République Algérienne Démocratique et Populaire

Ministère d'Enseignement Supérieur et de la Recherche Scientifique

Université ABBAS LAGHROUR- KHENCHELA-

Département MI

Mémoire

Présenté en vue d'obtenir le diplôme de

Master en Informatique (*LMD*)

Spécialité : *Génie logiciel et systèmes distribués.****(GLSD)***

Conception D'une Application Basée Sur La Classification Pour Recherche Des Employés Dans Le Domaine De Télétravail

Encadrer Par

Dr AYADI ABDELGHAFAR

Présenté Par

MECHRI KARIMA

ARAB SALSABIL

Année universitaire : **2020/2021**

Dédicace

J'ai le plaisir de dédie ce travail

A mon père ***ARAB DJAMEL*** *et à ma chère mère* ***YASMINA*** *pour votre amour et la confiance que vous avez placée en moi depuis mon très jeune âge, votre soutien de tout ordre et pour vos encouragement, vos peur pour moi et tout bon moment avec vous…*

A mon oncle adorée ***YACINE ARAB*** *que Dieu lui fasse miséricorde*

A mes frères ***YASSER YACINE*** *et* ***AMIR*** *qui mon soutenu par leur amour et leur encouragements que dieu vous protège.*

A mes sœur ***HADIL*** *et* ***ALAA*** *pour vos encouragements, votre amour et votre aide, que dieu vous protège …*

A ma meilleur amie ***SANA AISSAOUI*** *A tous les moments passés avec toi, en gage de ma profonde estime pour l'aide que tu m'as encouragé.*

A ma grand-mère ***SALMI HENIA*** *et mon grand-père* ***ARAB BELKACEM***

A mes cousins ***DHIA EDDIN, BAHA EDDIN*** *et mon chère* ***MAHDI***

A tout ma famille spécialement ***SAMIA , FATIHA ,HANAN, OMAR*** *et* ***SORIA***

Ma chère binôme ***MECHRI KRIMA*** *et à toute sa famille.*

A mes amis spécialement ***REGHIS HAYAM***

A tous mes collègues

Merci pour votre aides votre soutiens et vos prières.

SALSABIL

Dédicace

Je dédie cette thèse :

*A mon père **ELHADJ MECHRI** et A l'âme de ma mère **AICHA BOUGATTAYA** pour leur amour inestimable, leurs sacrifices, leur confiance, leur soutien et toutes les valeurs qu'ils ont su m'inculquer.*

*A mon cher frère **KAMEL MECHRI***

A tous les moments d'enfance passés avec toi mon frère, en gage de ma profonde estime pour l'aide que tu m'as apporté. Tu m'as soutenu, réconforté et encouragé. Puissent nos liens fraternels se consolider et se pérenniser encore plus.

*A mes sœurs et en particuliers **AHLEM, WASSILA, KHADIDJA, IMEN** pour leurs encouragements durant tout mon parcours.*

*Ma chère binôme **SALSABIL ARABE** et à toute sa famille.*

A TOUTE MA FAMILLE

Aucun langage ne saurait exprimer mon respect et ma considération pour votre soutien et encouragements. Je vous dédie ce travail en reconnaissance de l'amour que vous m'offrez quotidiennement et votre bonté exceptionnelle. Que Dieu le Tout Puissant vous garde et vous procure santé et bonheur.

Karima

Remerciement

*C*e travail est le fruit de la combinaison d'efforts de plusieurs personnes. Tout d'abord nous remercions Dieu Tout-Puissant qui par sa grâce nous a permis d'arriver au bout de nos efforts en nous donnant santé, force et courage et en nous faisant nous entourer de personnes merveilleuses pour lesquelles nous tenons à remercier. Nous remercions:

*D*oyen de la Faculté des Mathématiques et de l'Information pour ses conseils et les efforts de son équipe pour assurer la qualité de la formation.

*N*otre Directeur de mémoire, Dr AYADI ABDELGHAFAR pour son encadrement sans faille, son soutien moral, sa rigueur au travail, ses multiples conseils, ses orientations et sa disponibilité malgré ses multiples occupations.

*T*ous les professeurs d'informatique, pour la qualité de l'enseignement et leurs conseils qui nous ont permis de poursuivre notre parcours universitaire jusqu'à présent.

*N*ous tenons à exprimer nos sincères remerciements aux membres du jury qui nous ont honorés de l'évaluation de ce travail.

*N*otre remerciement les plus chaleureux vont à tous nos camarades au Master 2 GLSD de l'Université *ABBES LAGHROUR KHENCHELA*, ainsi que toute notre autre camarade de cette Université pour leur présence dans les moments difficiles et les excellents moments que j'ai passés avec eux tout au long de cette année.

*A*u terme de ce travail, nous tenons à remercier tous ceux qui ont contribué de près ou de loin à la mise en œuvre de ce mémoire.

Résumé

Le télé- recrutement joue un rôle important dans le télétravail en raison des facilités, de la crédibilité et de l'intégrité qu'il offre dans l'emploi, et pour réaliser le télé recrutement , nous avons utilisé des supports automatisés, en particulier des classifications qui fournissent de nombreux algorithmes qui classent le travailleur en fonction de ses caractéristiques et considèrent l'étendue de sa qualification pour travailler à son insu Ou en regardant son nom, ces algorithmes nous permettent d'obtenir des résultats proches pour faciliter la recherche et la mise en place de nombreux entretiens d'embauche par l'employeur.

Parmi les algorithmes de classification, nous avons choisi les K plus proches voisins (KNN) et les avons appliqués à un ensemble de données de personnes pour réaliser le télé-recrutement grâce au travail à distance.

Table of Contents

Chapitre 2 : la classification supervisée et non supervisée

Liste De Figures

Liste des tableaux

INTRODUCTION GÉNÉRALE

INTRODUCTION GÉNÉRALE

Le télétravail est une organisation de travail qui est actuellement en pleine expansion. Le Code du travail le définit comme une organisation du travail qui effectue un travail en dehors des locaux de l'employeur, mais en raison des technologies de l'information et de la communication, il peut être effectué régulièrement et volontairement chez l'employeur dans le cadre du contrat de travail. Le télétravail est une forme d'organisation qui permet de réduire le coût de la structure de l'entreprise : l'espace requis pour les activités de l'entreprise est réduit, donc d'autres coûts sont également réduits : électricité, assurance chantier, etc.

Le télétravail généralisé complexifie les campagnes de recrutement. Pourtant, il devient un critère de choix des candidats. Pour les entreprises en quête de talents, un remaniement du processus d'embauche est nécessaire, des entretiens à l'accueil des recrues. Dans ce cas en a choisi la classification pour fait cette idée.

Il Ya deux types d'approches de classification automatique peuvent être distingué : La classification supervisée (**comme** : *Arbres de décision, KNN, Naïve Bayes, Les réseaux de neurones, SVM*) et la classification non supervisée (**comme** : *K-moyennes, C-moyennes floues, Classification hiérarchique*). Ces deux méthodes diffèrent sur la façon dont les classes sont générées.

En effet dans le cas de la classification non supervisée, les classes sont calculées automatiquement par la machine, par contre, dans la classification supervisée, on a classé des individus dans des classes définies a priori.

Pour mettre en œuvre le processus de recrutement à distance, nous avons utilisé l'algorithme KNN (basée sur la détection des synonymes), qui a facilité le processus de sélection des employés, car nous avons donné une liste de personnes, y compris celles qui sont acceptées et celles qui ne le sont pas, selon les données (âge, sexe, wilaya, Expérience).

Le processus de sélection des employés se déroule en plusieurs étapes : nous donnons une liste de personnes divisée en deux catégories : la première catégorie est des employés acceptables et la deuxième catégorie est des employés inacceptables.

On utilise ces deux catégories pour classer une personne selon ses données en calculant la distance euclidienne entre cette personne et les personnes dans les données, puis on range ces distances par ordre croissant, puis on extrait les salariés les plus proches de lui en précisant la valeur de k par lequel nous déterminons la majorité des employés (Soit acceptée ou rejetée) Si la majorité est acceptée, alors la personne est acceptée, et si la majorité est rejetée, alors la personne est rejetée.

CHAPITRE 01 : LE TELETRAVAIL

CHAPITRE 1

Le Télétravail

Sommaire

INTRODUCTION

Aujourd'hui, il n'y en a pas deux qui diffèrent de l'importance de l'informatique dans le développement technologique dans tous les domaines d'études, scientifiques, sociaux, politiques et surtout économiques, car l'informatique a grandement contribué au développement des institutions économiques de tous les côtés, mais en 2020 , après l'épidémie qui a terrifié le monde entier, COVID-19, le niveau économique a baissé dans tous les pays, en raison des précautions de quarantaine et de prévention des maladies, car dans certains pays la vie s'est presque arrêtée pour la sécurité du citoyen Le développement des médias automatisés était le seul débouché qui pouvait sauver la situation économique misérable à laquelle étaient exposés divers pays du monde. Où il a eu recours à une solution de travail à distance pour atténuer cette crise, et comme on sait l'ampleur du développement de la technologie et de ses moyens qui permettent au salarié de travailler chez lui sans mettre sa vie et celle de sa famille en danger Dans ce chapitre, nous aborderons le travail à distance et son rôle dans le projet achevé.

I. Le Télétravail

1. Histoire

L'individualisation, la flexibilité et le développement des technologies de l'information et de la communication (TIC) font partie des nombreux concepts qui caractérisent le monde du travail actuel. C'est dans ce contexte de transformation du travail que s'est développé, principalement de manière informelle, le télétravail 2006.

Historiquement, le travail à distance est apparu à deux fois dans le monde du travail. Entre 1973 et 1989, Le télétravail est considéré comme un projet technologique permettant de mener des activités professionnelles en dehors de l'entreprise et de limiter la consommation énergétique des déplacements professionnels. Cependant, ce nouveau concept est apparu trop tôt, la technologie n'est pas encore mature et la protection de l'information est la priorité absolue, le travail à distance est donc considéré comme un échec.

C'est dans les années 1990 que le télétravail a vraiment émergé et a été considéré comme la réponse à la croissance économique et au développement des technologies de l'information et de la communication. En fait, la combinaison du développement des TIC et de la réduction des coûts a en fait changé les règles. L'expansion d'Internet a apporté une contribution importante au développement des pratiques de bureau à distance, qui ne sont plus considérées comme des projets techniques, mais comme des outils utiles pour l'autonomie des employés qui souhaitent gérer leur travail.

Il ne fait aucun doute que l'intensification des technologies de l'information et de la communication et la volonté des entreprises de faire des salariés des indépendants sont au cœur du développement du télétravail. Gérer les activités des entreprises distantes sans gestionnaire chargé de superviser les activités des employés distants. **[1]**

Années 2000.

En 2002, un accord a été trouvé au niveau européen pour renforcer les droits et la sécurité de ce nouveau monde du travail. Au niveau national, le président de la République, Emmanuel Macron, a finalement intégré le personnel du télétravail en réponse au décret Macron du 22 septembre 2017 définissant le télétravail comme ayant le statut et les droits du télétravail.

2. Définition

Le télétravail est défini comme l'utilisation des technologies de l'information et de la communication (TIC) - téléphones intelligents, tablettes, ordinateurs portables et ordinateurs de bureau - pour effectuer des tâches en dehors des locaux de l'employeur.

En d'autres termes, le télétravail désigne le travail effectué à l'aide des TIC en dehors de l'emplacement de l'employeur. **[2]**

Le télétravail doit faire l'objet d'un accord volontaire entre l'employeur et l'employé. En plus de l'entente sur le lieu de travail (domicile du salarié ou ailleurs), l'entente doit également préciser plusieurs autres aspects, à savoir : la durée et l'heure du travail, le mode de communication utilisé, les tâches à accomplir, le mécanisme de contrôle des travaux achevés, et Méthode de rapports.

En règle générale, la définition du travail à distance n'inclut pas les travailleurs de l'économie de plate-forme: par exemple, les travailleurs indépendants qui travaillent principalement à domicile ne sont généralement pas considérés comme des travailleurs à distance, mais peuvent être classés comme des travailleurs domestiques sur le lieu de travail. Convention sur le travail domestique de l'Organisation internationale du travail, 1996 (n° 177). **[3]**

Selon Business Europe (ex-UNICE), l'Union européenne de l'artisanat et des petites et moyennes entreprises (UEAPME), le Centre européen des entreprises à participation publique (CEEP) et la Confédération européenne des syndicats (CES)

Le télétravail est définir par :

« Une forme d'organisation et/ou de performance au travail, utilisant

Informations, dans le cadre d'un contrat ou d'une relation de travail, dans lesquelles le travail,

Elle peut également se faire dans les locaux de l'employeur, mais en dehors de ces locaux

Endroit régulier ». **[4]**

D'après qu'est ce qu'on comprit

Le télétravail désigne « toute forme d'organisation du travail dans laquelle, dans le cadre d'un contrat de travail ou d'un contrat de travail, un travail pouvant être effectué chez l'employeur est effectué régulièrement et volontairement en dehors de ces locaux par des salariés utilisant l'information et la communication technologie.

3. Les types de télétravail

Le télétravail peut être mis en œuvre selon plusieurs modalités, notamment:

■ **Espace de bureau partagé**: l'employé travaille à distance une partie ou la plupart du temps, et au bureau principal pour le reste. Lorsqu'il travaille au bureau principal, l'employé occupe un espace de travail non dédié, attribué pour une utilisation ponctuelle, et ne dispose pas d'un espace de bureau réservé, qui resterait vacant lorsqu'il télétravail.

Bureau éphémère: ce concept s'apparente à celui d'espace de bureau partagé, mais l'employé doit réserver une place à l'avance**. [4]**

■ **Télé centre**: installation qui offre des postes de travail et d'autres équipements de bureau aux employés de plusieurs organisations. Ce type de télétravail est utile dans la mesure où il permet d'offrir une technologie plus pointue que le bureau à domicile; on estime toutefois qu'il est en déclin en raison de la généralisation des ordinateurs portables, des smartphones et des réseaux à très haut débit.**[4]**

■ **Espace bureautique collaboratif**: environnement de travail virtuel, où les employés peuvent travailler en collaboration grâce à un réseau informatique, même s'ils se trouvent dans des endroits distincts. Les télétravailleurs nomades, qui travaillent au moins dix heures par semaine hors de leur lieu principal d'emploi, y compris avec leur téléphone mobile durant leurs déplacements; **[4]**

■ **Le télétravail supplémentaire** qu'exécutent les personnes qui travaillent ponctuellement à domicile le soir ou le week-end, généralement pour tenir les échéances durant les périodes de pointe **[4]**

4. Le rôle de travail à distance

- Le télétravail correspond à une organisation du travail qui permet aux salariés de mener des activités en dehors des locaux professionnels.
- Le télétravail est une méthode d'organisation pour réduire les dépenses de l'entreprise.
- Le télétravail est le moyen le plus efficace pour le gouvernement de lutter contre la propagation et la propagation du COVID-19 tout en maintenant la continuité des activités à mesure que les connexions physiques diminuent.

- Hors contexte sanitaire, faire du télétravail permet parfois de concilier vie professionnelle et vie personnelle, d'avoir une qualité de travail qui peut être supérieure et d'aménager le travail en s'adaptant à la situation du salarié

5. Les avantages et les inconvénients

5.1-Les avantages

5.1.1- Pour les employés :

- Meilleur équilibre entre vie professionnelle et vie privée
- Réduisez les embouteillages entre la maison et le bureau
- Possibilité de trouver du travail à distance en dehors de votre environnement sans entrave
- Bien qu'ils aient une mobilité limitée en raison d'une maladie ou d'un handicap, ils peuvent toujours travailler à domicile.
- Réduire les déplacements (contribution environnementale possible)
- Flexibilité des horaires
- Augmentation potentielle de la motivation et de la productivité.
- Une plus grande autonomie de travail
- relâcher la pression **[5]**

5.1.2- Pour les employeurs :

- Organisation de l'espace (décoration, concentration, décentralisation, délocalisation)
- Économisez sur les frais de déplacement entre le domicile et le travail
- Augmenter l'attractivité (l'image) de l'entreprise permet d'embaucher et de fidéliser les salariés.
- Augmentation de la productivité
- Plus de flexibilité dans les activités et services commerciaux **[5]**

5.1.3- Pour les entreprises :

- d’élargir leur vivier de travailleurs qualifiés
- Limiter l'impact de l'infection (les employés malades peuvent continuer à travailler en dehors des lieux de travail normaux)
- Réduisez les dépenses en offrant un espace suffisant pour tous les employés, en particulier les dépenses immobilières.
- Augmentation de la productivité
- Réduire la consommation d'énergie et les émissions de dioxyde de carbone

- Satisfaire aux exigences législatives concernant l'emploi des personnes handicapées et d'autres groupes défavorisés
- Réduire l'absentéisme et les fluctuations;
- Améliorer le moral des employés ;
- Compléter la stratégie de développement durable de l'entreprise
- Mieux gérer les événements qui s'étendent sur plusieurs fuseaux horaires
- Améliorez leur adaptabilité culturelle**. [4]**

5.2- Les inconvénients

5.2.1- Pour les employés :

- Les travailleurs sont plus isolés socialement
- Diminution de l'information et de la communication formelles et informelles
- Le risque de manquer des opportunités de carrière.
- Il peut y avoir des conflits entre les rôles professionnels et personnels
- Le besoin «d'auto-motivation» et de gestion du temps s'est accru. **[5]**

5.2.2- Pour les employeurs :

- Contrôle plus délicat, nécessité de trouver de nouvelles formes de gestion
- Augmentation des coûts d'accompagnement, de recrutement, etc.
- Perte potentielle d'engagement et de loyauté
- Difficulté de la communication interne
- La socialisation des nouveaux employés devient plus difficile. **[5]**

5.2.3- Pour les entreprises :

- **Culture d'entreprise:**

La culture d'entreprise est un élément important pour rendre les employeurs bons - la façon dont ils traitent les employés, les récompenses pour le travail acharné, etc. Lorsque vous n'avez pas de collaborateurs dans un certain fuseau horaire, il est difficile pour vos employés de se sentir inclus ou de faire partie de l'équipe. Cela peut nuire à leur éthique professionnelle et à leur engagement envers votre produit ou votre entreprise**.**

- **L’exclusion:**

En travaillant à domicile, les employés peuvent se sentir exclus de l'équipe, ce qui peut entraîner une baisse de leurs performances et rendre plus difficile la réalisation des objectifs de l'entreprise.

