Revue Africaine d'Environnement et d'Agriculture
https://www.rafea-congo.com
2025 ; 8(1), 133-145
ISSN (Print) : 2708-7743 **eISSN** : 2708-5422 **Dépôt légal** : JL 3.01807-57259

Apprentissage artificiel basé sur la combinaison des classifieurs automatiques par vote majoritaire pour la prédiction de clients crédibles dans le secteur bancaire

Armel Mbenza Makiadi Oumar*[1,4], Rebecca Raïssa Baketi[2], Christian Kilikwa[2], Fidèle Lukelwa Lubela[2], Gloria Thsibola Kabeya[4] Denis Kutelama[4], Maria Amisi Mstayabu[4], Pierre Kafunda Katalay[3]

[(1)]Institut Supérieur Pédagogique de la Gombe. BP 3580 Kinshasa/Gombe (RDC). E-mail: mcoumarh@ispgombe.ac.cd
[(2)] Institut Supérieur d'Etudes Agronomiques de Mangai. BP 855 Kinshasa XI (RDC).
[(3)] Université de Kinshasa. Faculté des Sciences. BP 190 Kinshasa XI (RDC).
[(4)] Université de Kinshasa. Faculté des Sciences Economiques et de Gestion. BP 190 Kinshasa XI (RDC).

Reçu le 15 septembre 2024, **accepté** le 20 mars 2025, **publié en ligne** le 29 mars 2025
DOI : https://dx.doi.org/10.4314/rafea.v8i1.14

RESUME

Description du sujet. Le secteur bancaire accorde une grande importance à la préservation des clients en raison de la prédominance des modes de gestion axées sur la "relation client" en assistant le banquier dans sa quête de clients fiables.
Objectifs. L'étude vise à anticiper le risque de remboursement de crédit des futurs emprunteurs de la banque afin de ne donner le crédit qu'aux emprunteurs forts potentiels, qui sont considérés comme des clients fiables.
Méthodes. La prédiction a été effectuée en utilisant des classifieurs automatiques ainsi que l'approche de vote majoritaire qui impliquent la fusion de plusieurs classifieurs de manière autonome afin d'obtenir le classement automatique d'un nouvel individu en utilisant une fonction mathématique qui relie le jeu de données de départ à un ensemble de classes d'arrivée.
Résultats. Les résultats obtenus ont montré que 70 % des cas traités sont des dossiers de prêt satisfaisants, en raison de variables explicatives qui présentent une répartition de données sur la variable d'intérêt (Risque). Les dossiers de prêt non satisfaisants constituent le pourcentage restant, soit 30 %.
Conclusion. Les performances des systèmes de classification automatique sont considérablement améliorées par la combinaison de classifieurs. Ce dispositif intelligent apporte un éclaircissement aux banques concernant les décisions de donner ou non des prêts aux emprunteurs.

Mots-clés : Classifieur, vote majoritaire, combinaison, prêt, banque.

ABSTRACT

Machine learning based on the automatic classifiers combination by the voting mainstream for the prediction of trustworthy customers in the banking sector

Description of the subject. The banking sector places a high emphasis on customer retention because of the popularity of management styles that are centered on "relation clients," which aid bankers in retaining their reliable clientele.
Objectives. The study attempts to forecast the risk of credit payback from prospective borrowers who are considered reliable clients in order to refrain from granting credit to them.
Methods. Both automatic classifiers and the majoritarian voting approach have been employed for the prediction. This approach uses a mathematical function that applies the initial data game to a set of arriving classes to automatically classify a new individual by combining numerous classifiers in an autonomous manner.
Results. According to the results, 70% of the handled cases are categorized loan files because of explanatory variables that display a distribution of data on the interest variable (risk). The remaining 30% of instances are unsatisfactory loan files.
Conclusion. When classifiers are combined, automatic classification systems perform noticeably better. This clever system gives banks clarity when deciding whether to provide or refuse loans to borrowers.

Keywords: Classifier, majority vote, combination, loan, bank.

1. INTRODUCTION

Le terme "crédit" est issu du latin "creditum", qui signifie "croire" ou "avoir confiance" (Charles, 2012). Accorder un crédit, c'est accorder confiance, mais c'est également donner librement la disponibilité affective et immédiate d'un bien réel ou d'un pouvoir d'achat, en échange de la promesse que le même bien ou l'équivalent vous sera restitué dans un délai donné, généralement avec la rémunération du service fourni et du risque couru (Charles, 2012). Il existe un risque de perte partielle ou totale lié à la nature même de ce service.

Différentes approches ont été élaborées afin de répondre aux exigences des systèmes déjà en place dans différentes applications, comme la prédiction de la crédibilité d'un client. Étant donné que les banques opèrent dans des environnements extrêmement concurrentiels, il est difficile de conserver un client fidèle (Lyonnet, 2006).

La rentabilité d'un établissement de crédit désigne sa capacité à générer des bénéfices adéquats de son activité, après avoir déduit les coûts liés à cette activité, afin de maintenir son activité de manière durable (Charles, 1967). Le crédit est inévitablement lié à une notion de rentabilité et de risque. Dans le cadre de l'activité bancaire, ces deux aspects restent étroitement liés. Il n'est pas toujours avisé de chercher toujours plus de profits sur les prêts bancaires, car cela a des répercussions (Nouy, 1993). Selon la politique de chaque institution financière, il est possible de choisir entre une préférence pour la qualité ou le volume de crédit (Charles, 2012). Les répercussions de cette décision stratégique sont importantes car elle établit les principes directeurs de la banque et sa politique de prêt (Sylvie-Nuria Noguer, 2018). Il est essentiel de gérer de manière optimale le risque et la rentabilité afin que la banque puisse générer un maximum de profits tout en minimisant les pertes (Grawitz, 2001).

