les librairies suivantes :

5.3.1.1 Numpy

NumPy (diminutif de Numerical Python) est une librairie destiné au langage Python qui permet de stocker et effectuer des opérations sur les données .D'une certaine manière, les tableaux Numpy sont comme les listes en Python, mais Numpy permet de rendre les opérations beaucoup plus efficaces, surtout sur les tableaux de large taille. Les tableaux Numpy sont au cœur de presque tout l'´écosystème de data science en Python.[38]

FIGURE 5.2: Logo Numpy

5.3.1.2 Matplotlib

Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques. Elle permet l'exportation des graphiques sous plusieurs formes (PNG, JPEG, PDF...) et elle est doté d'une 'User Graphical Interface' qui permet de zoomer et explorer les graphiques facilement.

FIGURE 5.3: Logo Matplotlib

5.3.1.3 Pandas

Pandas est une bibliothèque ´écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de s´séries temporelles. Les principales structures de données sont les s´séries (pour stocker des données selon une dimension - grandeur en fonction d'un index), les DataFrames (pour stocker des données selon 2 dimensions - lignes et colonnes), les Panels (pour représenter des données selon 3 dimensions, les Panels4D ou les DataFrames avec des index hiérarchiques aussi nommés MultiIndex (pour représenter des données selon plus de 3 dimensions - hypercube). [38]

FIGURE 5.4: Logo Pandas

5.3.1.4 Jupyter Notebook

Jupyter Notebook est une application Web open source qui vous permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Les utilisations incluent : le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore.[9]

5.4 Réalisation et Implémentation

Dans cette partie, on utilisera Jupyter Notebook et le langage Python afin de réaliser le système conçu dans le chapitre précédent :

FIGURE 5.5: Logo Jupyter

5.4.1 La distance LCP

Le préfixe commun le plus long pour un tableau de chaînes est le préfixe commun entre 2 chaînes les plus dissemblables.

— Distance = 1 - LLCP(S1,S2)/ $\sqrt{S1 * S2}$

Apres faire l'ordre décroissent de les distances. Voilà un exemple :

- Input = {AAAAAHHGGFHJFHGFHHHJGHFJ}
- Output =

```
Name: DISTANCE, Length: 4112, dtype: float64
> O > A
le LCP est = 0
La distance LCP ente 2 sequence =  1.0
0        1.0
1        1.0
2        1.0
3        1.0
4        1.0
        ...
4107     1.0
4108     1.0
4109     1.0
4110     1.0
4111     1.0
Name: DISTANCE, Length: 4112, dtype: float64
```

FIGURE 5.6: La distance LCP

5.4.2 Application de KNN

Donc on peut appliquer la méthode de la classification KNN dans les étapes suivantes :

- Étape 1 : Sélectionnez le nombre K de voisins
- Étape 2 : Calculez la distance (la distance LCP)
- Étape 3 : résultat de prédiction
- Étape 4 : validation de modèle

Chapitre 6

Conclusion général

6.1 Conclusion général

Le terme d'exploration de données fait généralement référence à l'analyse de données sous différents angles, à la conversion de ces données en informations utiles et à l'établissement de relations entre des données ou des modèles de découverte de données. En raison des avantages de ses multiples méthodes de classification, il a été exploité dans de nombreux domaines économiques et politiques. Il est exposé au vol et à la fraude, qui ont causé un déficit important pour certains pays et touché les secteurs commerciaux, qui sont à la base des pays.

Les grands pays cherchent à développer des moyens de lutter contre la fraude bancaire et à améliorer la méthode la plus importante, qui consiste à utiliser les comportements des clients pour anticiper la survenance de la fraude.

Bibliographie

[1] study.com(https ://study.com/academy/lesson/what-is-bank-fraud-definition-prevention.html) :.

[2] (https ://www.avg.com/fr/signal/identity-theft). 04/05/2021.

[3] Berry law (https ://jsberrylaw.com/blog/bank-fraud-definition-penalties/) :. 04/05/2021.

[4] Difs(https ://www.michigan.gov/difs/0,5269,7-303–458212–,00.html). 04/05/2021.

[5] Ivestopedia(https ://www.investopedia.com/terms/m/moneylaundering.asp). 04/05/2021.

[6] National check fraud center. 1995/2011.

[7] Sequential pattern mining(https ://www.cc.gatech.edu/ hic/cs7616/pdf/lecture13.pdf. 2013.

[8] A comparative study of sequential pattern mining algorithm. jawahar. s assistant pro.fessor, rvs college of arts and science, coimbatore. 2015.

[9] Le langage de programmation python. 2019.

[10] Jupyternotebook (project jupyter | home).on :. 24/06/2021.

[11] Data mining, fouille de données : Concepts et techniques. jiawei han et micheline kamber. intelligent database systems research lab school of computing science simon fraser university- canada. Février2006.

[12] Apriori algorithm in data mining : Implementation with examples. May 30,2021.

[13] Dr. Abdelhamid Djeffal. . Introduction aux données séquentielles (master 2 informatique de l'optimisation et de la décision). 2020/2021.