- **Responsabilité :**

Le plus grand obstacle à la gestion du travail à distance est la responsabilité. Il est difficile de savoir si votre équipe travaille réellement ou si certaines personnes n'en ont pas profité pour rester à la maison et regarder la télévision.

- **La micro-gestion :**

Lorsque votre équipe a besoin d'une surveillance continue, la micro-gestion du travail à distance peut échouer, ce qui est beaucoup plus difficile, voire impossible, pour les travailleurs à distance.

6. Modélisation du concept de télétravail

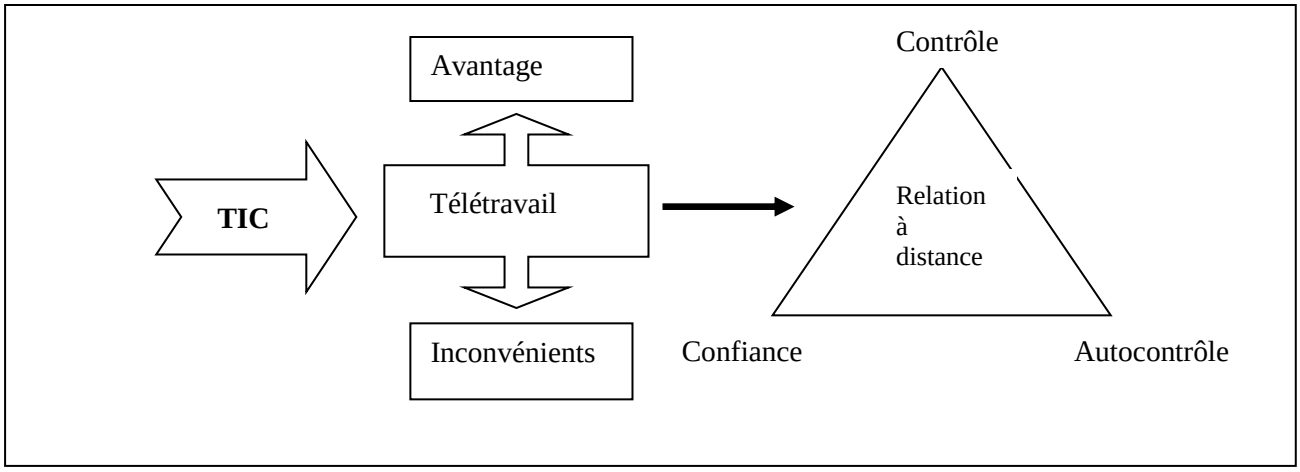


Figure 01 : Modélisation du concept de télétravail

Par souci de clarté, nous proposons un modèle pour le concept de télétravail. Le modèle illustre les concepts suivants: technologies de l'information et de la communication (TIC), avantages et inconvénients, que nous pouvons classer comme influence, relation de distance et trois variables manquantes. Pas encore mentionné: contrôle, confiance en soi et maîtrise de soi. Jusqu'à présent, nous avons souligné le rôle principal des TIC dans le télétravail. Sans eux, cette forme de travail serait difficile. De plus, nous sommes préoccupés par les conséquences positives et négatives de cette forme de travail, qui ont de nombreuses conséquences, de sorte que le travail à distance est

toujours controversé. D'autres s'inquiètent des inconvénients que cela entraîne et sont sceptiques quant à son intégration dans la société. Cependant, l'aspect le plus intéressant de cette recherche est la relation à distance que nous devons maintenant entretenir avec les subordonnés. **[1]**

7. La mise en place de télétravail

- la protection des données et la protection informatique (sécurité informatique)
- le cadre de travail de télétravail (bon pratique, équipement, accord/charte)
- les rythmes de travail (gestion des pauses ; droit de déconnexion)
- la formation (de manager de télétravailleurs)
- l'organisation du travail
- les éléments financiers (cout additionnel, cout moins)
- le suivi individuel (formation, suivi la performance)
- la cohésion d'entreprise (la communication)
- les parties prenantes (salarié futur, fournisseur, client)
- le management d'équipe (la dynamique d'équipe).
- l'infrastructure technique (maintenance, signature électronique) **[6]**

II. Le télé-recrutement

1. Définition

Le **télé-recrutement** (recrutement à distance) fait partie du travail à distance. Les employeurs utilisent la technologie moderne pour sélectionner les employés au bureau ou à domicile selon leurs propres conditions, sans avoir besoin d'entretiens, pour réduire le temps et les coûts, et pour éviter propagation du nouveau coronavirus.

« Le télé-recrutement est une opportunité pour réduire les coûts et les délais de recrutement ».

« Dans la plupart des cas, les solutions de recrutement à distance peuvent optimiser la gestion des applications et devraient permettre de gagner plus de 80 % de temps.».

2. Le rôle de télé-recrutement

- Faciliter le processus de sélection des travailleurs.
- Sélection selon les critères requis.
- Aide à réduire la propagation de l'épidémie COVID-19.
- Réduire les coûts et raccourcir les délais.

- Améliorer la gestion des travaux.

3. Les avantages et les inconvénients

3.1. Les avantages

- Transparence et crédibilité dans la sélection des travailleurs (élimination de la corruption, la médiation, …).
- Offrir aux personnes ayant des besoins particuliers la possibilité de trouver un travail convenable sans déménager.
- Eviter les rassemblements en entreprise pour un entretien d'embauche.
- Minimiser le temps.
- Faciliter la sélection pour l'employeur et assurer une sélection appropriée selon des conditions appropriées.
- Encourager le Télétravail.
- Prévention du virus COVID-19.

3.2. Les inconvénients

- Incertitude sur l'état de santé (physique et psychologique) du travailleur
- Incapacité de vérifier l'authenticité des informations (usurpation d'identité)
- Incapacité de savoir si le travailleur a un casier judiciaire

4. Le mécanisme de télé-recrutement

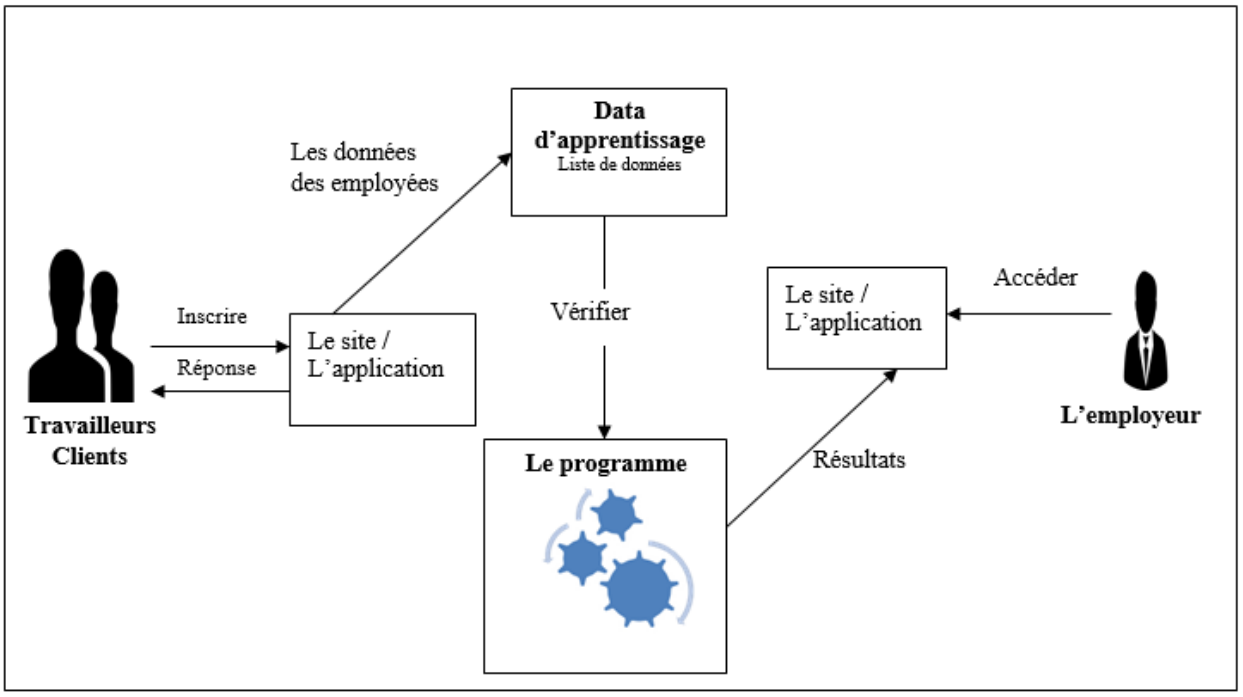


***Figure 02*:** le mécanisme de télé-recrutement

A travers un site internet ou une application, n'importe qui peut se connecter et mettre ses données, âge, adresse, expérience, etc...

Avec la même application /site qui analyse ses données et les présente aux employeurs pour prendre une décision appropriée concernant le travailleur, qu'il soit acceptable ou non, afin de faciliter les processus de recrutement et de faciliter le travail pour les employeurs.

CONCLUSION

Avec les événements COVID-19 qui ont obligé le monde à travailler à distance, ainsi que le recrutement à distance via des sites Web ou des applications qui permettent aux employeurs de choisir à distance dans des conditions spécifiques, afin d'éviter de mener des entretiens d'embauche pour un grand nombre de personnes afin de faciliter la sélection et l'évitement de La propagation de l'épidémie de Corona et assurer le confort du travailleur et de l'employeur en réduisant les coûts de transport et autres.

A travers cette fin de notre projet d'étude, l'employeur sélectionne ses salariés à distance sans entretien selon les conditions qu'il se fixe (âge, spécialité, sexe...etc.), la sélection dépend des technologies de l'information, notamment data Mining, nous utiliserons la classification et l'algorithme de sélection qui nous permet d'atteindre l'objectif souhaité, nous entrerons plus en détail dans le deuxième chapitre.

CHAPITRE 02 : CLASSIFICATION SUPERVISEE ET NON SUPERVISEE

CHAPITRE 2

Classification supervisée et non supervisée

Sommaire

INTRODUCTION

Les environnements informatiques modernes permettent de générer et d'archiver de grandes quantités de données numériques, en utilisant au moins 30 ans de connaissances sur les objets écrits avec eux. **[7]**

Cependant, le problème de l'acquisition de ces connaissances va bien au-delà des compétences analytiques d'une personne, car ces éléments de connaissance ou « particules » sont répartis sur de gros volumes de données souvent complexes. En revanche, depuis le début des années 1990, il n'a pas su en tirer pleinement parti pour favoriser le développement de nouvelles disciplines scientifiques. "Data Mining". .Communauté de base de données. **[7]**

Du point de vue du positionnement scientifique, la fouille de données (DM) est à l'interface de l'informatique et des statistiques. La méthode DM utilise à la fois des statistiques et des méthodes d'apprentissage automatique. Son but est de développer des algorithmes qui peuvent apprendre à résoudre des problèmes à travers des exemples de solutions pour créer un modèle qui représente les données, mais en même temps, il est robuste aux erreurs de développement. Il est utilisé pour traiter de grandes quantités de données et considérer Un algorithme efficace pour des problèmes plus ouverts que ceux proposés précédemment dans l'analyse de données statistiques **[7]**.Tâches et méthodes d'exploration des données (en particulier comme suit : règles d'association, réseaux de neurones artificiels, exploration de réseau, arbres de décision).

1. Data Mining

1.1. Définition

Ils existent plusieurs définitions de data Mining :

• Le Data Mining est l'analyse de grandes ensembles de données observationnelles pour découvrir des nouvelles relations entre elles et de les reformuler afin de les rendre plus utilisables de la part de ses propriétaires **[8]**.

• Le Data Mining est un domaine interdisciplinaire utilisant dans le même temps des techniques d'apprentissage automatiques, de reconnaissance des formes, des statistiques, des bases de données et de visualisation pour déterminer les manières d'extraction des informations de très grandes bases de données **[9].**

• Le Data Mining est un processus inductif, itératif et interactif dont l'objectif est la découverte de modèles de données valides, nouveaux, utiles et compréhensibles dans de larges Bases de Données **[10].**

1.2. Les étapes du processus de data Mining

1) Collecte des données : la combinaison de plusieurs sources de données, souvent hétérogènes, dans une base de données **[11] [12].**

2) Nettoyage des données : la normalisation des données : l'élimination du bruit (les attributs ayant des valeurs invalides et les attributs sans valeurs) **[11] [12].**

3) Sélection des données : Sélectionner de la base de données les attributs utiles pour une tâche particulière du data Mining **[13].**

4) Transformation des données : le processus de transformation des structures des attributs pour être adéquates à la procédure d'extraction des informations **[14].**

5) Extraction des informations (Data Mining): l'application de quelques algorithmes du Data Mining sur les données produites par l'étape précédente (Knowledge Discovery in Databases, ou KDD) **[12] [13].**

6) Visualisation des données : l'utilisation des techniques de visualisation (histogramme, camembert, arbre, visualisation 3D) pour exploration interactive de données (la découverte des modèles de données) **[14] [12].**

7) Evaluation des modèles : l'identification des modèles strictement intéressants en se basant sur des mesures données **[11].**

1.3. Les tâches du Data Mining

Contrairement aux idées reçues, l'exploration de données n'est pas une panacée qui peut résoudre

toutes les difficultés ou besoins des entreprises, mais elle peut formaliser de nombreux problèmes intellectuels, économiques ou commerciaux en un seul. Les tâches suivantes :

1) Classification.
2) Estimation.
3) Prédiction.
4) Groupement par similitudes.
5) Segmentation (ou clustérisassions).
6) Description.

Aucun des outils d'exploration de données décrits ci-dessous ne peut résoudre divers problèmes. Chaque outil a ses propres caractéristiques et utilisations. **[15]**

2. Machine Learning

2.1. Définition

L'apprentissage automatique (Machine Learning) est une région de sondage en télétraitement qui claustration des méthodes d'identification et de mise en œuvre de systèmes et algorithmes par auxquelles une poupée peut apprendre, ce région a journellement été spectateur à l'intelligence artificielle et plus typiquement l'intelligence computationnelle. L'intelligence computationnelle est une accoutumance d'analyse de atout qui gouge transport la élaboration inné de modèles analytiques. Autrement dit, permettant à une poupée d'élaborer des concepts, d'évaluer, vivre des décisions et arranger les options futures. **[16]**

L'ensemble du processus d'apprentissage nécessite un ensemble de données comme suit :

- Ensemble de données pour l'entrainement : c'est la base de connaissance utilisée pour entrainer, notre l'algorithme d'apprentissage, pendant cette phase, les paramètres du modèle peuvent être réglés (ajustés) en fonction des performances obtenues.
- Ensemble de données pour le test : cela est utilisé juste pour évaluer les performances du modèle sur les données non- vues.

La théorie de l'apprentissage utilise des outils mathématiques dérivés de la théorie des probabilités et de la théorie de l'information. Cela vous permet d'évaluer l'optimalité de certaines méthodes par rapport aux autres. On peut citer trois types d'algorithme d'apprentissage automatique :

- Apprentissage supervisé.
- Apprentissage non supervisé.
- Apprentissage par renforcement.