De nos jours, il est envisageable pour un programme d'intelligence artificielle d'aider les spécialistes dans la prise de décision dans un environnement complexe et évolutif, comme l'analyse du marché financier ou, le cas échéant, la prédiction de la crédibilité des clients. En s'engageant sur cette voie, quelques interrogations ont orienté cette recherche : (i) Faut-il anticiper le comportement d'un client inconnu ? (ii) Est-il justifié de classer un client en fonction de son statut ? (iii) Pourquoi diminuer le risque de remboursement de crédit ?

Un dispositif automatisé pourrait aider à repérer les erreurs d'analyse afin d'améliorer la fiabilité des résultats. Ainsi, cette étude pourrait constituer une solution appropriée lorsqu'il s'agit de décider si un emprunteur doit obtenir ou non un crédit (Barger, 2003).

La gestion du risque de crédit est en effet en constante évolution, étant donné la complexité des risques liés à l'activité de prêt. Il est essentiel pour les établissements de crédit de maximiser la gestion des risques afin de réduire les pertes financières et temporelles. Cela peut être réalisé en utilisant une intelligence artificielle qui peut classer un client en fonction de sa qualité de dossier en matière de crédit bancaire.

Les données de grandes dimensions sont générées par cette classification automatique qui découle des combinaisons de classifieurs, ce qui offre une grande probabilité d'obtenir des résultats stables et fiables. Les caractéristiques discriminantes et pertinentes peuvent être extraites grâce à une analyse discriminante linéaire (LDA, ADL) afin de diminuer le temps d'exécution, les redondances et le bruit.

L'étude vise à anticiper le risque de remboursement de crédit des futurs emprunteurs de la banque afin de ne donner le crédit qu'aux emprunteurs forts potentiels qui sont considérés comme des clients fiables. Cette étude permet au banquier de préserver les sommes destinées aux clients en difficulté de remboursement pour les concentrer sur d'autres objectifs.

2. MATÉRIEL ET MÉTHODES

2.1. Site d'étude

Cette étude repose sur des données simulées tirées d'une expérience des spécialistes du domaine bancaire. Quelques données secondaires ont été tirées de l'expérience réalisée par Eco Bank, une banque spécialisée en Afrique subsaharienne (Baketi, 2021). Elle propose des services de banque de détail, de banque de grande clientèle et d'investissement, ainsi que des services bancaires transactionnels aux États, aux établissements financiers, aux multinationales, aux entreprises locales, aux petites et moyennes entreprises (PME) et aux particuliers.

2.2. Crédit bancaire

Dans le secteur bancaire, un prêt bancaire consiste à offrir des fonds à une date ou une période spécifique contre l'engagement de rembourser moyennant une rémunération. Il s'agit d'un prêt qui est conclu par un contrat entre un emprunteur et un prêteur. Les banques jouent un rôle essentiel dans la fourniture de crédit, que ce soit aux individus ou aux entreprises (Louis-Ferdinand Céline, 1985).

2.3. Typologie des crédits bancaires

Différents genres de prêts bancaires peuvent être différenciés en fonction du critère sélectionné par l'analyste selon l'objet, la durée et les caractéristiques du prêt (Lasserre *et al*., 2012).

a. Selon l'objet du crédit

Les crédits pour les particuliers (personnes physiques) : (i) Crédit-bail (leasing, location-vente), (ii) Crédit à la consommation (affecté, personnel, revolving), (iii) Crédit immobilier (Epargne logement).
Les crédits pour les entreprises et les professionnels : (i) Crédit d'exploitation (escompte, faculté de caisse, affacturage, Credoc), (ii) Crédit d'investissement (prêt d'équipement, crédit-bail).

b. Selon la durée du crédit

La durée de crédit pour personne physiques : (i) Crédit à très court terme (jusqu'à 3 mois), (ii) Crédit à court terme (jusqu'à 2 ans), (iii) Crédit à moyen terme (jusqu'à 7 ans), (iv) Crédit à long terme (jusqu'à 20 ans, (v) Crédit à très long terme (au-delà de 20 ans, voire perpétuel.

c. Selon la forme du crédit (les caractéristiques)

Selon la forme de crédit (Lasserre Capdevielle et Storck, 2012*)*, on a : (i) Monnaie : En Monnaie nationale, en devises étrangers ; (ii) Mode d'amortissement : constant, à annuité constante ou remboursable infinie ; (iii) Type de taux : A taux fixe, à taux variable ou indexé, à taux variable capé ; (iv) Mécanisme : permanent ou revolving, sur ligne de crédit ; (v) Contrat : sur compte débiteur, sur contrat de prêt, emprunt obligation, en pool ; (vi) Garantie : En blanc, garanti.

d. Remboursement du crédit

Le second processus s'articule autour des activités suivantes (www.legifrance.gouv.fr, 2020) : (i) Appel d'échéances ; (ii) Remboursement d'échéances normales ; (iii) Remboursement anticipé des échéances ; (iv) Clôture et délivrance des mains levées ; (v) e- Recouvrement.