[14] M. TOMMASI : R.GILLERON. Découverte de connaissances à partir de données,. 2000.

[15] D. Abdelhamid. Cour de classification,. *Université de Biskra*, 2019.

[16] Djeffal Abdelhamid. Ouassaf atika.(automatic bank fraud detection using support vector machines). 2014.

[17] R. Agrawal, R. ; Srikant. ., "mining sequential patterns," proceedings of the eleventh international conference on data engineering,. 1995.

[18] Sassi. Amina. Une approche basée agent pour la fouille de données, magister en informatique,. *Université de , batna*, 2013.

[19] Mr. MOUNA Azzedine. Mémoire de magister en informatique, titre : datamining distribue dans les grilles : approche règles d'association,. *Université des sciences et technologie, d'Oran,*, 2012/2013.

[20] D. Boukraâ. Classes master 1 siad et ilm.cours en fouille de données-. *Université de Jijel , 2019/2020.*, 2019/2020..

[21] Chami. Djazia. Une plate forme orientée agent pour le data mining, magister,. *Université de , batna*, 2010.

[22] Dr. Abdelhamid DJEFFAL. Fouille des motifs séquentiels. *(Master 2 SIOD).*, 2017-2018.

[23] J. P. G. DONG. Sequence data mining, springer edition,. 2007.

[24] President Edward J. Potter. Psi fraud solutions in (customer authentication : The evolution of signature verification in financial institutions) journal of economic crime management summer. 2002.

[25] Bernard ESPINASSE. information et de la fouille de données. Septembre 2008.

[26] Mickaël Fabrègue. Extraction d'informations synthétiques à partir de données séquentielles application à l'évaluation de la qualité des rivières. *doctorat, Strasbourg*, 2014.

[27] B. N. Fatima. Fake news detection using machine learning,. 2019/2020..

[28] C. GROUIN. Les techniques de la fouille de données, inalco,. 2009/2010.

[29] Jiawei Han and Micheline Kamber. Data mining concepts and techniques. *Diane Cerra San Francisco.*, .

[30] B.Mortazavi-Asl Q. Chen U.Dayal J.Han, J.Pei and M.-C. Hsu. Freespan : Frequent pattern projected sequential pattern mining. *In Proceedings of the 6th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining. ACM, New York, 355–359,, 2000.

[31] J.Han. BIDE J.Wang. : Efficient mining of frequent closed sequences. in proc. of 2004 int. *Conf. on Data Eng.*, . 2004, Boston, MA. 79–90.

[32] Layadi Kanza. Extraction des motifs séquentiels. 2016.

[33] Stefan Lopez-Rojas, Edgar Alonso ; Axelsson. Banksim : A bank payments simulator for fraud detection research inproceedings. 2014.

[34] K. R. Seeja M. Zareapoor and A. M. Alam. Analyzing credit card : fraud detection techniques based on certain design criteria. *International Journal of Computer Application,*, 2012.

[35] Bharat Chaudhari Manan Parikh and Chetna Chand. A comparative study of sequential pattern mining algorithms, volume 2, issue 2,. feb-2013.

[36] Abdelhak Mansoul. fouille de données biologiques : étude comparative et expérimentation. *magister, Oran*, 2010.

[37] Alice Marascu. Extraction de motifs séquentiels dans les flux de données,. *Docteur en Sciences, France*, 2009.

[38] Nick McClure. Tensor flow machine learning cookbook. packt publishing ltd. 2017.

[39] P. PREUX. Fouille de données : Notes de cours,. *Université de Lille 3*, 09-oct-2008.

[40] R. Agrawal R. Srikant. Mining quantitative association rules in large relational table., proc. of the acmsigmod. *Conference on Management of Data, Montreal, Canada,*, 1996.

[41] N. VENKATESAN S. PRABHU. Data mining and warehousing. *New Age International (P) Ltd., Publishers, New Delhi*, 2007.

[42] Time-Series and Sequence Data. Chapter 8 : Mining stream.

[43] M. W. N. TOLOFON. Etude et mise en place d'un système basé sur le machine learning pour la détection de fraudes monétiques,. 2019/2020.

[44] P. F.-. Viger. An introduction to sequential pattern mining, posted. 2012017-03-08.

[45] Stephan Kovach Wilson Vicente Ruggiero. Online banking fraud detection based on local and global behavior. 2011.

[46] KIPROTICH JOHN YEGO. The impact of fraud in the banking industry : A case of standard chartered bank. 2016.

[47] Shin-Yi Wu Kwei Tang Yen-Liang Chen, Mi-Hao Kuo. ‖discovering recency, frequency, and monetary (rfm) sequential patterns from customers' purchasing data‖,. *Electronic Commerce Research and Applications*, Electronic Commerce Research and Applications.

[48] O. R. ZAÏANE. Principles of knowledge discovery in databases,. *University of Alberta,*, 1999.

[49] M.J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering 12(3), 372-390*, , 2000.

[50] M' zali hassen. les règles d'association séquentielles. 2006.

[51] Zheng Zhu. "data mining survey. *ver 1.1009*.