2.1.1. Apprentissage supervisé

L'apprentissage supervisé est la tâche d'apprentissage automatique la plus simple et la plus connue. Il est basé sur un certain nombre d'exemples pré classifiés, dans lesquels est connu à priori la catégorie à laquelle appartient chacune des entrées utilisées comme exemples. Dans ce cas, la question cruciale est le problème de généralisation, après l'analyse d'un échantillon d'exemples, le système devrait produire un modèle qui devrait fonctionner pour toutes les entrées possibles. L'ensemble de données pour l'entrainement, est constitué de données étiquetées, c'est-à-dire d'objets et de leurs classes associées. Cet ensemble d'exemples étiquetés constitue donc l'ensemble d'apprentissage. Afin de mieux comprendre ce concept, prenons un exemple : un utilisateur reçoit chaque jour un grand nombre d'e-mails, certains sont des e-mails d'entreprises importants et d'autres sont des e-mails indésirables non sollicités ou des spam. Un algorithme supervisé sera présenté avec un grand nombre d'e-mails qui ont déjà été étiquetés par l'utilisateur comme spam ou non spam. L'algorithme fonctionnera sur toutes les données étiquetées, faire des prédictions sur l'e-mail et voir si c'est un spam ou non. Cela signifie que l'algorithme examinera chaque exemple et fera une prédiction pour chacun pour savoir si l'e-mail est un spam ou pas. La première fois, l'algorithme fonctionne sur toutes les données non étiquetées, la plupart des e-mails seront mal étiquetés car il peut fonctionner assez mal au début. Cependant, après chaque exécution, l'algorithme compare sa prédiction au résultat souhaité (l'étiquette). Au fur et à mesure, l'algorithme apprendra à améliorer ses performances et sa précision.

Dans l'exemple que nous avons utilisé, nous avons décrit un processus dans lequel un algorithme apprend à partir de données étiquetées (emails qui ont été catégorisés comme spam ou non-spam). Dans certains cas, le résultat n'est pas nécessairement discret et il se peut que nous n'ayons pas un nombre fini de classes dans lesquelles classer nos données. Par exemple, nous essayons peut-être de prédire l'espérance de vie d'un groupe de personnes en fonction de paramètres de santé préétablis. Dans ce cas, comme le résultat est une fonction continue (nous pouvons spécifier une espérance de vie comme un nombre réel exprimant le nombre d'années que la personne devrait vivre), nous ne parlons pas d'une tâche de classification mais plutôt d'un problème de régression. Dans un problème de régression, l'ensemble d'apprentissage est une paire formée par un objet et une valeur numérique associée. Il existe plusieurs algorithmes d'apprentissage supervisé qui ont été développés pour la classification et la régression. Parmi tous, les arbres de décision, les règles de décision, les réseaux de neurones et les réseaux bayé siens.

2.1.2. Apprentissage non supervisé

La deuxième classe d'algorithmes d'apprentissage automatique est appelée apprentissage non supervisé, dans ce cas, nous n'étiquetons pas les données au préalable, nous laissons plutôt l'algorithme arriver à sa conclusion. Ce type d'apprentissage est important car il est beaucoup plus commun dans le cerveau humain que l'apprentissage supervisé. Les algorithmes d'apprentissage non supervisé sont particulièrement utilisés dans les problèmes de clustering, dans lesquels, étant donné une collection d'objets, nous voulons être en mesure de comprendre et de montrer leurs relations. Une approche standard consiste à définir une mesure de similarité entre deux objets, puis à rechercher tout groupe d'objets plus similaires les uns aux autres, par rapport aux objets des autres clusters. Par exemple, dans le cas précédent des e-mails spam/ non spam, l'algorithme peut être capable de trouver des éléments communs à tous les spam (par exemple, la présence de mots mal orthographiés). Bien que cela puisse fournir une classification meilleure qu'aléatoire, il n'est pas clair que les spam/non spam puissent être facilement séparés. **[16] [17] [18]**

2.1.3. Apprentissage par renforcement

L'apprentissage par renforcement est une approche de l'intelligence artificielle qui met l'accent sur l'apprentissage du système à travers ses interactions avec l'environnement. Avec l'apprentissage par renforcement, le système adapte ses paramètres en fonction des réactions reçues de l'environnement, qui fournit ensuite un retour d'information sur les décisions prises. Par exemple, un système qui modélise un joueur d'échecs qui utilise le résultat des étapes précédentes pour améliorer ses performances, est un système qui apprend avec le renforcement. La recherche actuelle sur l'apprentissage avec renforcement est hautement interdisciplinaire et comprend des chercheurs spécialisés dans les algorithmes génétiques, les réseaux de neurones, la psychologie et les techniques de contrôle. **[16] [18]**

I. LA CLASSIFICATION

1. Définition

La classification est la responsabilité la davantage commune du Data Mining et qui semble concerner une Obligation humaine. Afin de comprendre notre vie quotidienne, certains sommeils régulièrement Classifiés, catégorisés et évalués **[19]**.

La classification consiste à étudier les caractéristiques d'un nouvel objet pour lui Attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant. L'objectif est de créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées **[20].**

Quelques exemples de l'utilisation des tâches de classification dans les domaines de recherche et commerce sont les suivants :

• Déterminer si l'utilisation d'une carte de crédit est frauduleuse.

• Diagnostiquant si une certaine maladie est présente **[21].**

• Déterminer quels numéros de téléphone correspondent aux fax **[19].**

1.2. La différence entre la classification supervisée et la classification non supervisée

La antinomie là-dedans lequel les nettoyages situations est la science des classes pendant le cas non supervisée, on article une fouille à l'absolu ensuite en qu'en supervisée on a un patron d'exorde En supervisée on connait un informateur de performance : le montant de mal classé, cependant en non supervisée? Difficile de valider les résultats.

1.3. Les étapes d'une classification

1. Choix des données.
2. Calcul des similarités entre les n individus à partir des données initiales.
3. Choix d'un algorithme de classification et exécution.
4. L'interprétation des résultats. **[22]**

2. Classification supervisée et non supervisée

2.1. Classification Supervisée

2.1.1. Définition

La classification supervisée est une technique largement utilisée avec différentes applications dans la vie réelle **[23].** Elle permet de générer des règles de classification (modèle) à partir d'un jeu données classées à priori et d'un algorithme d'apprentissage automatique adéquat. Ces règles seront utilisées pour classer les nouvelles instances.

L'apprentissage supervisé est généralement effectué dans le contexte de la classification et de la régression **[24]**.

- **Classification**: La classification peut être utilisée lorsque la classe est discrète et que l'objectif est de prédire l'une des valeurs mutuellement exclusives dans la variable cible. Un exemple est la cote de crédit, où la prédiction finale est de savoir si l'individu est responsable du crédit. Les algorithmes populaires sont les arbres de décision, les classificateurs naïfs de Bayes, les machines vectorielles auxiliaires, les réseaux de neurones et les méthodes d'ensemble. **[24]**
- **Régression**: La régression traite de la variable cible continue, contrairement à la classification, qui fonctionne avec une variable cible discrète. Par exemple, pour prévoir la température extérieure des prochains jours, on utilisera la régression ; tandis que la classification sera utilisée pour prédire s'il pleuvra ou non. De manière générale, la régression est un processus qui estime la relation entre les caractéristiques, c'est-à-dire la manière dont la variation d'une caractéristique modifie la variable cible. **[24]**

2.1.2. Les Algorithmes De Classification Supervisée

Selon le but et le type de formation, il existe différents algorithmes d'extraction de connaissances. Les deux algorithmes d'apprentissage peuvent être différents dans le type d'apprentissage, c'est-à-dire qu'ils sont différents dans le type d'objectif, et ils sont également différents dans la façon dont ils apprennent. Nous examinerons quelques méthodes populaires utilisées dans l'apprentissage automatique.

- Arbres de décision
- K Nearest Neighbors (k plus proches voisins.)
- SVC linéaire (classificateur de vecteur de support)
- Régression logistique

- Naïve Bayes
- Les réseaux de neurones
- Régression linéaire
- Support Vector Machine (SVM)
- Arbres de régression

A. Arbres de décision

1. Définition

Les arbres de décisions sont des outils d'aide à los Angeles décision qui permettent selon des variables discriminantes de répartir une populace d'individus en groupes homogènes en fonction d'un objectif connu. Les arbres de décision sont des outils puissants et populaires pour los Angeles type et los Angeles prédiction. Un arbre de décision permet à partir des données connues sur le problème de donner des prédictions par réduction, niveau par niveau, du domaine des solutions. Chaque nœud interne d'un arbre de décision permet de répartir les éléments à classifier de façon homogène entre ses différents fils en portant sur une variable discriminante de ces éléments. Les branches qui représentent les liaisons entre un nœud et ses fils sont les valeurs discriminantes de los Angeles variable du nœud. Et en fin, les feuilles d'un arbre de décision représentent les résultats de los Angeles prédiction des données à classifier **[25]**

2. Algorithme d'apprentissage d'arbre de décision

Données : un échantillon Ω de m enregistrements étiquetés

Initialisation : arbre vide ; nœud courant : racine ; échantillon courant : Ω

Répéter

Décider si le nœud courant est terminal

Si le nœud courant est terminal alors

Étiqueter le nœud courant par une feuille

Sinon

Sélectionner un test et créer le sous-arbre

FinSi

Nœud courant : un nœud non encore étudié

Échantillon courant : échantillon atteignant le nœud courant

Jusqu'à

Production d'un arbre de décision

Sortie : arbre de décision **[26]**

3. Exemple

La **figure 3** illustre un exemple simple d'arbre de décision qui est présenté dans l'ouvrage de Quinlan **[27]** .Ici on cherche à classer une population d'individus en deux classes par rapport à un jeu {jouer, ne pas jouer} à partir des prévisions météorologiques (***Tableau 1***).

Temps	Température	Humidité	Vent	Tennis ?
Ensoleillé	Chaude	Elevée	Faux	Non
Ensoleillé	Chaude	Elevée	Vrai	Non
Couvert	Chaude	Elevée	Faux	Oui
Pluvieux	Modérée	Elevée	Faux	Oui
Pluvieux	Fraîche	Normale	Faux	Oui
Pluvieux	Fraîche	Normale	Vrai	Non
Couvert	Fraîche	Normale	Vrai	Oui
Ensoleillé	Modérée	Elevée	Faux	Non
Ensoleillé	Fraîche	Normale	Faux	Oui
Pluvieux	Modérée	Normale	Faux	Oui
Ensoleillé	Modérée	Normale	Vrai	Oui

Tableau 1 : Données "weather" **[27]**

Dans ce tableau on a : Temps {ensoleillé, couvert, pluvieux}

Température {chaud, modéré, frais}

Humidité {élevée, normale}

Vent {VRAI, FAUX}

Alors l'arbre dans cet exemple est :

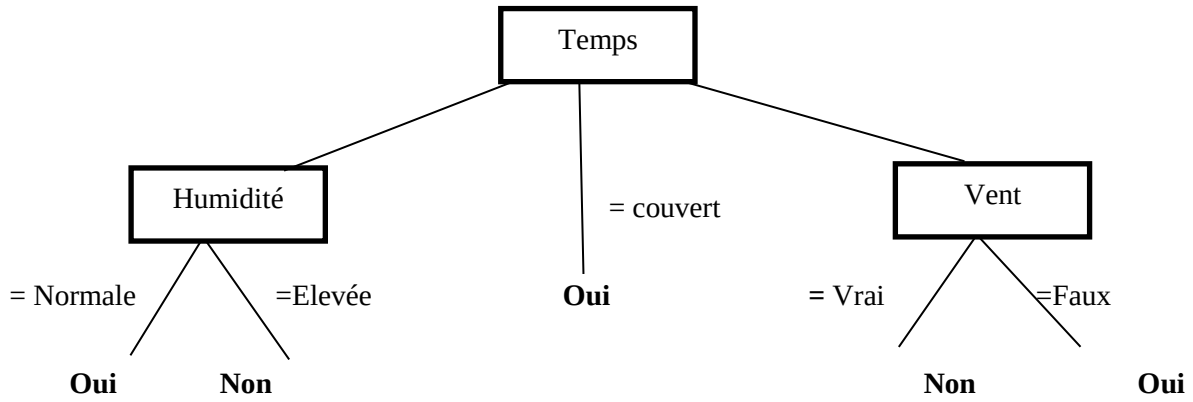


Figure 3 : Exemple d'arbre de décision sur les données "weather" **[27]**

B. KNN (K- plus proche voisin)

1. Définition

L'algorithme des *k- principalement proches voisins*, traduction de k-*Nearest Neighbor (kNN)* en anglais, est une méthode d'apprentissage à base d'instances. Il ne comporte pas de phase d'apprentissage en tant que telle. Les documents faisant partie de l'ensemble d'apprentissage sont seulement enregistrés. Lorsqu'un nouveau document à classer arrive, il est comparé aux documents d'apprentissage à l'aide d'une mesure de similarité. Ses *k* plus proches voisins sont alors considérés : on observe leur catégorie et celle qui revient le plus parmi les voisins est affectée au document à classer. C'est là une version de base de l'algorithme que l'on peut raffiner. **[28]**

2. Le principe de la méthode KNN

Le principe de la méthode KNN est de trouver k plus proches voisins, à partir de l'échantillon d'apprentissage, à une nouvelle instance qu'on cherche à classer. La classe de la nouvelle instance est la classe majoritaire (la plus représentée) parmi ces k voisins. Dans le cas d'une régression, la valeur de sortie est une valeur continue qui peut être, par exemple, la moyenne des valeurs des k voisins. Il existe plusieurs fonctions pour calculer la distance entre deux voisins, notamment, la distance euclidienne, la distance de Manhattan, la distance de Minkowski, la distance de Jaccard, etc. Dans ce qui suit, noud définissons la distance euclidienne et la distance de Manhattan**. [29]**

3. Ecriture algorithmique

L'algorithme 1 ci-après décrit les différentes étapes pour classer un nouveau document avec kNN en utilisant une mesure de similarité (ex. cosinus) pour sélectionner ses voisins les plus proches et le vote majoritaire pour calculer sa classe **[28]**

Début

Paramètre : Le nombre k de voisins

Données : Un échantillon de n exemples d'apprentissage Ω = (ω 1,...., ωn)

La classe d'un exemple ω est Y(ω), Y= {C1,C2, …Cm}

Entrée : un enregistrement X

Pour chaque exemple ω **faire**

Calculer la distance d(X,ω) ;

Fin

KNN= les k plus proches voisins de X qui minimise la distance d ;

Pour chaque {ω∈KNN} **faire**

Calculer les scores des classes ;

Fin

Attribuer Y(X) à la classe ayant le plus grand score;

Sortie : la classe de X est Y(X)=Cj ;

Fin. [26]

4. Exemple

Un exemple de classification par kNN est présenté dans la figure *5* Dans ce schéma, la première classe est représentée par un cercle et la deuxième par un carreau. La nouvelle instance à classifier est sous la forme d'une croix. En utilisant un classifieur 3NN (kNN avec $k = 3$), les 3 plus proches voisins de la nouvelle instance appartiennent. **[30]**

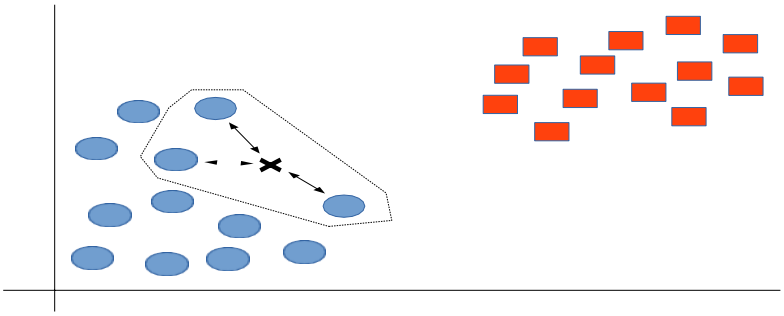


Figure 4: Un exemple de classification par 3NN

(a) k trop petite (3NN) : sur-apprentissage

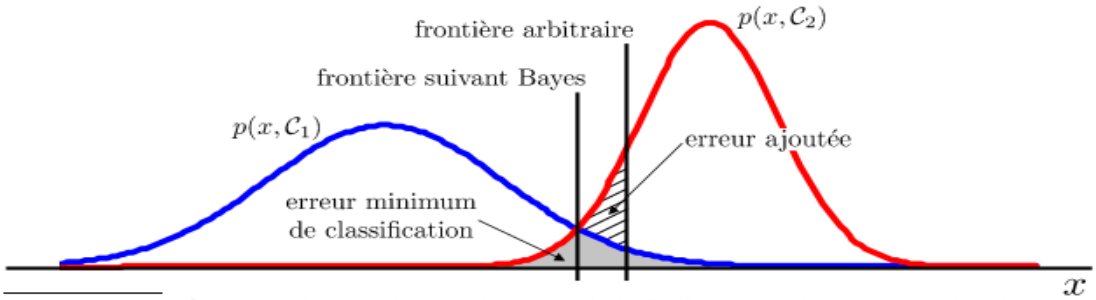
(b) k trop grande (8NN) : sous-apprentissage

Figure 5 : Importance du choix de la valeur k dans la classification par KNN **[30]**

C. Naïve Bayes

1. Définition

Naïve Bayes est un classificateur probabiliste simple. Il calcule un ensemble de probabilités en comptant la fréquence et les combinaisons de valeurs dans un jeu de données. L'algorithme utilise le théorème de Bayes et suppose que tous les attributs sont indépendants compte tenu de la valeur de la variable classe. Cette hypothèse d'indépendance conditionnelle est rarement valable dans les applications du monde réel, d'où la caractérisation naïve. Cependant, l'algorithme tend à bien fonctionner et à apprendre rapidement dans divers problèmes de classification supervisée. **[29]**



Note : La surface grisée représente la probabilité d'erreur obtenue par la règle de décision de Bayes. La surface hachurée représente la probabilité d'erreur ajoutée à la précédente par le choix d'une autre frontière de décision.