Le recours à la troisième phase s'opère généralement selon la logique suivante (Nouy, 1993) : (i) Constatation des impayés ; (ii) Renégociation des conditions ; (iii) Classement des crédits en créance douteuse ; (iv) Contentieux (Gérard et Philippe, 2012).

2.4. Matériel technique

L'apprentissage artificiel ou machine learning est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques pour donner aux ordinateurs la capacité d'apprendre à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune (sage-automatique.html, 2020).

L'apprentissage artificiel fait référence au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques (https://www.techno-science.net/glossaire-definition/Apprentis, 2022).

L'apprentissage artificiel est une tentative de comprendre et de reproduire cette faculté d'apprentissage dans des systèmes artificiels (Figure 1). Il s'agit, très schématiquement, de concevoir des algorithmes capables, à partir d'un nombre important d'exemples (les données correspondant à l'expérience passée), d'en assimiler la nature afin de pouvoir appliquer ce qu'ils ont ainsi appris aux cas futurs (Markoff, 2011).

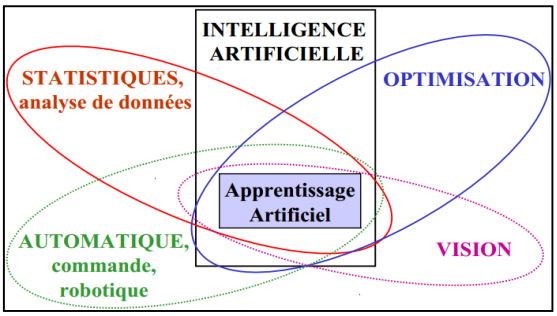

Figure 1. Nature multidisciplinaire de Machine learning (MINES ParisTech, mai 2011)

2.5. Types d'apprentissages

Apprentissage supervisé

L'apprentissage supervisé est une technique artificielle où l'on cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des « exemples » (en général des cas déjà traités et validés) où l'on agit sur les données tout en ayant les hypothèses ou les informations supplémentaires fournies sur les données par le superviseur (Turing, 2023).

En cas de prédétermination des classes et de connaissance des exemples, le système acquiert la capacité de classer en utilisant un modèle de classement (Richard, 2001); on parle alors d'apprentissage supervisé (ou d'analyse discriminante). Un expert (ou oracle) doit s'assurer de bien étiqueter des exemples au préalable comme illustré dans la figure 2 ci-dessous. L'apprenant peut donc découvrir ou approximativement identifier la fonction qui permet d'attribuer la bonne "étiquette" à ces exemples (Zakia Messaoudi, 2020). Il n'est pas nécessaire d'associer une donnée à une classe

unique, mais plutôt à une probabilité d'appartenance à chacune des classes prédéterminées (Bousquet *et al*., 2004).

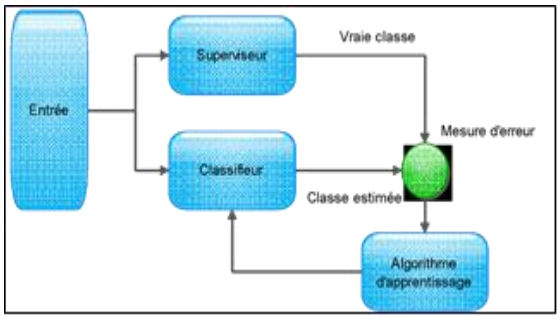

Figure 2. Schéma de l'apprentissage supervisé (Kafunda, 2017)

Les méthodes utilisées sont nombreuses parmi lesquelles il y a : (i) Boosting, (ii) Réseau des neurones (Montanez-Barrera, 2022), (iii) SVM (Support Vector Machine) ou MVS (ou Machines à Vecteurs Supports), (iv) Arbres de décision, (v), Etc.

Apprentissage non-supervisé

Lorsque le système ou l'opérateur ne possède que des échantillons, mais non étiquetés, et que le nombre de classes et leur nature n'ont pas été préétablis, on parle d'apprentissage non supervisé ou de répartition. Aucun spécialiste n'est disponible ni nécessaire. La structure plus ou moins dissimulée des données doit être découverte par l'algorithme lui-même. L'apprentissage en clustering est un algorithme non supervisé la figure 3 ci-dessous présente une modélisation d'un apprentissage non supervisé.

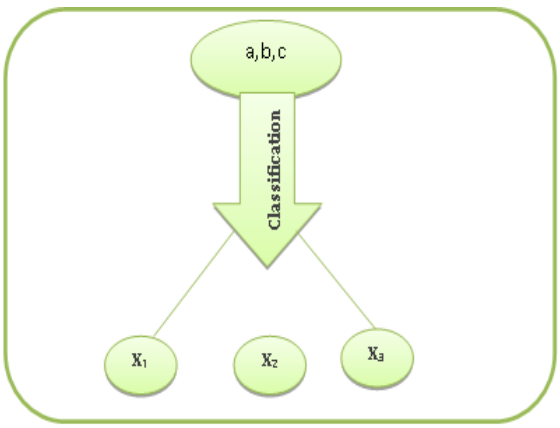

Figure 3. Apprentissage non supervisé

2.6. Combinaison des classifieurs automatiques

La combinaison d'informations consiste à regrouper des informations issues de plusieurs sources afin d'améliorer la prise de décision (Sylvie-Nuria, 2018).

Dans cette recherche, le résultat est obtenu en combinant divers classifieurs, notamment le réseau neuronal, la machine supportée par des vecteurs (SVM), l'arbre décisionnel et bien d'autres. Ces données des classifieurs peuvent être issues de diverses sources ou posséder des attributs variés, tirés des mêmes données d'origine.