Figure 6: Règle de décision de Bayes pour un problème à deux classes. **[31]**

D. Les réseaux de neurones

1. Définition

Le réseau neuronal formel ou artificiel (ARN) est un modèle mathématique qui simule la structure et la fonction des réseaux neuronaux biologiques. Ce sont des outils largement utilisés pour la classification, l'évaluation, la prévision et la segmentation. Vous pouvez également trouver la même solution que Bayésien., pas de conditions particulières **[32] [33]** En fait, nous allons prouver que la sortie du réseau de neurones est une estimation de la probabilité postérieure des membres de la classe. Dans ce chapitre, nous sommes limités aux réseaux de neurones conçus pour les tâches de scoring et de classification, tels que les perceptrons multicouches (PMC) ou les perceptrons multicouches (MLP). Lors de l'introduction des réseaux de neurones formels, nous commencerons par quelques définitions hiérarchiques des réseaux de neurones, puis nous présenterons l'algorithme d'apprentissage utilisé pour entraîner de tels réseaux : l'algorithme de rétro propagation de gradient. **[26]**

1.1. Le neurone biologique

Un neurone est une cellule constituée principalement de trois parties **(Figure 7)** : ce sont les dendrites, le soma et l'axone. Les dendrites collectent les signaux venant d'autrescellules au niveau de points de contact avec les autres neurones appelés synapses.

L'information est ensuite acheminée vers le corps cellulaire ou soma qui recueille et concentre l'ensemble des informations reçues par les dendrites. Les réseaux de

neurones artificiels sont des architectures artificielles inspirées à partir d'un tel fonctionnement. **[26]**

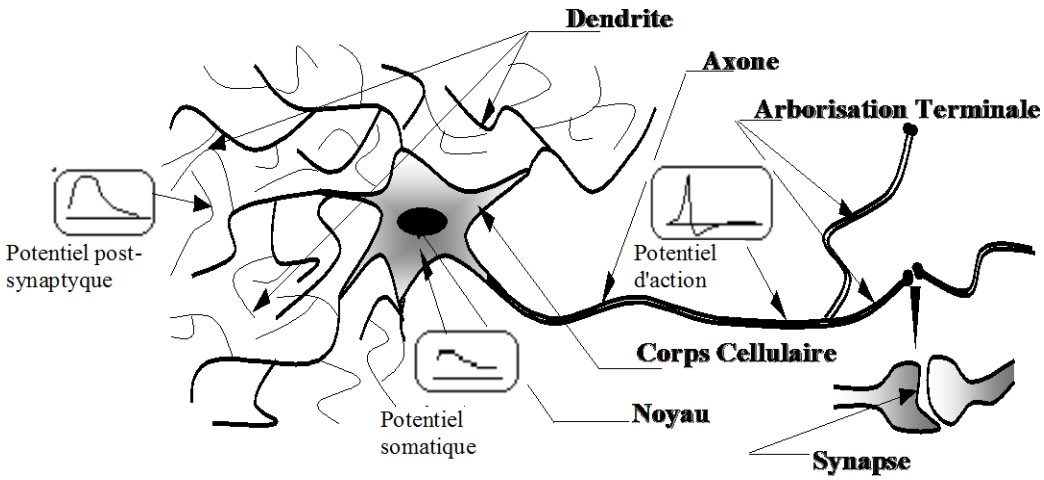


Figure 7: Neurone typique de vertébré **[26]**

1.2. Le neurone formel

Dès l'établissement du fonctionnement réel d'un neurone biologique, plusieurs modèles ont été proposés, dont le but principal est de refléter ce fonctionnement. Le plus important est celui de MC Culloch et Pitts établit en 1943**. [26]**

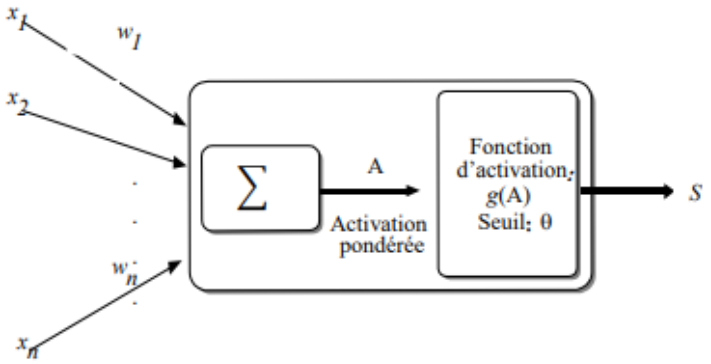


Figure 8*:* Neurone modélisation générale **[26]**

2. Exemple

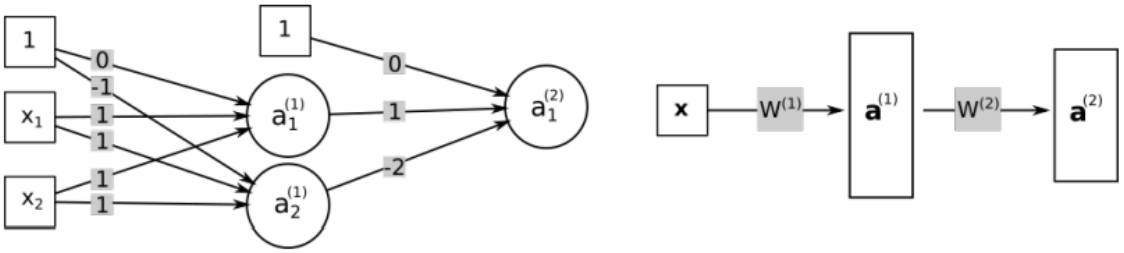


Figure 9: Réseau de neurones représentant la fonction OU EXCLUSIF avec deux représentations graphiques. À gauche, chaque neurone est représenté par un cercle. Les poids sont représentés sur les connexions entre les neurones. De même, les biais représentés par la connexion entre une constante (carrés) et chaque neurone. À droite, dans ce style graphique, chaque couche est représentée par un rectangle. Les matrices de paramètres peuvent être indiquées sur les connexions entre les couches. L'avantage de cette seconde représentation est d'être plus compact que la première. **[34]**

3. Les trois types de couches de réseau de neurones

- **Couche d'entrée :** les neurones de cette couche reçoivent des valeurs d'entrée du réseau et les transmettent aux neurones cachés. Chaque neurone se voit attribuer une valeur. Donc, il ne plantera pas.
- **Couche cachée :** chaque neurone de cette couche reçoit des informations des couches préc
- édentes, effectue une sommation pondérée, puis la convertit en sa fonction d'activation (généralement une fonction sigmoïde), puis envoie la réponse au neurone dans le yuan d'en-tête suivant. ...
- **Couche de sortie :** Le rôle de la couche cachée est le même, la seule différence entre les deux types de couches est que la sortie des neurones de la couche de sortie n'est distribuée à aucun autre neurone **[15]**

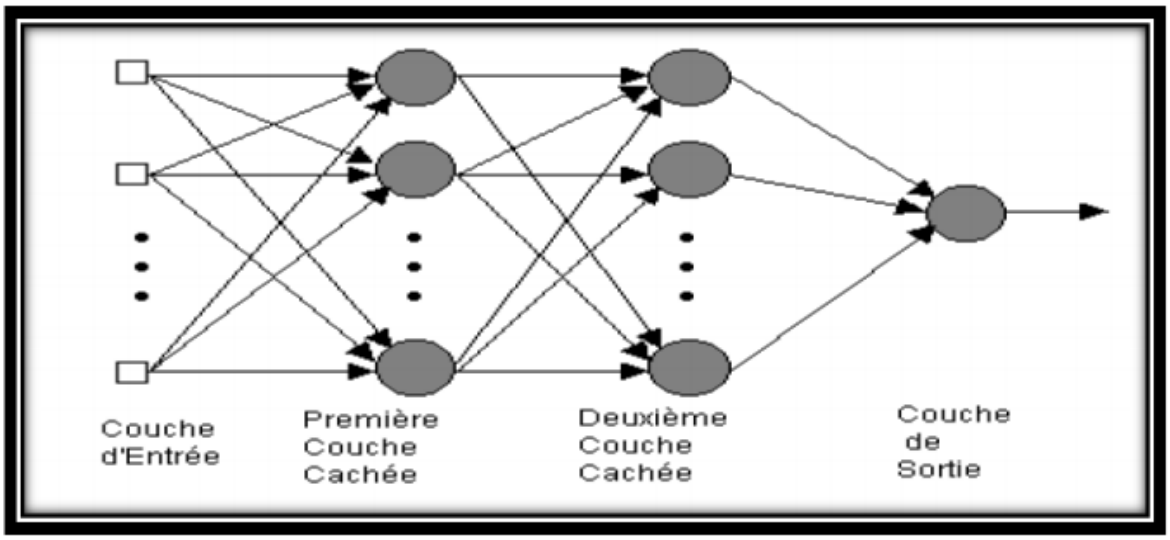


Figure 10 : Types de couches en Réseaux de Neurones **[35]**

E. Support Vector Machine (SVM)

1. Définition

Le classificateur SVM développé par Vladimir Vapnik en 1995 est un classificateur puissant et a fait ses preuves dans divers domaines. Le principe est de projeter des données qui ne peuvent pas être linéairement divisées dans un autre espace de dimension supérieure, où elles peuvent être partagées avec d'autres noyaux. Le but du SVM binaire est de trouver le meilleur hyperplan séparant deux classes en maximisant la distance. Cette distance s'appelle la marge. En utilisant la classification binaire, **Figure 11** L'hyperplan est une ligne droite et le point le plus proche est le seul point utilisé pour définir le jeu, appelé vecteur de support. **[29]**

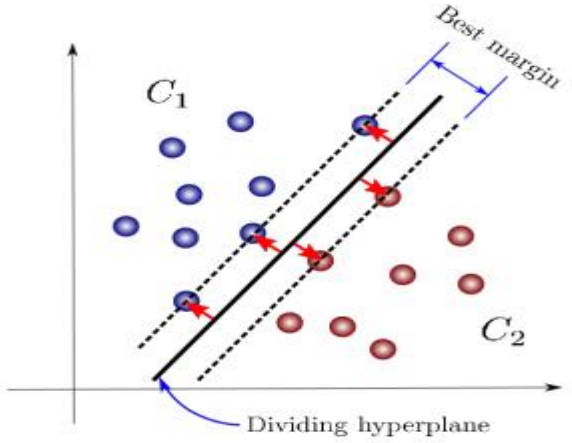


Figure 11: Schéma explicatif d'un séparateur à vaste marge **[36]**

2. Algorithme Général

Algorithm 1: Feature Ranking using SVM

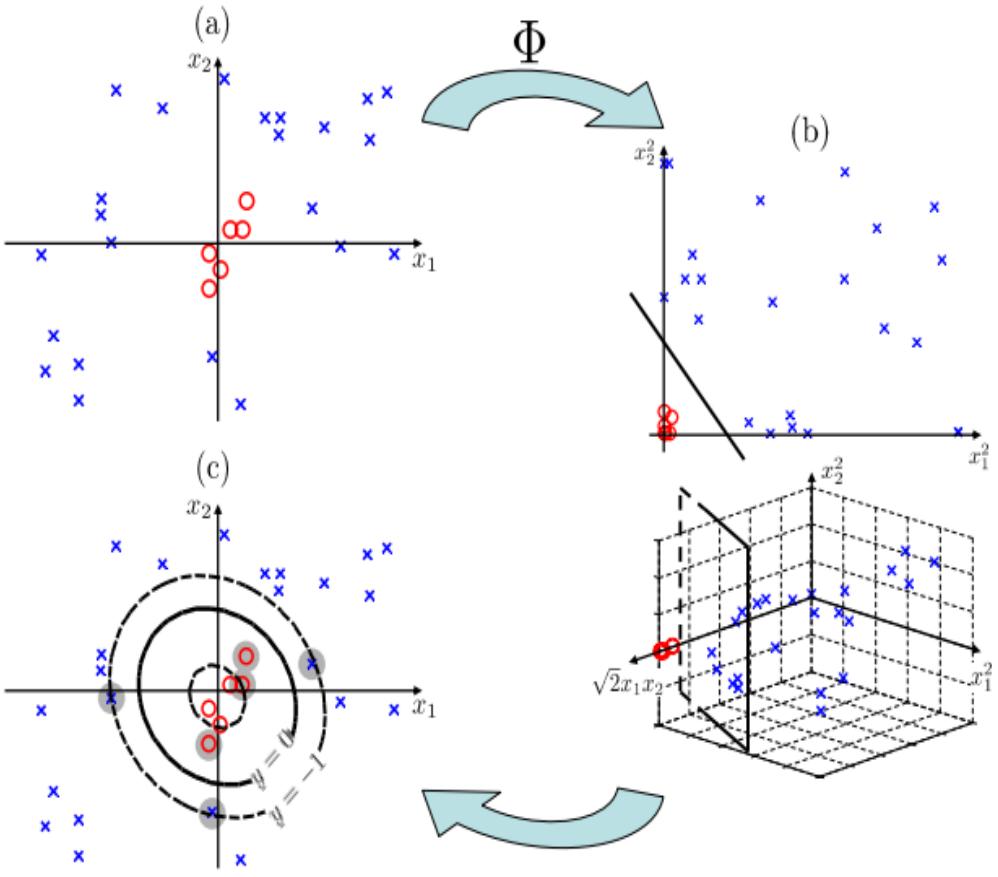
Data: Training sets F and c

Result: Sorted features list $sorted_F$

```
C ← GetParameter(F, c) ;          // Find best parameter for SVM
list_F ← TrainModel(F, C) ;       // Get weight of each feature
sorted_F ← Sort(list_F);
return sorted_F
```

[37]

3. Exemple



Note : (a) Observations représentées dans l'espace d'origine. (b) Observations et hyperplan représentés dans l'espace de redescription (deux et trois dimensions). (c) Retour à l'espace d'origine et expression de la fonction discriminante.

Figure 12 : Illustration de la transformation de l'espace par une fonction noyau (dans le cadre des support vector machines) sur un exemple de discrimination non linéaire. **[38]**

2.1.3. Avantages et Inconvénients des algorithmes de Classification supervisée

Dans le Tableau, nous résumons les avantages et les inconvénients de ces méthodes :

L'algorithme	Les avantages	Les Inconvénients
Arbres de décision	• Les arbres de décision sont capables de produire des règles compréhensibles. • Les arbres de décision effectuent la classification sans exiger beaucoup de calcul • Les arbres de décision sont en mesure de manipuler à la fois les variables continues et catégorielles. **[39]**	• Manque de performance dans le cas de plusieurs classes; les arbres deviennent très complexes et ne sont pas nécessairement optimaux. • Demande beaucoup de temps de calcul lors de la construction (le choix du meilleur partitionnement) et l'élagage (la comparaison de sous-arbres). • Moins bonnes performances concernant les prédictions portant sur des valeurs numériques. **[40]**
KNN	• Il est très simple à mettre en œuvre ; il n'a besoin que de deux paramètres (K et la mesure de similarité). kNN ne fait pas d'apprentissage, il stocke tout simplement tous les exemples d'apprentissage. Il est également bien adapté à la catégorisation multi-classes puisque sa décision de classification est basée sur un voisinage de documents	• Le temps nécessaire pour calculer la similarité est énorme. En pratique, il est impossible de mettre en œuvre l'algorithme pour des dimensions élevées et des corpus d'exemples énormes. En conséquence, le coût de classification devient très élevé pour le plus proche voisin. En outre, le stockage mémoire

	similaires. **[22]**	augmente avec le nombre de documents d'entraînement. **[22]**
Naïve Bayes	• Le classificateur Bayésien naïf est une méthode d'apprentissage populaire pourla classification de textes car il est rapide et facile à mettre en œuvre et donne de bons résultats. • Il est caractérisé par sa robustesse vis-à-vis des donnéesmanquantes, sa vitesse de classification et d'apprentissage. **[22]**	• Une fois que l'espace d'apprentissage devient considérablement large, il est impossible d'interpréter le modèleconstruit. • Son hypothèse naïve affectela qualité des résultats, si les mots sont liés entre eux. **[22]**
Les réseaux de neurones	• Les réseaux de neurones sont des modèles non linéaires, ce qui les rend souples dans la modélisation des relations complexes du monde réel. Les réseaux de neurones sont en mesure d'estimer les probabilités à posteriori, qui fournissent la base pour établir la règle de classification et de l'analyse statistique. **[28]**	• Avec l'augmentation du nombred'entrée et les nœuds cachés, les paramètres nécessaires pour le réseau • neuronal augmentent également, ceci provoque le sur-apprentissage. **[28]**
SVM	• Par rapport à d'autres méthodes, SVM est moins sujet au surentraînement car la complexité du modèle n'a rien à voir avec la dimensionnalité	• L'efficacité dépend du choix de la fonction noyau : Puisqu'il n'y a pas une fonction noyau supérieure aux autres et le temps

	de l'espace attributaire. • Les méthodes basées sur SVM peuvent traiter de grands espaces d'attributs avec une excellente précision de classification. SVM fournit les meilleurs résultats aux niveaux du test et de la formation, est robuste en termes de nombre d'attributs et la classification est très rapide. **[28]**	requis pour l'apprentissage est relativement long parce qu'on est obligé d'expérimenter avec quelques fonctions noyaux candidates pour en trouver la meilleure qui nous convient le plus. • L'algorithme ne résiste pas aux valeurs manquantes puisqu'il a besoin de toutes ces dernières pour faire son calcul. **[28]**

Tableau 2 : Les avantages et les inconvénients d'algorithmes d'apprentissage supervisé

2.1.4. Comparaison des techniques d'apprentissage automatique supervisé

	Arbres de décision	**Les réseaux de neurones**	**Bayésien naïf**	**KNN**	**SVM**	**Apprenants de règles**
Précision en général	**	***	*	**	****	**
Vitesse d'apprentissage en fonction du nombre d'attributs et du nombre d'instances	***	*	****	****	*	**

Vitesse de classification	****	****	****	*	****	****
Tolérance aux valeurs manquantes	***	*	****	*	**	**
Tolérance aux attributs non pertinents	***	*	**	**	****	**
Tolérance aux attributs redondants	**	**	*	**	***	**
Tolérance aux attributs hautement interdépendants (par exemple, problèmes de parité)	**	***	*	*	***	**
Traiter les attributs discrets\binaires\continus	****	*** (pas discret)	*** (pas continu)	*** (pas directement discret)	** (pas discret)	*** (pas directement continu)
Tolérance au bruit	**	**	***	*	**	*
Gérer le risque de sur apprentissage	**	*	***	***	**	**
Tentatives d'apprentissage progressif	**	***	****	****	**	*
Capacité d'explication/	****	*	****	**	*	****

transparence des connaissances/classifications						
Gestion des paramètres du modèle	***	*	****	***	*	***

Tableau 3 : Comparaison des techniques d'apprentissage automatique supervisé [**41]**

2.2. Classification non supervisée

2.2.1. Définition

Les approches descriptives désignent l'ensemble des méthodes permettant d'organiser et d'identifier des tendances dans les données. Loin de la volonté de faire un état de l'art exhaustif de toutes les méthodes descriptives existantes, on s'intéresse dans cette section uniquement à l'une des moyennes utilisée pour décrire les données, à savoir, le clustering. Cette section en présente ses concepts clefs. **[42]**

- **Le Clustering** est une technique permettant de regrouper des instances similaires en clusters en fonction d'une mesure de distance. L'idée principale est de placer des instances similaires (c'est-à-dire proches les unes des autres) dans le même cluster, tout en gardant les points différents (c'est-à-dire les plus éloignés les uns des autres) dans des clusters différents. Un exemple de l'apparence des clusters est illustré dans le diagramme suivant: **[24]**

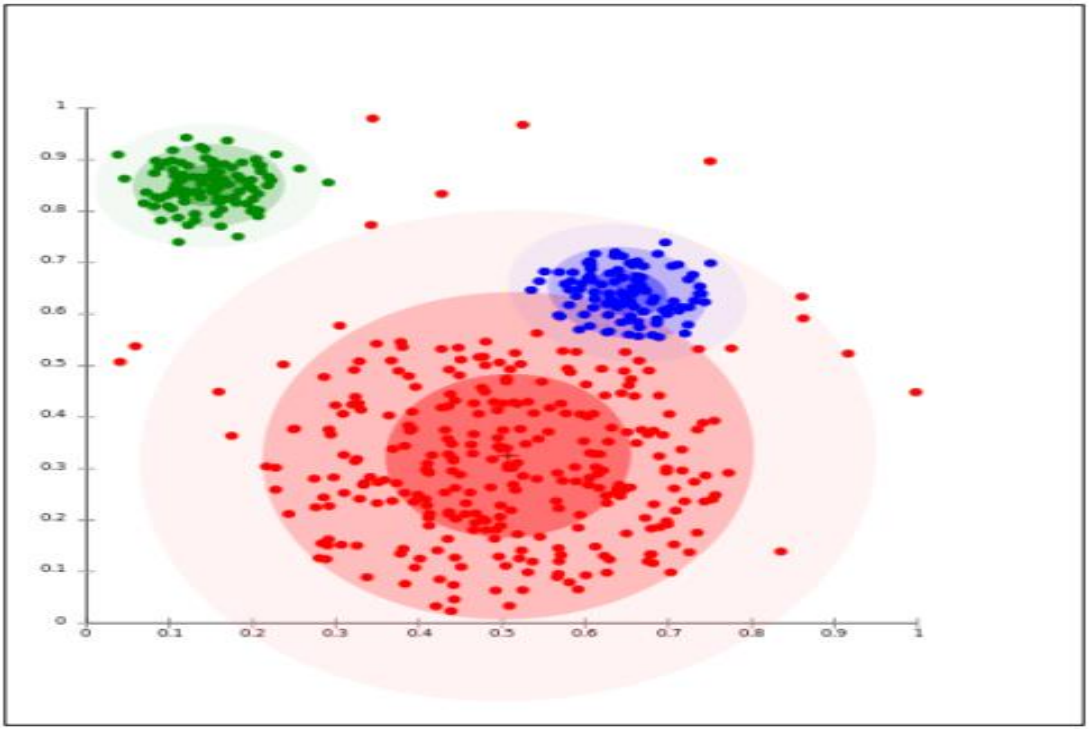


***Figure 13**:* un exemple de l'apparence des clusters est montré **[24]**

2.2.2. Les algorithmes de classification non supervisée

Voici une liste de certains algorithmes d'apprentissage automatique non supervisés:

- K-means clustering
- Fuzzy C-means/ C-moyennes floues
- Dimensionality Reduction (Réduction de la dimensionnalité)
- Principal Component Analysis (Analyse des composants principaux)
- Singular Value Decomposition (Décomposition en valeur singulière)
- Independent Component Analysis (Analyse en composantes indépendantes)
- Distribution models (Modèles de distribution)
- Hierarchical clustering (Classification hiérarchique)

F. K-means clustering(K-moyennes)

1. Définition

L'algorithme k-means est l'un des algorithmes de classification non supervisés les plus populaires et est appelé algorithme du centre mobile. Divisez l'ensemble de données en un nombre donné de K régions. Chaque groupe est représenté par sa valeur moyenne (le point central de la classe), c'est-à-dire que ses coordonnées sont la moyenne arithmétique de chaque valeur de mesure

séparée de tous les points du groupe. **[22]**

2. Principe

L'idée principale est de définir k points focaux arbitraires c1, c2,..., ck (k est le nombre de clusters a priori fixes, et chaque ci représente le centre de la classe).Ces points focaux doivent être placés à des endroits différents. La meilleure option est de les placer aussi loin que possible. L'étape suivante consiste à lier chaque point appartenant à l'ensemble de données au foyer suivant. Le groupe d'individus le plus proche de son ci, le nuage dynamique, est une généralisation de ce principe. Chaque groupe est représenté par un noyau, mais est plus complexe que la moyenne.

S'il n'y a pas de points à traiter, la première étape est terminée et un regroupement précoce est effectué. A ce stade, nous devons recalculer les k nouveaux groupes centroïdes mi qui ont remplacé ci à l'étape précédente (m j est le centroïde de la classe S j calculé avec la nouvelle classe obtenue), puis nous répétons ce processus jusqu'à atteindre un état stable. Aucune amélioration n'est possible, nous pouvons voir les k centres de gravité changer progressivement de position jusqu'à ce que d'autres changements se produisent, en d'autres termes, le centre de gravité ne bougera pas. **[43]**

3. Algorithme général

Choisir k moyennes $c_1, c_2, ..., c_k$ initiales (par exp au hasard)

1. Répéter :

   affectation de chaque point à son cluster le plus proche :

$$S_i^{(t)} = \left\{ x_i : \| x_j - m_i^{(t)} \| \leq \| x_j - m_{i*}^{(t)} \| \ for all \ i^* = 1, ..., k \right\}$$

   mettre à jour la moyenne de chaque cluster

$$m_i^{(t+1)} = \frac{1}{| S_i^{(t)} |} \sum_{x_j \in s_i^{(t)}} x_j$$

2. Jusqu'à : atteindre la convergence quand il n'y a plus de changement.

Fin.

[43]

4. Exemple

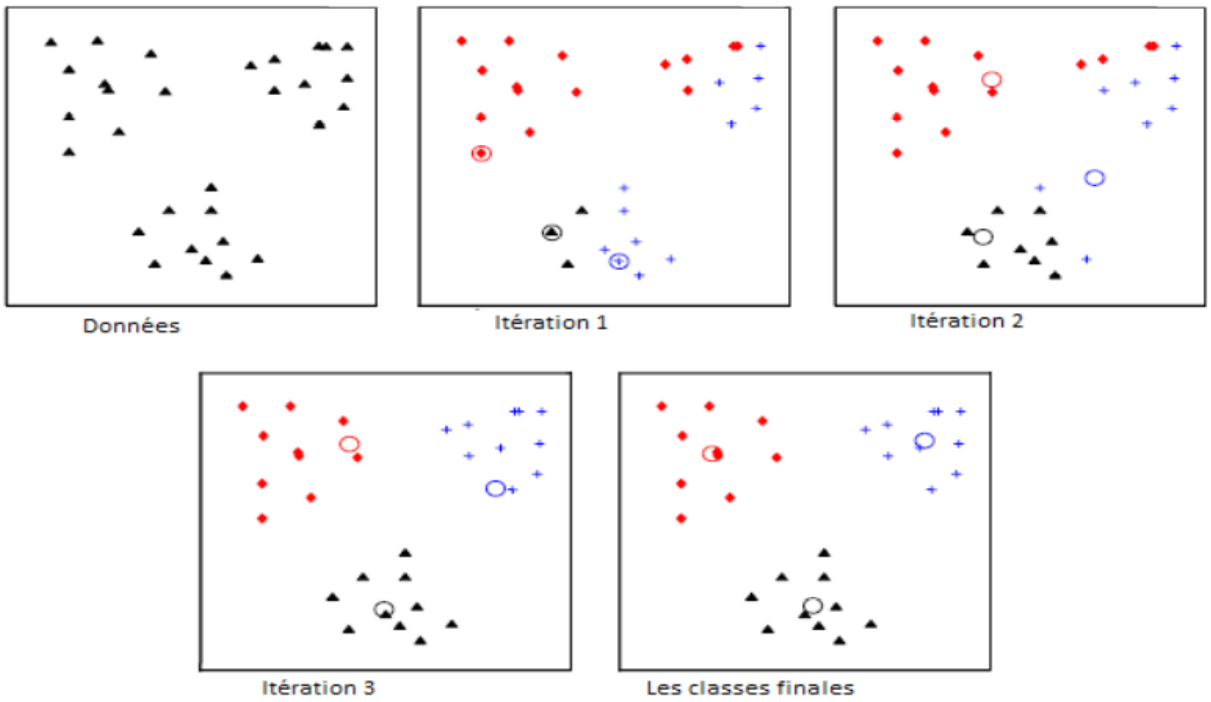


Figure 14 : Illustration de l'algorithme des K-moyennes sur un jeu de données défini dans $\mathbb{R}^2$ contenant 3 classes. L'étape Itération 1 correspond à l'étape d'initialisation, les étapes itérations 2 et 3 définissent l'évolution des centres de classe et l'étape des classes finales présente le résultat de l'algorithme après stabilisation des centres de classe **[45] [44].**

G. Fuzzy C-means/ C-moyennes floues

1. Définition

Une ik =1; ∀ α-coupe de A est le sous-ensemble classique des éléments ayant un degré d'appartenance supérieur ou égal à α α-coupe de A est le sous-ensemble classique des éléments ayant un degré d'appartenance supérieur ou égal à α i.

α –coupe(A) = $\{x \epsilon X | \mu_{A(x)} \geq \alpha\}$

FCM est basé sur la minimisation de la fonction objective en suivant un processus itératif

$$J_m(U,C) = \sum_i \sum_k (u_{i,k})^m . ||x_i - c_k||^2 ;$$

m>1 est un paramètre contrôlant le degré de floue (généralement m = 2).

xi est le ième élément des données mesurées.

C_k est le centre d'un classe k

Lorsque la partition devient stable, c'est-à-dire lorsqu'elle cesse de se développer entre deux itérations consécutives, l'algorithme FCM s'arrête. **[28]**

2. Principe

Fuzzy C-Means (FCM) est une technique de regroupement qui permet aux objets de données d'appartenir à deux ou plusieurs groupes. Cette méthode est dérivée de l'algorithme c-means**[46]** , qui est le même que l'algorithme k-means ci-dessus, développé par Dunn **[47]** en 1973 et amélioré par Bezdek **[48]** en 1981.

Il est largement utilisé dans la reconnaissance de formes, basé sur la minimisation de la fonction objectif suivante :

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \| x_i - c_j \|^2 \qquad 1 \le m < \infty$$

où m est un nombre réel (> 1), Ui j est le degré d'appartenance de xi dans le j ème Cluster, xi est le ième élément des données mesurées, c j est le centre d'un cluster et ||*|| est toute norme exprimant la similarité entre les données mesurées et le centre. Ce Partitionnement logique37 floue (fuzzy) est réalisé grâce à une optimisation itérative de la fonction objectif indiqué ci-dessus, avec la mise à jour de l'appartenance ui j et les centres des clusters c j. **[43]**

3. La différence entre fuzzy C-means et k-means

On peut résumer la différence entre fuzzy C-means et k-means dans la fonction d'appartenance d'un nuage de points dans deux clusters dans l'exemple suivant :

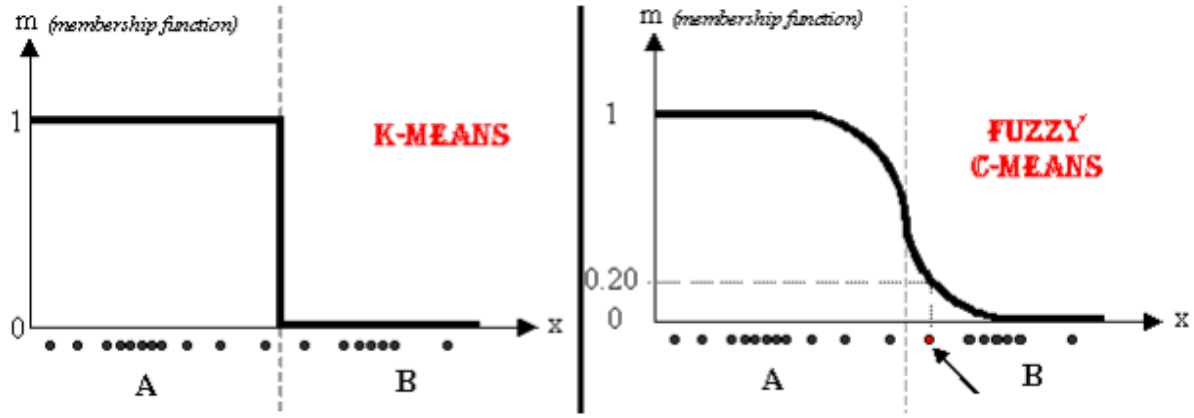


Figure 15 : Fonction d'appartenance dans kmeans/Fuzzy C-means **[43]**

Dans le cas de k-means un outil ne peut pas être là-dedans lequel règle clusters Simultanément, ce qui explique la Discrimination dichotomique parmi les clusters cependant en FCM il est éventuel qu'un outil appartient à règle ou riche clusters remplaçant inégaux pourcentages cad que le filon sont liés à quelque segment par la déviation d'une établi d'interdépendance, ce qui représente le tube floue de cet algorithme. Pour le faire, quelques-uns endettons certainement mettre une coulé appropriée nommée U leptocéphale les facteurs sont des nombres parmi 0 et 1,

et représentent la sellette d'interdépendance parmi les cœurs de filon et des clusters.

$$U_{MC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

Il est parce que démesuré de libeller que les initialisations étranges causent étranges mouvements de l'algorithme. En fait, il pourrait concourir enthousiasme le adéquat résultat, mais pratiquement pour une quantité agité d'itérations. **[43]**

4. Algorithme En Général de fuzzy C-means

1. Fixer les paramètres :
   a. Le nombre de classes ***K***
   b. Le seuil **s** représentant l'erreur de convergence
   c. Le degré flou **m**, généralement pris égal à 2
2. Initialiser les centres des ***K*** classes de manière aléatoire.
3. Mettre à jour la matrice ***U*** des degrés d'appartenance par la relation (2.10)
4. Mettre à jour le vecteur **K** des centres des classes par la formule (2.11)
5. Répéter les étapes 3 et 4 jusqu'à satisfaction du critère d'arrêt : **||I/(t) _ i/(t+1)|| < £,1** étant la t^{leme} itération.