D'une manière générale, les techniques de combinaison (ou fusion) des informations sont divisées en systèmes répartis et centralisés selon que les conclusions tirées par les différents experts, sont combinées, ou que les différentes sources d'informations sont utilisées ensemble par un mécanisme d'inférence simple (Bousquet et Boucheron, 2004). Il s'agit ici de combiner les résultats de classifieurs en proposant un schéma de combinaison des systèmes répartis. D'où, le terme « combinaison de classifieurs ».

Voici en quelque sorte les différents intérêts de la combinaison des classifieurs : (i) La distribution des caractéristiques sur des classifieurs adaptés ; (ii) L'exploitation de la complémentarité entre classifieurs ; (iii) La prise en compte des performances de chacun des classifieurs ; (iv) La réduction de l'importance des certains choix initiaux.

A cela, plusieurs chercheurs font les constats suivants : (i) Il n'existe pas de « meilleur » classifieur capable de traiter (apprendre) n'importe quelle distribution des données d'apprentissage ; (ii) Aucun classifieur ne peut discriminer suffisamment un ensemble important de classes ; (iii) Le « réglage » d'un classifieur est un problème extrêmement difficile (on procède souvent par essai/erreur).

Il existe trois types d'architectures de combinaisons de classifieurs :

(i) La combinaison séquentielle ou en série
Cette combinaison consiste en une structure en différents niveaux de décision qui permet de diminuer progressivement le nombre de classes envisageables. La prédiction est effectuée grâce à h(t), ce qui signifie que le classifieur qui prédit correctement sera pondéré positivement (+P) tandis que l'individu (l'instance) mal classé sera marqué par un signe négatif (-α). On appelle pondération le poids calculé en fonction de la distance entre le séparateur tracé par le premier classifieur. Selon Ng et Jordan (2002), un seul classifieur tient compte de la réponse fournie par le classifieur précédent à chaque niveau.
(ii) La combinaison parallèle
Les classifieurs opèrent indépendamment les uns des autres puis on fusionne leurs réponses respectives. L'idée de base est de chercher un point d'accord entre les classifieurs pour aboutir à une décision unique. Le problème de la combinaison de classifieur en parallèle est celui de savoir comment élaborer une réponse finale unique à partir des k résultats fournis par k classifieurs.

(iii) La combinaison hybride (mélange de la combinaison séquentielle et la combinaison parallèle)

L'approche hybride consiste à combiner à la fois des architectures séquentielles et parallèles afin de tirer pleinement avantage de chacun des classifieurs utilisés présente un exemple de combinaison hybride dans laquelle on combine un classifieur en série avec deux classifieurs en parallèle (Bikmukhametov, 2020).

(iv) Vote majoritaire

Le jugement majoritaire est une méthode de vote par valeurs (les électeurs attribuent une mention à chaque candidat et peuvent attribuer la même mention à plusieurs candidats) pour laquelle la détermination du gagnant se fait par la médiane plutôt que par la moyenne. Contrairement aux méthodes utilisant la moyenne, le jugement majoritaire utilise des échelles de mentions verbales plutôt que numériques pour évaluer les candidats. Cette possibilité permet, d'après les inventeurs du jugement majoritaire, d'offrir aux électeurs des mentions dont les acceptions sont plus homogènes parmi les électeurs (Tom Mitchell, 1997).

Le vote majoritaire consiste à faire voter de manière indépendante plusieurs classifieurs et à retenir la réponse ou la sortie majoritaire, correspondant à la meilleure solution, c'est-à-dire celle qui ne sera pas trop loin de réalité recherchée. Ceci est issu du fait que les classifieurs pris individuellement sont incapables d'assurer une unique vérité pour toutes les décisions.

Un vote de majorité est une méthode de combinaison de classifieurs ou l'on cherche à optimiser la pondération des votants. L'objectif est alors de construire un vote final plus performant et plus robuste que les votants individuels. Cependant, choisir des poids pertinents est une tâche parfois complexe.

Un système multi-classifieur (Multiple Classifier System : MCS en anglais) est constitué d'un ensemble de différents classifieurs et d'une fonction de décision pour combiner leurs sorties. La description d'un Système Multi-Classifieur suit les deux phases suivantes : (i) Générer un ensemble de classifieurs complémentaires qui peuvent être combinés pour arriver à une solution optimale ; (ii) Définir la fonction de combinaison pour donner une décision finale.

Le choix de la fonction de décision joue un rôle très important dans la conception d'un système multi-classifieur. La fonction de décision peut être conçue comme étant une fonction de combinaison, par conséquent la sortie du MCS reflète la décision de tout l'ensemble en utilisant par exemple le vote majoritaire, la somme pondérée, etc. ou bien comme étant une fonction de sélection dynamique d'un classifieur, dans ce cas, il faut avoir au moins un classifieur dans l'ensemble qui pourra classer correctement un individu à l'entrée (Artières, 2008)

Le regroupement des résultats se fait généralement par vote, c'est-à-dire par une combinaison linéaire des décisions suivie d'une décision ferme. En effet, pour un problème à K classes, on peut soit combiner des classificateurs partiels, qui ne savent par exemple que distinguer une classe d'une autre, soit combiner des classificateurs complets (Maloof , 2006).