[22]

5. Le FCM présente plusieurs inconvénients :

- Les niveaux d'adhésion sont relatifs, c'est-à-dire que l'adhésion d'une personne à une classe dépend de son adhésion à d'autres classes. Du centre de classe ils ne correspondent pas au centre réel ou typique
- les points atypiques (éloignés) peuvent avoir une valeur d'appartenance élevée et peuvent affecter de manière significative l'évaluation du centre de classe
- Ces algorithmes simulent des fluctuations dans l'étape de classification ou les classes l'ambiguïté entre est basée sur des règles de décision floues définies a priori. Au stade de la qualification, des résultats de classe ambigus ou non représentatifs affecteront la position du centre. **[22]**

H. Hierarchical clustering (Classification hiérarchique)

1. Définition

Une hiérarchie est un ensemble de parties imbriquées ; elle est généralement représentée par un arbre hiérarchique appelé dendrogramme, avec des objets en bas et des ensembles complets en haut. Il existe deux principaux types de méthodes de superposition : **descendante**, également appelée division, et **ascendante**, appelée agrégation. La première méthode, moins fréquemment utilisée, consiste à commencer par une classe grossière qui contient tous les objets, puis à la diviser en deux parties. Cette opération est répétée pour chaque itération jusqu'à ce que toutes les classes soient fusionnées. **[49]**

2. Le Principe

Le principe est simple : tout le monde forme d'abord une classe, c'est-à-dire n classes, donc on essaie de réduire itérativement le nombre de cette classe **newnbrclss < n**, de sorte qu'on fusionne deux classes à chaque étape (le choix de deux classes à fusionner est car ils ne sont pas similaires au sexe pour le rapprocher) ou ajouter un nouvel élément à la classe (si l'élément est plus proche de la classe que tout autre élément, alors l'élément appartient à cette classe). La valeur de dissemblance est appelée indice d'agrégation. Itérer et grandir d'itération en itération. Les algorithmes les plus populaires de ce type incluent : la classification ascendante hiérarchique (CHA), qui utilise le mot ascendant pour désigner des situations où tout le monde est présent. Ce sont des grappes matures, et nous essayons de les diviser en catégories de plus en plus grandes. Ainsi, le qualificatif « hiérarchique » signifie qu'il crée une structure hiérarchique. **[43]**

3. Algorithme CHA

Algorithme : Classification Ascendante Hiérarchique

Entrée : $\mathcal{Z}$

1. Initialiser les N classes c_k formées chacune d'une observation : $c_i = \{z_i\}$ et poser $d_C(c_i, c_{i'}) = d(z_i, z_{i'})$;
2. Fusionner les deux classes c_l et c_q les plus proches pour former une nouvelle classe $c_k = c_l \cup c_q$ tels que $d_C(c_l, c_q) = \min_{i,i'}(d_C(c_i, c_{i'}))$;
3. Calculer la distance entre la nouvelle classe c_k et les autres : $d_C(c_k, c_i)$ pour $i \neq l, q$;
4. Itérer : répéter n-1 fois les étapes 2 et 3 jusqu'à l'obtention d'une seule classe regroupant tous les objets.

Sortie : Un dendrogramme représentant les étapes de fusion des classes

[49]

Dans la figure suivante, on représente une illustration du principe de CHA et la hiérarchie finale obtenue où les liens hiérarchiques apparaissent clairement.

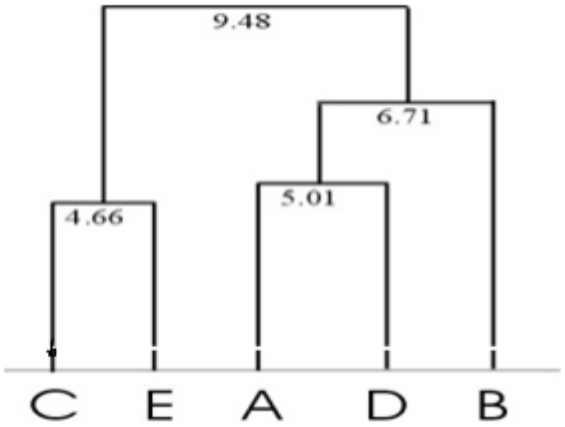


Figure 16: le dendrogramme de la hiérarchie H de la suite de partitions d'un ensemble {a, b, c, d, e} **[49].**

4. Les Avantage De Clustering

- La facilité pour traiter différentes formes de similarité ou de distance entre les objets.
- Leur application aux différents types d'attributs.

5. Les Limites De Clustering

Il existe de nombreux problèmes avec le Clustering, notamment :

- La méthode de regroupement actuelle ne répond pas pleinement (et en même temps) à tous les besoins. Par exemple, nous n'avons pas de variables continues (durée), mais des catégories nominales, comme les jours de la semaine. Dans ces cas, il est également nécessaire de comprendre le sujet afin de développer des regroupements appropriés.
- Traitant un grand nombre de mesures et une grande quantité de données, ce problème peut survenir en raison de la complexité du temps de calcul.
- L'efficacité de cette méthode dépend de la définition de « distance » utilisée.
- Si la métrique de distance n'existe pas, il faut la "définir", ce qui n'est pas toujours facile, surtout dans les espaces multidimensionnels.
- Les résultats de l'algorithme de clustering peuvent avoir différentes interprétations.
- De nombreux algorithmes de clustering nécessitent que vous spécifiiez le nombre de clusters à créer en entrée de l'ensemble de données avant d'exécuter l'algorithme. C'est-à-dire que si la valeur appropriée est connue à l'avance, la valeur appropriée doit être déterminée ; un problème est survenu et diverses méthodes ont été utilisées pour résoudre le problème. **[40]**

CONCLUSION

Après avoir parlé dans le premier chapitre du travail à distance et de son importance dans notre projet, nous sommes passés à l'explication des détails du Data Mining et de ses solutions et algorithmes pour résoudre les problèmes dans le domaine des nombreuses données qui nécessitent une classification. , nous avons abordé les classifications existantes et leurs algorithmes afin de choisir quels algorithmes Nous utilisons la classification des travailleurs selon les conditions fixées par l'employeur et nous seuls que la classification supervisée et l'algorithme du plus proche voisin sont la solution la plus appropriée pour nous aider classer les données des travailleurs.

Chapitre 03 : L’implémentation de l’algorithme KNN

CHAPITRE 3

L'implémentation de l'algorithme KNN

Sommaire

INTRODUCTION

*D*ans le courant de l'apprentissage automatique, différents types de classificateurs ont été mis au point, visant toujours à atteindre le plus haut niveau de précision et d'efficacité. Chaque classificateur a ses propres avantages et inconvénients. En CT, nous nous intéressons à l'algorithme k-plus proche voisin appelé kNN, qui a été introduit par Fix et Hodges en 1957 **[50]** et est devenu l'un des algorithmes de classification de texte les plus populaires. . Comme l'une des méthodes les plus recommandées parmi plus de dix méthodes de classification de texte **[51]**, **[52]**. Le chercheur Sébastian la recommande car elle est simple et comparable à la meilleure méthode SVM **[53]**, et outre ses bonnes performances, elle est également très facile à comprendre et à mettre en œuvre. **[54]**

1) Environnement et outils de mise en œuvre

1.1. Java

Java a été développée en C++. Quand il est né, son créateur l'a appelé "chêne", ce qui signifie chêne, ce qui signifie l'arbre qu'il a vu de la fenêtre du bureau lorsqu'il travaillait dans le laboratoire Microsystems de Sun, puis l'a renommé Java , Et ceci nom (nommer un langage de programmation n'est pas courant) n'est pas la première lettre d'une phrase ou d'une phrase, il a un certain sens, mais pour vous c'est juste un nom développé par le développeur de ce langage pour rivaliser avec d'autres noms. Il est livré avec de nombreux outils (JDK Java Développement Kit) et de nombreux packages : de nombreuses classes et ces différentes classes de base couvrent des domaines variés(E/S, GUI, réseau, etc.). Le grand nombre de "bibliothèques standard" explique en partie le succès de Java. La langue elle-même est dans le package java.Lang.

Le langage Java fournit un environnement interactif sur le World Wide Web, il est donc utilisé pour écrire des programmes éducatifs pour Internet, utiliser un logiciel de simulation informatique pour des expériences scientifiques et utiliser un logiciel de classe virtuelle pour l'apprentissage en ligne et l'apprentissage à distance. En plus d'être limité au réseau Elle nous, Java permet également de créer des programmes à usage personnel et commercial. Ces programmes sont implémentés via de nombreux programmes qui facilitent l'écriture de commandes, tels que les programmes NetBeans et Eclipse.

1.2. NetBeans

1.2.1. Définition

NetBeans IDE est un environnement de développement intégré (IDE) gratuit et open source qui peut être utilisé pour développer des applications de bureau, mobiles et Web. L'IDE prend en charge le l'applications dans plusieurs langages, notamment Java, HTML5, PHP et C++. IDE fournit une prise en charge intégrée pour l'ensemble du cycle de développement, de la création du projet au débogage, à l'analyse et au déploiement. L'IDE peut fonctionner sous Windows, Linux, Mac OS X et d'autres systèmes UNIX. L'IDE prend entièrement en charge JDK 7 et supérieur. Technologie d'extension Java. **[55]**

1.2.2. La version utilisée

La version utilisée pour développer notre application est IDE version 12.2

1.2.3. L'interface graphique de NetBeans

L'interface graphique de NetBeans est divisée en plusieurs fenêtres, menus et barres d'outils, et son apparence en mode conception est différente de son apparence dans les polices.

1.2.3.1. Apparence de l'interface en mode Design

Comme son nom l'indique, les modèles de conception sont conçus pour développer des interfaces utilisateur graphiques pour les applications.

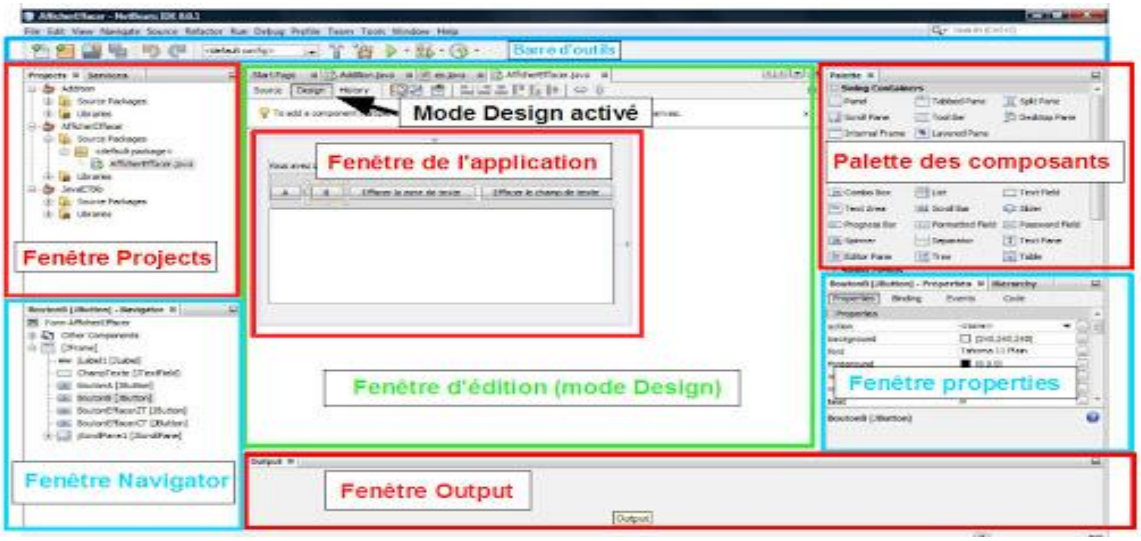


Figure 17 : L'interface graphique de NetBeans en mode Design

1.2.3.2. L'interface de NetBeans en mode Source

Le mode source est utilisé pour la programmation, dans ce cas, NetBeans affiche le texte du fichier source dans la fenêtre d'édition.

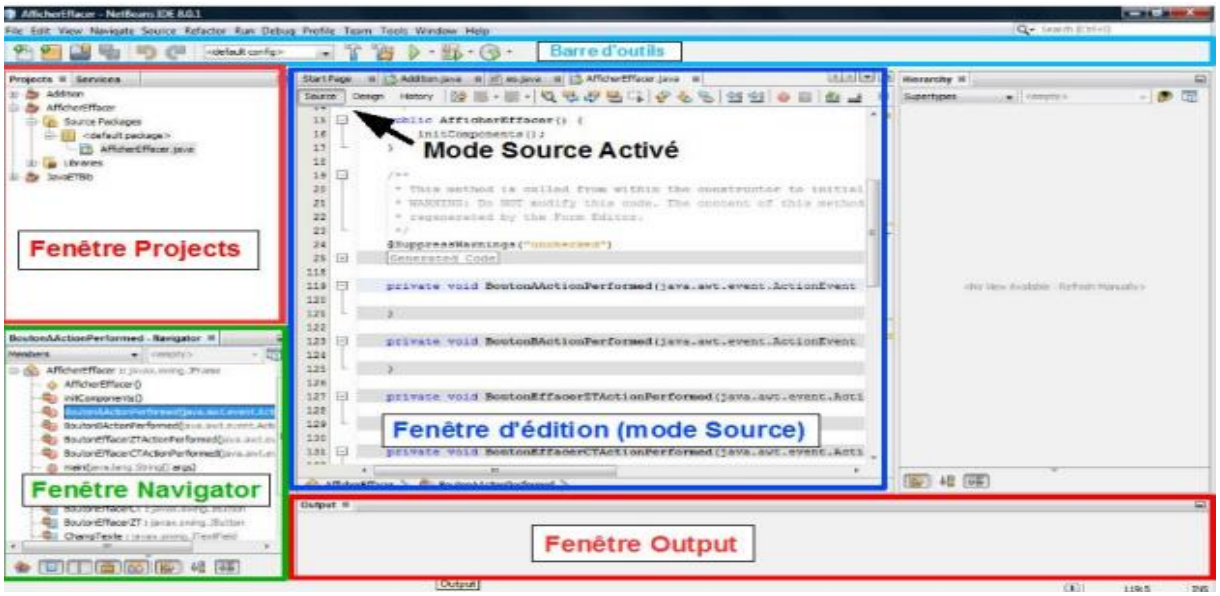


Figure 18 : L'interface graphique de NetBeans en mode Source

1.2.4. Les principales nouveautés de NetBeans 12.2

Dans cette nouvelle version, le support a été ajouté pour les nouvelles fonctions du langage Java introduites dans JDK 14 et JDK 15. De même, une surbrillance de code a été mise en place pour les mots clés « scellé », « non scellé » et « permis », en plus d'ajouter une mise en forme et une complétion de code pour le type « scellé », ainsi que la mise en forme pour les expressions « scellé » et "permis".

1.2.5. Les changements qui ressortent de cette nouvelle version

- Le code de support JavaFX a été étendu pour prendre en charge les objets immuables.
- Ajout du support pour les nouvelles fonctions PHP 8.
- Dépendances et infrastructure nettoyées pour les fonctions JavaScript et HTML
- Le compilateur javac est limité à une seule instance.
- Amélioration de la gestion des dépendances pour JavaScript et HTML.
- La prise en charge obsolète d'Oracle JET a été supprimée.
- Prise en charge améliorée de CSS3.

2) K- plus proche voisin (K-ppv)

2.1. Définition

L'algorithme des *K* plus proches voisins (noté K-PPV) en anglais (K Nearest Neighbors KNN)

L'algorithme KNN est un algorithme d'apprentissage supervisé qui classe les nouvelles instances en fonction des k voisins les plus proches. Le principe de base de l'algorithme KNN est de trouver les K instances les plus proches : calculer la similarité entre l'instance à classer et l'instance qui a été classée (l'instance de la bibliothèque d'apprentissage), trier les instances en fonction de leur distance, et prendre le premier k. Ensuite, nous trouvons la classe majoritaire parmi les k prochaines plus petites instances et l'attribuons aux instances non classées**. [58]**

Aussi « C'est une méthode très simple et directe. Il ne nécessite pas l'apprentissage, seules les données d'apprentissage doivent être enregistrées. Le principe de KNN est le suivant : Les données de catégorie inconnues sont comparées à toutes les données stockées. Pour les nouvelles données, nous sélectionnons la classe majoritaire parmi ses K voisins les plus proches en fonction de la distance sélectionnée (elle peut donc être lourde pour les grandes bases de données) ». **[57]**

Et d'autre façon …

« K-ppv est une méthode d'apprentissage basée sur des exemples qui n'inclut pas de phase d'apprentissage. Seul l'ensemble d'entraînement est enregistré. Lorsqu'un nouveau document arrive

pour classification, il est comparé au document de formation sur la base de la mesure de similarité. Considérons ensuite ses k plus proches voisins : nous observons leurs catégories et attribuons le document classifié à celui qui a la fréquence la plus élevée parmi les voisins. C'est la version de base de l'algorithme qui peut être améliorée ». **[54]**

Dans la reconnaissance de formes, l'algorithme du k-plus proche voisin (KNN) est une méthode non paramétrique de classification et de régression ; dans les deux cas, l'ensemble de données doit être alloué aux k espaces de caractéristiques identifiées lors de l'apprentissage La catégorie du voisin le plus proche. Le résultat dépend de l'utilisation de l'algorithme à des fins de classification ou de régression :

a. Classification KNN

Dans la classification KNN, le résultat est la classe d'appartenance. Les entités en entrée sont classées selon la plupart des statistiques de la classe d'appartenance de leurs k voisins les plus proches (k est généralement un petit entier positif). k = 1, l'objet se voit attribuer une classe appartenant à son plus proche voisin

b. Régression KNN

La régression KNN est une technique non paramétrique qui approxime intuitivement la relation entre les variables indépendantes et les résultats continus en faisant la moyenne des observations dans la même plage. La taille du voisinage doit être déterminée par l'analyste, ou elle peut être sélectionnée par validation croisée pour choisir la taille qui minimise l'erreur quadratique moyenne.

2.2. Principe de fonctionnement de KNN

La classification des nouvelles instances est effectuée en calculant la distance euclidienne entre la représentation vectorielle de l'instance à classer et la représentation des instances regroupées et classées, en sélectionnant l'instance la plus proche et en l'affectant à la classe majoritaire. **[58]**

2.3. Le principe de la méthode de l'algorithme K Plus Proche Voisin est le suivant

- On note *x* une nouvelle instance décrit par un vecteur de *p* attributs $x= \{x_1, x_2, \ldots..x_p\}$.
- Chaque instance I*i* de l'ensemble d'entrainement est sous forme d'un couple,

  $Ii =< a_{1i}, a_{2i}, \ldots., a_{pi} , c_i >$ tel que $(a_{1i} , a_{2i} , \ldots., a_{pi})$ représente un vecteur de *p* attributs de l'instance I*i*, c_i représente la classe de l'instance I_i. On trouve alors, parmi l'ensemble d'instances de la base d'apprentissage, les *k* plus proches voisins de *x* et on associe à *x* la classe majoritaire parmi ses *k* voisins les plus proche dans la base d'apprentissage. **[58]**

2.4. Pseudo code algorithme KNN [58]

Algorithme : KNN
Argument :
X : l'instance non classé
B : L'ensemble de toutes les instances (la base d'apprentissage)
Y_i : la classe d'instances de **B**
Y_{ij} : les instances de **B** appartenant la classe **Y_i**
D : les distances
K : le nombre du plus proche voisin
Entrées : X, B, D, k
Sortie : Y = classe de **X**
Début :

1. Pour chaque instance **Y_{ij}** dans **B**
   Calculer **D (X,Y_{ij})**
2. Classer les distances par ordre croissant
3. Compter le nombre d'occurrences de chaque classe **Y_i** parmi les **k** plus proche selon l'ordre
4. **Y**= la classe plus fréquent
5. Retourner **Y**

Fin.

Cette méthode dépend donc des deux éléments suivant :

1. Le nombre de voisins K
2. La distance entre deux instances **[58]**

Le résultat dépend du réglage de ces paramètres.

- Pour le premier paramètre, on prend k comme un nombre de voisins tel que la valeur k est généralement impair.
- Pour le deuxième paramètre, la méthode nécessite une distance pour mesurer la proximité entre l'instance x à classer et chaque instance de l'ensemble d'apprentissage. Lorsque les attributs sont numériques, la distance euclidienne est généralement utilisée. **[58]**

2.5. Les distances possibles en KNN

***a)* La distance Euclidienne**

Soit deux données d_1, d_2 de coordonnées respectives (x_1, y_1) et (x_2, y_2)

Distance (d_1, d_2) = $\sqrt{(x1-x2)^2+(y1-y2)^2}$

***b)* La distance de Manhattan**

Soit deux données d1 (x1, y1), et d2 (x2, y2)

Distance (d_1, d_2) = | x_1 - x_2 | + | y_1 - y_2 |

c) **La distance de Tchebychev**

Soit deux données $d_1(x_1, y_1)$ et $d_2(x_2, y_2)$

Distance (d1, d2)= max (|x1-x2|, |y1-y2|).

2.6. Les avantages et les inconvénients de la méthode des k plus proches voisins

2.6.1. Avantages

La méthode des k plus proches voisins représente des avantages tels que :

- L'algorithme KNN peut résister aux données bruitées.
- La méthode du k plus proche voisin est efficace lorsque la quantité de données est importante et incomplète.
- Cette méthode est l'une des plus simples de tous les algorithmes d'apprentissage automatique. **[56]**

2.6.2. Inconvénients

La méthode des k plus proches voisins comporte des inconvénients tels que :

3) Architecture général de la plate-forme

- Besoin de déterminer la valeur du numéro du voisin le plus proche (paramètre k).
- Le temps de prédiction est très long puisqu'on doit calculer la distance de tous les exemples.
- Cette méthode prend beaucoup de mémoire car elle nécessite beaucoup de mémoire pour traiter les paquets de données. **[56]**

Dans cette section, nous allons présenter notre plate-forme. Nous décrirons notamment son Classification dans le domaine de télé-recrutements**.**

3.1. Pseudo code (Algorithme de classification par KNN) de télé-recrutements

Cet algorithme KNN est proposé dans cette section. Nous avons implémenté dans notre projet. Nous donnons ci-dessous son pseudo code.

Algorithme KNN ;

Paramètre : **K** : nombre de voisin,
Age, Wilaya ,Nboutils ,Sexe, Expérience: **data_app**,

P : la longeur de data_app,
D : la distance

Début

1. **Pour** i de 1 a n faire
   Lire (data_app[i])
   Fin pour

2. **Pour** i=1 a P faire
3. Faire la normalisation sur data_app [i]
   Data normalized[i] = Data[i]-MinData[i] / MaxData[i]-MinData[i]
   Fin pour
4. Lire(K)
5. Lire individu[j]
   Pour i de 1 a n faire
   Calculer la distance D (data_app[i],individu[j])

   Fin pour

6. Fonction Trier les distances en ordre croissant selon le D
7. Compter le nombre d'occurrences de chaque classe parmi les K plus proche selon l'ordre (calcule de la classe majoritaire)
8. Classer individu [j]

Fin

9. Fonction prédire la classe de l'individu
10. Retourner la classe de l'individu est ce que Accepté ou non accepté

Fin

3.2. Data d'apprentissage

Dans cette application les le data d'apprentissage est un ensemble des employés accepté o non accepté. Tell que Le tableau suivant présente la classification des employés par catégorie (accepté o non accepté).

Nom	Prénom	Age	Wilaya	Nb outils	Sexe	Expérience	Classe
BOUDJLAL	MAISSA	22	KHENCHELA	3	F	2	Accepté
ATTIA	MOUSSA	30	ADRAR	5	M	4	Non accepté
BOUDRAA	FATIMA	20	TLEMCEN	1	F	3	Accepté
HEFIAN	AMANI	33	CONSTANTINE	0	F	2	Non accepté
SAIHI	SARA	32	KHENCHLA	5	F	0	Non accepté
BORBACHE	ZOHAIR	25	ANNABA	2	M	1	Non accepté
SEKAONI	ZAKARIA	27	BATNA	0	M	5	Accepté
AGABA	HAMZA	20	BATNA	1	M	2	Accepté
DJBAILI	KHALED	38	CONSTANTINE	6	M	0	Non accepté
CHERAB	ROFIADA	44	TLEMCEN	4	F	1	Non accepté

Tableau 4 : *classification des employés par catégorie (accepté o non accepté)*

3.3. Les étapes d'application

3.3.1. Calculer la normalisation

- Notre data d'apprentissage utilisée en java est :

```
List<MyObject> myObjects = new ArrayList<>();

myObjects.add(new MyObject(22, 40 ,3 ,0 , 2, "accpté"));
myObjects.add(new MyObject(30, 1  ,5 ,1 , 4, "NoNaccpté"));
myObjects.add(new MyObject(20, 13 ,1 ,0 , 3, "accpté"));
myObjects.add(new MyObject(33, 25 ,0 ,0 , 2, "NoNaccpté"));
myObjects.add(new MyObject(32, 40 ,5 ,0 , 0, "NoNaccpté"));
myObjects.add(new MyObject(25, 23 ,2 ,1 , 1, "NoNaccpté"));
myObjects.add(new MyObject(27, 5  ,0 ,1 , 5, "accpté"));
myObjects.add(new MyObject(20, 5  ,1 ,1 , 2, "accpté"));
myObjects.add(new MyObject(38, 25 ,6 ,1 , 0, "NoNaccpté"));
myObjects.add(new MyObject(44, 13 ,4 ,0 , 1, "NoNaccpté"));
Knn knn = new Knn(myObjects);
for (MyObject myObject : knn.myObjects) {
    System.out.println(myObject);
}
```

- Pour normaliser notre data on a applique :

La normalisation de **x** est :

$$X_{Normalised} = x - \min(x) / \max(x) - \min(x)$$

- Dans notre programme java on a calculé de normalisation avec le code suivant :

```
//normaliser un objet

    MyObject normeObject(MyObject myObject1) {
        myObject1.age = (myObject1.age - smallsetAge) / (biggestAge - smallsetAge);
        myObject1.nbr_outils = (myObject1.nbr_outils - smallestnbr) / (biggestnbr - smallestnbr);
        myObject1.experience = (myObject1.experience - smallestexpr) / (biggestexpr - smallestexpr);
        return myObject1;
    }
//normaliser tout les attributs de data

    void normalizeData() {
        for (int i = 0; i < myObjects.size(); i++) {
            myObjects.set(i, normeObject(myObjects.get(i)));
        }
    }
```

- Le résultat après la normalisation

```
*******************************************************************************
le tableau après la normalization
*******************************************************************************
MyObject{age=0.5, wilaya=40, nbr_outils=0.8333333333333334, sexe=femme, experience=0.0, myClass=NoNaccpté} 2.0420747891408007
MyObject{age=0.08333333333333333, wilaya=40, nbr_outils=0.5, sexe=femme, experience=0.4, myClass=accpté} 1.7643262799783441
MyObject{age=0.2916666666666667, wilaya=5, nbr_outils=0.0, sexe=hommme, experience=1.0, myClass=accpté} 1.384537628396009
MyObject{age=1.0, wilaya=13, nbr_outils=0.6666666666666666, sexe=femme, experience=0.2, myClass=NoNaccpté} 2.094321024951895
MyObject{age=0.5416666666666666, wilaya=25, nbr_outils=0.0, sexe=femme, experience=0.4, myClass=NoNaccpté} 1.7159383568311666
MyObject{age=0.4166666666666667, wilaya=1, nbr_outils=0.8333333333333334, sexe=hommme, experience=0.8, myClass=NoNaccpté} 1.22499999999999
MyObject{age=0.0, wilaya=13, nbr_outils=0.16666666666666666, sexe=femme, experience=0.6, myClass=accpté} 1.7567528441859874
MyObject{age=0.75, wilaya=25, nbr_outils=1.0, sexe=hommme, experience=0.0, myClass=NoNaccpté} 1.577819346087786
MyObject{age=0.20833333333333334, wilaya=23, nbr_outils=0.3333333333333333, sexe=hommme, experience=0.2, myClass=NoNaccpté} 1.648568402517
MyObject{age=0.0, wilaya=5, nbr_outils=0.16666666666666666, sexe=hommme, experience=0.4, myClass=accpté} 1.5640270315936216
*******************************************************************************
```

3.3.2. Calcule de la distance

- Le calcule de distance sachant qu'on utiliser la distance Euclidienne :

$$\text{Distance} = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

```
//Calculer la distance eucldienne
public double distance(MyObject o, int coef) {
    double distance = 0;
    if (coef == 1) {
        distance += Math.abs(age - o.age);
        if (o.wilaya == wilaya) {
            distance += 0;
        } else {
            distance += 1;
        }
        distance += Math.abs(nbr_outils - o.nbr_outils);
        distance += Math.abs(sexe - o.sexe);
        distance += Math.abs(experience - o.experience);
    } else {
        distance += Math.pow(age - o.age, coef);
        if (o.wilaya == wilaya) {
            distance += 0;
        } else {
            distance += 1;
        }
  // pour la distance de wilaya: si le méme wilaya de new object alore wilaya=0 sinon wilaya=1
        distance += Math.pow(nbr_outils - o.nbr_outils, coef);
        distance += Math.pow(sexe - o.sexe, coef);
        distance += Math.pow(experience - o.experience, coef);

        distance = Math.pow(distance, 1.0 / coef);
    }
    return distance;
}
```

- L'appel de la distance dans la classe « KNN.java »

```
//calculer la distance entre les données utilisé La distance Euclidienne

    void setDistances(MyObject myObject, int coef) {

        données = new HashMap<>();
        for (int i = 0; i < myObjects.size(); i++) {
            MyObject myObject1 = myObjects.get(i);
            double distance = myObject.distance(myObject1, coef);
           données.put(myObject1, distance);
        }

    }
```

3.3.3. Trier les distances dans l'ordre croissant :

- Le code suivant faire le tri :

```
// trier les données

    void sort() {
        Map<MyObject, Double> sorted = données.entrySet().
                stream().
                sorted(Map.Entry.comparingByValue()).
                collect(Collectors.toMap(Map.Entry::getKey, Map.Entry::getValue, (e1, e2)
                        -> e1, LinkedHashMap::new));

        données = sorted;
    }
```

- Le résultat de trie de data selon la distance dans l'ordre croissant :

A. En cas (âge=40, wilaya=40, nombre d'outil informatique= 0, sexe= femme (0), expérience=0)

```
*******************************************************************************
 la distance de chaque employer dans l'ordre croissant
*******************************************************************************
MyObject{age=0.5, wilaya=40, nbr_outils=0.8333333333333334, sexe=femme, experience=0.0, myClass=NoNaccpté} 0.8975274678557508
MyObject{age=0.08333333333333333, wilaya=40, nbr_outils=0.5, sexe=femme, experience=0.4, myClass=accpté} 0.986154146165801
MyObject{age=0.5416666666666666, wilaya=25, nbr_outils=0.0, sexe=femme, experience=0.4, myClass=NoNaccpté} 1.1158267985867898
MyObject{age=1.0, wilaya=13, nbr_outils=0.6666666666666666, sexe=femme, experience=0.2, myClass=NoNaccpté} 1.2297244497131143
MyObject{age=0.0, wilaya=13, nbr_outils=0.16666666666666666, sexe=femme, experience=0.6, myClass=accpté} 1.4429907214608908
MyObject{age=0.20833333333333334, wilaya=23, nbr_outils=0.3333333333333333, sexe=hommme, experience=0.2, myClass=NoNaccpté} 1.594282318509
MyObject{age=0.0, wilaya=5, nbr_outils=0.16666666666666666, sexe=hommme, experience=0.4, myClass=accpté} 1.6977108770995792
MyObject{age=0.75, wilaya=25, nbr_outils=1.0, sexe=hommme, experience=0.0, myClass=NoNaccpté} 1.7340543372237343
MyObject{age=0.2916666666666667, wilaya=5, nbr_outils=0.0, sexe=hommme, experience=1.0, myClass=accpté} 1.8147734783652139
MyObject{age=0.4166666666666667, wilaya=1, nbr_outils=0.8333333333333334, sexe=hommme, experience=0.8, myClass=NoNaccpté} 1.87298039379902
*******************************************************************************
```