Après avoir obtenu les prédictions des classes de chaque classifieur, la combinaison par vote majoritaire simple consiste à choisir la classe la plus proposée par les classifieurs. Un résultat de rejet est généré si toutes les classes ont le même nombre de votes. La condition ici est que chaque classifieur doit donner un résultat différent des deux autres (Dash, 2022).

Pour prendre en considération l'importance que peut avoir un classifieur par rapport aux autres, nous utilisons une pondération afin de représenter cette notion d'importance. Alors, la combinaison consiste à faire la somme du produit des classifieurs avec les pondérations qui leurs sont associées (Mitchell, 2005).

La majorité des combinaisons de classifieur appliquent un seul algorithme dans l'apprentissage de base, ce qui se traduit par une homogénéité chez tous les classifieurs. Les classifieurs homogènes font référence aux classifieurs du même type, avec des qualités similaires (Guo, 2022).

D'autres techniques utilisent différents classifieurs, ce qui crée des ensembles hétérogènes. Les classifieurs hétérogènes peuvent être de diverses catégories (Markoff, 2011). Diverses techniques de combinaison de classifieurs séquentiels, parallèles et hybrides ont été suggérées, telles que le bagging, le boosting et le staking, dont le bagging et le boosting sont les plus répandues. Pour notre travail, nous ne traiterons que du Boosting car c'est ce qui est étudié.

2.7. Mesures de performance d'un classifieur

Pour une entrée donnée, un classifieur peut générer les réponses suivantes (http://www.college-de-france.fr/site/stanislas-dehaene/_course.htm , juin 2022) : (i) Un rejet (pour indiquer que le classifieur n'a pas pu identifier cette entrée) ; (ii) Une reconnaissance (dans ce cas, il identifie bien l'entrée, et il lui attribue sa classe appropriée) ; (iii) Une substitution (le classifieur attribue une autre classe à l'entrée).

Complexité d'un classifieur

L'objectif de la théorie de la complexité consiste à déterminer si un problème peut être efficacement

résolu par un ordinateur. En général, l'utilisation de certains algorithmes est limitée par le temps de calcul excessif, ce qui signifie que la mesure de complexité la plus précieuse est la complexité temporelle (Ng et Jordan, 2022).

Complexité temporelle

La complexité temporelle d'un algorithme A est la fonction Temps où Temps (x) est le nombre d'instructions exécutées pendant le calcul A(x). On aimerait définir la complexité temporelle d'un problème comme étant la complexité temporelle de l'algorithme A, le plus efficace pour résoudre mais il est possible de démontrer qu'on peut toujours rendre un algorithme A plus efficace. La complexité temporelle doit être asymptotique. Pour toute fonction f, la complexité temporelle d'un langage L est en O(f) s'il existe un algorithme A qui décide L et des constantes n_0, n_1, …..n_n (https://www.college-de-france.fr/site/yann-lecun/Recherches-sur-l-intelligence-artificielle.htm, juin 2022).

$$\forall x|x| > n_o \Rightarrow Temps_A(x) \leq cf(|x|)$$

3. RÉSULTATS

Le dataset initial comprend 1000 entrées qui possèdent 20 attributs catégoriels/symboliques élaborés par Hofmann (2002). En ce qui concerne cette étude, chaque entrée correspond à une personne qui contracte un prêt auprès d'une banque. Ainsi, chaque personne est classée en fonction de son dossier prêt selon la description des données (sage-automatique.htmlhttps://www.techno-science.net/glossaire-definition/Apprentis (consulté en février 2022).

La compréhension de l'ensemble des données d'origine est quasiment impossible en raison de son système complexe de catégories et de symboles. Un petit script Python a été développé afin de le convertir en un fichier CSV accessible. Plusieurs colonnes sont tout simplement négligées, soit parce qu'elles ne sont pas significatives, soit parce que leurs descriptions sont peu claires.

3.1. Implémentation en python

Importation des bibliothèques et du dataset

```
# Bibliothèques utilisées
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
#Importiion des données

df_credit = pd.read_csv("german_credit_dataok.csv",index_col=0)
```

3.2. Visualisation des données

```
print(df_credit.info())
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1000 entries, 0 to 999
Data columns (total 10 columns):
Age 1000 non-null int64
Sex 1000 non-null object
Job 1000 non-null int64
Housing 1000 non-null object
Saving accounts 817 non-null object
Checking account 606 non-null object
Credit amount 1000 non-null int64
Duration 1000 non-null int64
Purpose 1000 non-null object
Risk 1000 non-null object
dtypes: int64(4), object(6)
memory usage: 85.9+ KB
None
```

3.3. Explorations

Examen de la variable cible et de sa distribution.

```
import plotly.offline as py
```

```
#ce code, nous permet de travailler sur la version offline de plotly
py.init_notebook_mode(connected=True)
# C'est comme "plt" de matplot
import plotly.graph_objs as go
# Il est utile d'obtenir des outils plotly
import plotly.tools as tls
# la bibliothèque sera utilisée pour ignorer certains avertissements
import warnings
# Pour faire le compteur de certaines fonctionnalités
from collections import Counter

trace0 = go.Bar(
    x = df_credit[df_credit["Risk"]== 'good']["Risk"].value_counts().index.values,
    y = df_credit[df_credit["Risk"]== 'good']["Risk"].value_counts().values,
    name='Bon dossier de prêt')

trace1 = go.Bar(
    x = df_credit[df_credit["Risk"]== 'bad']["Risk"].value_counts().index.values,
    y = df_credit[df_credit["Risk"]== 'bad']["Risk"].value_counts().values,
    name='Mauvais dossier de prêt')
```

La figure 4 ci-dessous offre une vue d'ensemble du risque lié au remboursement des prêts, basée sur notre base de données.