B. En cas (âge=21, wilaya=25, nombres d'outils informatique= 7, sexe=homme(1), expérience=5)

```
*******************************************************************************
 la distance de chaque employer dans l'ordre croissant
*******************************************************************************
MyObject{age=0.4166666666666667, wilaya=1, nbr_outils=0.8333333333333334, sexe=hommme, experience=0.8, myClass=NoNaccpté} 1.22499999999999
MyObject{age=0.2916666666666667, wilaya=5, nbr_outils=0.0, sexe=hommme, experience=1.0, myClass=accpté} 1.384537628396009
MyObject{age=0.0, wilaya=5, nbr_outils=0.16666666666666666, sexe=hommme, experience=0.4, myClass=accpté} 1.5640270315936216
MyObject{age=0.75, wilaya=25, nbr_outils=1.0, sexe=hommme, experience=0.0, myClass=NoNaccpté} 1.577819346087786
MyObject{age=0.20833333333333334, wilaya=23, nbr_outils=0.3333333333333333, sexe=hommme, experience=0.2, myClass=NoNaccpté} 1.648568402517
MyObject{age=0.5416666666666666, wilaya=25, nbr_outils=0.0, sexe=femme, experience=0.4, myClass=NoNaccpté} 1.7159383568311666
MyObject{age=0.0, wilaya=13, nbr_outils=0.16666666666666666, sexe=femme, experience=0.6, myClass=accpté} 1.7567528441859874
MyObject{age=0.08333333333333333, wilaya=40, nbr_outils=0.5, sexe=femme, experience=0.4, myClass=accpté} 1.7643262799783441
MyObject{age=0.5, wilaya=40, nbr_outils=0.8333333333333334, sexe=femme, experience=0.0, myClass=NoNaccpté} 2.0420747891408007
MyObject{age=1.0, wilaya=13, nbr_outils=0.6666666666666666, sexe=femme, experience=0.2, myClass=NoNaccpté} 2.094321024951895
*******************************************************************************
```

3.3.4. Les K plus proche voisins

- Dans ce cas k=3

```
String myClass = knn.getNclosestObject(3);
System.out.println("K plus proche voisins.....; are ");
```

- Remplir la liste qui contient les K plus proche objets :

```
    //remplir la list qui comport les K plus proche object
    NnearsestObjects.add(myObject);
    count++;
}
```

- L'affichage de 3 plus proches (les trois premiers éléments après le trier de data selon les distances)

A. En cas (âge=40, wilaya=40, nombre d'outil informatique= 0, sexe= femme (0), expérience=0)

```
K plus proche voisins.....; sont
        MyObject{age=0.5, wilaya=40, nbr_outils=0.8333333333333334, sexe=, experience=0.0, myClass=NoNaccpté}
        MyObject{age=0.08333333333333333, wilaya=40, nbr_outils=0.5, sexe=, experience=0.4, myClass=accpté}
        MyObject{age=0.5416666666666666, wilaya=25, nbr_outils=0.0, sexe=, experience=0.4, myClass=NoNaccpté}
*********************************************************************************
```

B. En cas (âge=21, wilaya=25, nombres d'outils informatique= 7, sexe=homme(1), expérience=5)

```
*********************************************************************************
K plus proche voisins.....; sont
*********************************************************************************
        MyObject{age=0.4166666666666667, wilaya=1, nbr_outils=0.8333333333333334, sexe=hommme, experience=0.8, myClass=NoNaccpté}
        MyObject{age=0.2916666666666667, wilaya=5, nbr_outils=0.0, sexe=hommme, experience=1.0, myClass=accpté}
        MyObject{age=0.0, wilaya=5, nbr_outils=0.16666666666666666, sexe=hommme, experience=0.4, myClass=accpté}
*********************************************************************************
```

3.3.5. La classe majoritaire

-Pour prédire la classe majoritaire après le trie et le choix de K a partir de pseudo code suivants :

```
//recupere la class majournant depuis données ordonnee

    public String getNclosestObject(int max) {
        NnearsestObjects = new ArrayList<>();
        sort();
        int count = 0;
        List<String> close = new ArrayList<>();
        for (MyObject myObject : this.données.keySet()) {
            if (count >= max) {
                break;
            }
```

- Récupère la classe majoritaire :

```
    //écrire la classe majoritaire
    String mostOcc = close.stream()
            .reduce(BinaryOperator.maxBy((o1, o2) -> Collections.frequency(close, o1)
                    - Collections.frequency(close, o2))).orElse(null);
    return mostOcc;
    }
```

- La classe s'affiche de forme suivant :

```
    @Override
    public String toString() {
        String sexeOfPerson = (sexe == 1) ? "hommme" :  "femme";
        String str = "MyObject{" + "age=" + age + ", wilaya=" + wilaya +
                ", nbr_outils=" + nbr_outils + ", sexe=" + sexeOfPerson + ", experience=" + experience;
        if (myClass != null) {
            str += ", myClass=" + myClass;
        }
        str += '}';

        return str;
    }
```

A.

- L'affichage d'une classe majoritaire d'après les données suivants (âge=40, wilaya=40, nombre d'outils informatique =0, sexe= femme (0), expérience=0)
- On a saisir les données comme ce suit :

```
**********************************************************
new Object
**********************************************************
Age

40
Wilaya
40
Nombre des outils informatique
0
Sexe (1 for males otherwise femmales )
0
Experience
1
**********************************************************
```

La classe de l'employé suivant est :

```
*******************************************************************************
class of new object
*******************************************************************************

class of new object MyObject{age=40.0, wilaya=40, nbr_outils=1.0, sexe=femme, experience=0.0} is NoNaccpté
```

B.

- L'affichage d'une classe majoritaire d'après les données suivants (âge=21, wilaya=25, nombre d'outils informatique = 7, sexe= homme (1), expérience=5)
- On a saisir les données comme ce suit :

```
*******************************************************************************
new Object
*******************************************************************************
Age
21
Wilaya
25
Nombre des outils informatique
7
Sexe (1 for males et 0 for femmales )
1
Experience
5
*******************************************************************************
```

- La classe de l'employé suivant est :

```
*******************************************************************************
class of new object
*******************************************************************************

class of new object MyObject{age=21.0, wilaya=25, nbr_outils=5.0, sexe=hommme, experience=7.0} is accpté
```

CONCLUSION

Dans ce chapitre, nous avons présenté le volet technique de notre application. Nous avons défini les environnements matériels et logiciel utilisés pour réaliser notre plate-forme ainsi que l'application développée en termes de conception et d'implémentation l'algorithme KNN.
Nous avons illustré par quelques pages le code de la plate-forme de télé recrutement.

CONCLUSION GÉNÉRALE

CONCLUSION GENERALE

L'algorithme Nearest Neighbor est un algorithme d'apprentissage supervisé qui classe les nouveaux cas en fonction des voisins les plus proches

Dans notre application nous déterminons le nombre de voisins K et calculons les distances, à partir desquelles nous pouvons savoir si le nouveau travailleur est qualifié pour travailler ou non Comme nous l'avons vu dans l'application, l'algorithme de KNN est facile, simple et plus efficace pour les données méga Mais le résultat n'est pas nécessairement correct 100%, car l'employée peut être accepté au lieu d'être rejeté, ou vice versa. Comme tous les algorithmes de classification de KNN a des inconvénients, dont l'un est que plus le nombre de voisins est grand, plus le résultat est différent, et aussi ce dernier consomme une très grande quantité de mémoire, et c'est pourquoi d'autres algorithmes sont apparus qui dépendent de la classification pour éviter les erreurs de KNN comme K-means et l'arbre de décision… etc.

En guise de perspectives pour notre travail, nous proposons les orientations suivantes :

Intégrer cette application a un site web de tel recrutement.

Intégrer à la plate-forme de nouvelles méthodes d'apprentissage et mettre en œuvre les outils nécessaires à leur comparaison.

Bibliographie

[1] ALVES CACHAPELA, LAURIE, ORIANNE, JEAN-FRANÇOIS, Le télétravail. Confiance et autocontrôle : alternative ou partenaires du contrôle à distance, diplôme de Master en Sciences du Travail, université de liège faculté (Faculté des Sciences Sociales), 2015-2016.

[2] Monique Haicault,Alain,,Travail a distance et/ou le télétravail ,les formes d'emploi,nouveaux contenus de travail des logiques contradictoires**,** juillet 1998

[3] © Organisation internationale du Travail 2020,Le télétravail durant la pandémie de Covid-19 et après,Première édition, 2020, Imprimé en Suisse

[4] BUREAU INTERNATIONAL DU TRAVAIL, Difficultés et avantages du télétravail pour les travailleurs et les employeurs dans les secteurs des TIC et des services financiers, Genève, 2016

[5] Michel Welrave, Comment introduire le télétravail ?,University of Antwerp Antwerpen, Belgium

,janvier 2010

[6]Génération Industrie, Télétravail : nouveau mode d'organisation Processus et outils RH pour faciliter sa mise en œuvre,22/06/2020

[7] FREDERIC PENNERATH, Méthodes d'extraction de connaissances à partir de données modélisables par des graphes. Application à des problèmes de synthèse organique. Thèse Doctorat de l'université Henri Poincaré – Nancy 1 ,2009

[8] D. HAND, H. MANNILA et P. SMYTH, Principles of Data Mining,MIT Press, Cambridge, MA, 2001.

[9] P. CABENA, P. HADJINIAN, R. STADLER, J. VERHEES et A. ZANASI, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.

[10] E-G. TALBI, Fouille de données (Data Mining) : Un tour d'horizon, Laboratoire d'Informatique Fondamentale de Lille.

[11] O. R. ZAÏANE, Principles of Knowledge Discovery in Databases, CMPUT690, University of Alberta, 1999.

[12] Ph. PREUX, Fouille de données : Notes de cours, Université de Lille 3, 9 octobre 2008.

[13] G. DONG, J. PEI, Sequence Data Mining, Springer Edition, 2007.

[14] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007.

[15] Mr Khaled BENALI , Contributions au domaine de Data Mining par l'utilisation des Ontologies, diplôme de Doctorat , université des sciences et de la technologie d'oran, 2018

[16] ZACCONE Giancarlo, MD REZAUL Karim, MENSHAWY Ahmed. Deep learning with tensorflow, 2017

[17] DJOKHRAB Ala eddine. Planification et optimisation de trajectoire d'un robot manipulateur à 6 ddl par des techniques neuro_floues. 2015

[18] PARIZEAU Marc. Réseaux de neurones. 2004

[19] M. J. BERRY, G. S. LINOFF, Data Mining Techniques For Marketing, Sales, and Customer Relationship, Management, Second Edition, 2004.

[20] M. J. BERRY, G. S. LINOFF, Mastering Data Mining: The Art and Science of Customer Relationship Management, 2000.

[21] D.T. LAROSE, Discovering Knowledge In Data: An Introduction to Data Mining, Central Connecticut State University, 2005.

[22] Mr. BEKKARI FOUAD , La logique floue pour Classification Des Feuilles de vigne, Memoire Master Academique , UNIVERSITE KASDI MARBAH OUARGLA, 2016

[23] Kalmegh, Gupta, A,Learning Apache Mahout Classification, Birmingham, UK. Packt Publishing. 2015.

[24] Bostjan-Kaluza-Machine-Learning-in-Java-Packt-Publishing-2016.

[25] G. CALAS, Études des principaux algorithmes de data mining, Spécialisation Sciences Cognitives et Informatique Avancée, France.

[26] SENOUSSI Hafida, Mr Z. AHMED FOUATIH, Sélection de Données pour l'Apprentissage des Réseaux de Neurones, Arbres de Décision et les k-Plus Proches Voisins : Application en Diagnostic de Pannes, Université des Sciences et de la Technologie d'Oran Mohamed BOUDIAF, 27 avril 2015.

[27] Quinlan, J. R. C4.5: Programs for Machine Learning. M. Kaufmann, San Francisco, 1983.

[28] A. AMINE, B. ATMANI, A. BENYETTOU, H. BELBACHIR, F. KHELFI, B. BELDJILALI, Contribution A La Catégorisation De Textes Et A L'extraction D'information, University D'oran , 2013.

[29] Dr BOUKLI HACENE Sofiane , Pr BELALEM Ghalem, Dr BERRABAH Djamel, Pr. ELBERRICHI Zakaria, privacy Preserving Classification of Biomedical Data, THESE DE DOCTORAT, UNIVERSITE DJILLALI LIABES, 17/07/2019

[30] Mohamed BOUAZIZ, M. Georges LINARÈS , Réseaux de neurones récurrents pour la classification de séquences dans des flux audiovisuels parallèles, l'Université d'Avignon et des Pays de Vaucluse, le 6 décembre 2017.

[31] Mathieu FEUILLOY, Daniel Schang, Étude d'algorithmes d'apprentissage artificiel pour la

prédiction de la syncope chez l'homme, Thèse de doctorat, École Doctorale STIM, 8 juillet 2009.

[32] ZURADA, J. M. Artificial neural systems. West publishing company, 1992.

[33] HERAULT, J. Réseaux neuronaux et traitement de signal. France: Hermes. 1994.

[34] Corentin HARDY , Marc TOMMASI, Sébastien MONNET, Contribution au développement de l'apprentissage profond dans les systèmes distribués, THÈSE DE DOCTORAT, L'UNIVERSITE DE RENNES 1, 8 avril 2019.

[35] KARIMA HOUACINE. Commande neuro-floue d'une machine asynchrone dans une chaine de population d'un véhicule électrique. Thèse de Doctorat à l'université Mouloud Mammeri de Tizi-Ouzou, juin, 2016.

[36] REZGUI ZOHRA , Mme. Héla OUAILI MALLEK ,Détection et classification de visages pour la description de l'égalité femme-homme dans les archives télévisuelles, niversité de Carthage , 19 Novembre 2019 .

[37] Mohd Hanafi Ahmad Hijazi , Image Classification: A Study in Age-related Macular Degeneration Screening, Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy, July 2012

[38] Mathieu FEUILLOY, Daniel Schang, Étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme, Thèse de doctorat, Doctorale STIM, 8 juillet 2009.

[39] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007.

[40] B. LAVOIE, Arbres de décisions, Synthèse de lectures, Séminaire sur l'apprentissage automatique, Programme de Doctorat en Informatique Cognitive, Université du Québec à Montréal, 15 mars 2006.

[41] DMITRII GMYZIN , A comparison of supervised machine learning classification techniques and theory-driven approaches for the prediction of subjective mental workload,Institute of Technology for the degree of M.Sc. in Computing (Stream), January 2016.

[42] Mlle. Oumaima ALAOUI ISMAILI , Clustering prédictif Décrire et Prédire simultanément, Thèse De Doctorat, L'université Paris–Saclay, 10/11/16

[43] KOUDRI MOHAMMED , Monsieur MOURTADA BENAZOUZ, maitre assistant De l'Informatique à l'université, univ-tlemcen.Memoire.

[44] Mory OUATTARA , M. BADRAN Fouad , Développement et mise en place d'une méthode de classification multi-bloc Application aux données de l'OQAI,Docteur du Conservatoire National des Arts et Métiers, 18 mars 2014

[45] Jain A. K : Data clustering : 50 years beyond k-means. Pattern Recognition Letters , 2010

[46] Ball, G.H. et Hall, D.J.ISODATA, an Iterative Method of Multivariate Analysis and Pattern Recognition. Behavior Science, 153, 1967.

[47] J. C. Dunn : "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics ,no 3, pp 32-57. 1973.

[48] J. C. Bezdek : "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York. 1981.

[49] Mory OUATTARA, Développement et mise en place d'une méthode de classification multi-bloc Application aux données de l'OQAI , THÈSE DE DOCTORAT, CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS,18 mars 2014.

[50] Fix, E., & Hodges, J. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties . International Statistical Review , 1989.

[51] Yang, Y. An evaluation of statistical approaches to text categorization. information retrieval , 1999.

[52] Yang, Y., & Liu, X. A re-examination of text categorization methods. Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR), 1999.

[53] SEBASTIANI, F. Machine learning in automated text categorization . ACM computing surveys , 2002.

[54] BARIGOU Fatiha, CONTRIBUTION À LA CATÉGORISATION DE TEXTES ET À L'EXTRACTION D'INFORMATION, université d'Oran, diplôme de doctorat, 2012/2013

[58] HALICH AMEL, Classification supervisée à base de KNN avec pondération d'attributs par L'Algorithme Génétique, diplôme de Magister, Université des Sciences et de la Technologie Houari Boumediene, 02/02/2015

[57] FAÏCEL CHAMROUKHI, Classification supervisée : Les K-plus proches voisins, Université de Caen

[56] LABIAD ALI, sélection des mots clés basée sur la classification et l'extraction des règles d'association, juin 2017

[55] Iqbal, M., Abid, M. M., Waheed, U., Alam Kazmi, S. H. (2017). Classification of Malicious Web Pages through a J48 Decision Tree, aNaïve Bayes, a RBF Network and a Random Forest Classifier forWebSpam Detection. International Journal of u-and e-Service, Science and Technology, 10(4), 51-72.