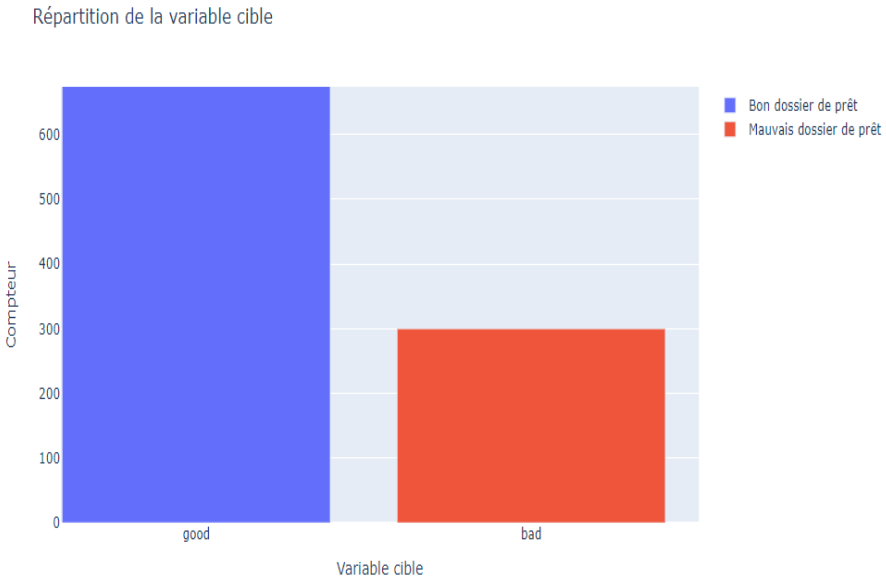

Figure 4. Répartition de la variable cible

A partir de la figure 5 ci-dessous, il est évident que les sommes les plus importantes ont été attribuées aux prêteurs ayant des durées les plus longues. Cette période varie entre [12, 18 et 24] mois.

```
import plotly.figure_factory as ff
import numpy as np
# Add histogram data
x1 = np.log(df_good['Credit amount'])
x2 = np.log(df_bad["Credit amount"])
# Group data together
hist_data = [x1, x2]
group_labels = ['Bon dossier', 'Mauvais dossier']
# Create distplot with custom bin_size
fig = ff.create_distplot(hist_data, group_labels, bin_size=.2)
plt.figure(figsize = (8,5))
g= sns.distplot(df_good['Credit amount'], color='r')
g = sns.distplot(df_bad["Credit amount"], color='g')
```

```
g.set_title("Montant du crédit par Répartition des fréquences", fontsize=15)
plt.show()
```

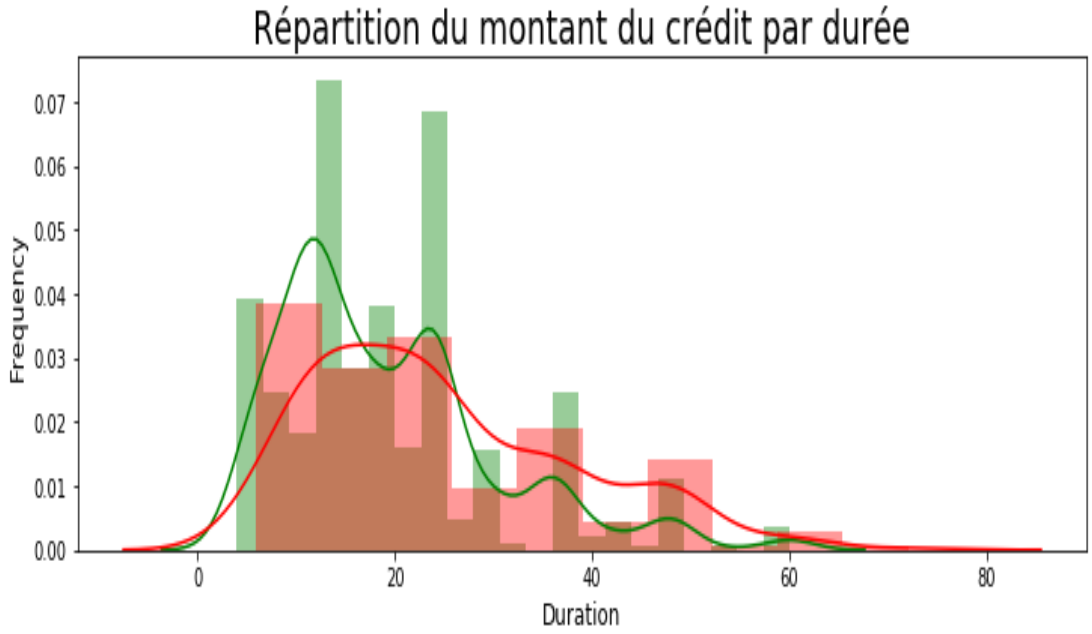

Figure 5. Répartition du montant du crédit par durée

3.4. Le prétraitement des données

```
# preparation des modèles
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('RF', RandomForestClassifier()))
models.append(('SVM', SVC(gamma='auto')))
models.append(('XGB', XGBClassifier()))
# evaluation de chaque modèle à tour de rôle
results = []
names = []
scoring = 'recall'
for name, model in models:
    kfold = KFold(n_splits=10, random_state=seed)
    cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

Il est observé par la figure 6 ci-dessous que la plupart des modèles présentent une faible valeur de retour, cependant, les meilleurs résultats ont été obtenus avec les modèles CART, NB et XGBoost. Quelques modèles ont été mis en place et tentés de les régler de manière simple.

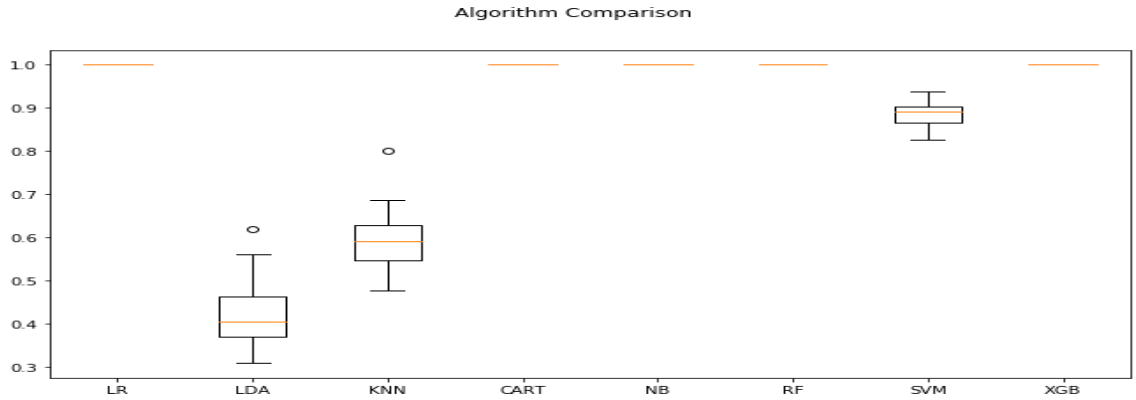

Figure 6. Combinaison des classifieur automatique

3.5. Modèle 1 : forêt aléatoire pour prédire le pointage de crédit

```
rf = RandomForestClassifier(max_depth=None, max_features=10, n_estimators=15, random_state=2)
#trainning with the best params
rf.fit(X_train, y_train)
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
        max_depth=None, max_features=10, max_leaf_nodes=None,
        min_impurity_decrease=0.0, min_impurity_split=None,
        min_samples_leaf=1, min_samples_split=2,
        min_weight_fraction_leaf=0.0, n_estimators=15, n_jobs=None,
        oob_score=False, random_state=2, verbose=0, warm_start=False)
#Tester le modèle
#Prédire avec notre modèle
y_pred = rf.predict(X_test)
# Vérification les résultats obtenus
print(accuracy_score(y_test,y_pred))
print("\n")
print(confusion_matrix(y_test, y_pred))
print("\n")
print(fbeta_score(y_test, y_pred, beta=2))
1.0 = 100% pour l'entrainement
```

Matrice de confisions

```
[[178  0]
 [ 0  72]]
```

Dans la matrice de confisions ci haut, il est évident que tous les individus ont été correctement classés. Il n'y a pas eu de faux rejets, ce qui démontre une précision absolue. En effet, 178 personnes sont classées comme ayant des dossiers de prêt satisfaisants et les 72 autres personnes ont des dossiers de prêt insatisfaisants. Ainsi, aucun individu n'a été mal classifié.

3.6. Model 2 : le modèle gaussien

```
from sklearn.utils import resample
from sklearn.metrics import roc_curve
# Criando o classificador logreg
GNB = GaussianNB()
# Fitting with train data
model = GNB.fit(X_train, y_train)
# Printing the Training Score
print("Training score data: ")
print(model.score(X_train, y_train))
Training score data: 1.0
```

Le score de la base d'entrenement est de 100%

```
y_pred = model.predict(X_test)
print(accuracy_score(y_test,y_pred))
print("\n")
print(confusion_matrix(y_test, y_pred))
print("\n")
print(classification_report(y_test, y_pred))  1.0 = 100%
```

Matrice de confisions

```
[[178  0]
[  0 72]]
```

La matrice de confisions du modèle gaussien indique une amélioration de la prédiction, comme dans le cas précédent, 178 dossiers de prêt positifs et 72 dossiers de prêt négatifs. L'exactitude mathématique est à 100 %.

Vérification de la courbe ROC

```
y_pred_prob = model.predict_proba(X_test)[:,1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
# Tracer la courbe ROC
plt.plot([0, 1], [0, 1], 'k--')
plt.plot(fpr, tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('le modèle ROC Curve')
plt.show()
from sklearn.metrics import roc_auc_score
y_pred_prob = model.predict_proba(X_test)[:,1]
print(f"ROC_AUC score: {roc_auc_score(y_test, y_pred_prob)}")
```

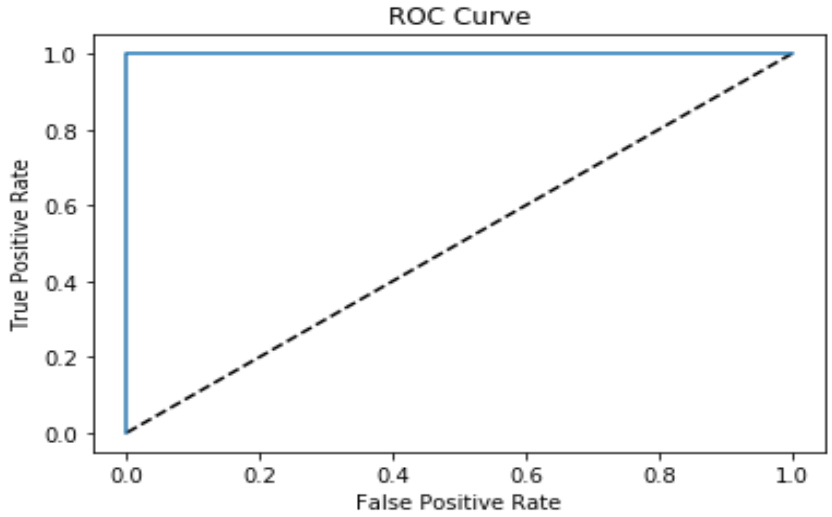

Figure 7. Courbe Roc

Matrice de confisions

```
[[149  29]
 [ 41  31]]
```

Sur la figure 7 en revanche, il est à noter que les individus ont été correctement classés avec 41 cas de réjections injustifiées pour les dossiers de prêt satisfaisants et 29 cas pour les dossiers insatisfaisants. Ainsi, l'exactitude mathématique est revue à 72 % pour ce classifieur.

3.7. Modèle 3 : Modèle CART : arbres de décision

```
model_xg = CARTClassifier(random_state=2)

grid_search = GridSearchCV(model_xg, param_grid=param_test1, cv=5, scoring='recall')
grid_search.fit(X_train, y_train)
GridSearchCV(cv=5, error_score='raise-deprecating',
    estimator=XGBClassifier(base_score=None, booster=None, callbacks=None,
    colsample_bylevel=None, colsample_bynode=None,
    colsample_bytree=None, early_stopping_rounds=None,
    enable_categorical=False, eval_metric=None, gamma=None, gpu_id=None,
    grow_policy=None, importance_type=...,
    tree_method=None, use_label_encoder=False, validate_parameters=None,
    verbosity=None),
    fit_params=None, iid='warn', n_jobs=None,
    param_grid={'max_depth': [3, 5, 6, 10], 'min_child_weight': [3, 5, 10], 'gamma': [0.0, 0.1, 0.2, 0.3, 0.4],
'subsample': [0.75, 0.8, 0.85], 'colsample_bytree': [0.75, 0.8, 0.85]},
    pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
    scoring='recall', verbose=0)
grid_search.best_params_
{'colsample_bytree': 0.75,
 'gamma': 0.0,
 'max_depth': 3,
 'min_child_weight': 3,
 'subsample': 0.75}
```

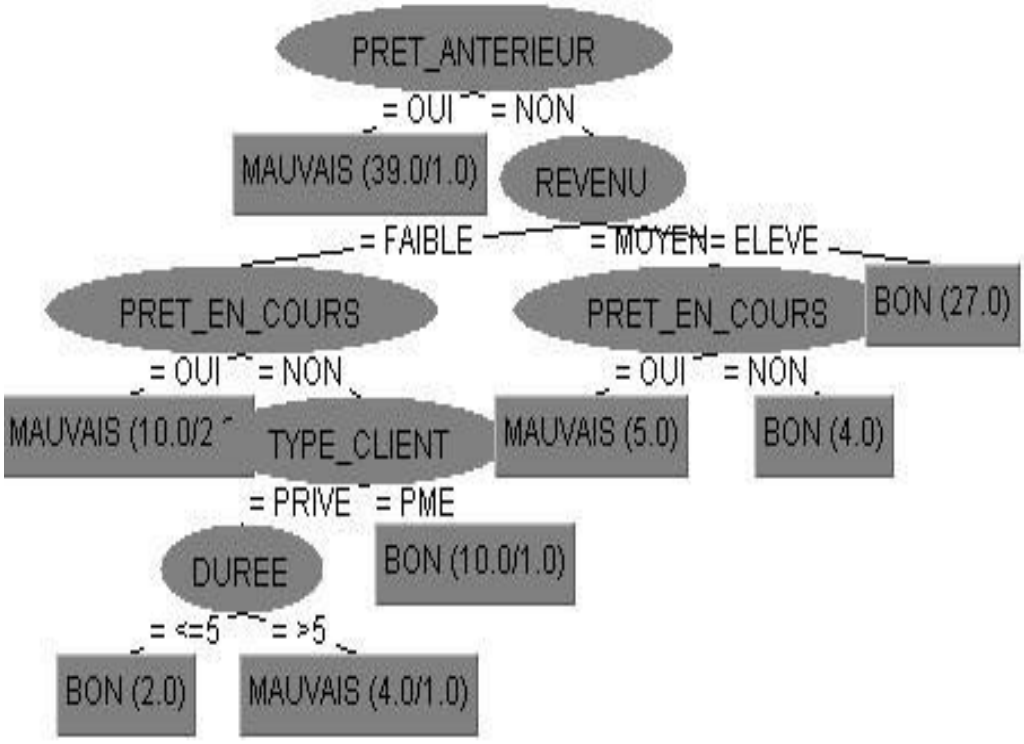

Figure 8. Le graphe de l'arbre de décision

L'arborescence décisionnelle de la base d'entraînement est illustrée par le graphique présenté dans la figure 8. Cela indique une orientation idéale pour un classement judicieux du nouvel emprunteur en se basant sur le critère principal qui est le prêt précédent. Par conséquent, pour chaque demande de prêt, la machine contrôle si le bénéficiaire possède un prêt précédent non réglé dans les délais prévus. Si c'est le cas, le dossier est recalé. En revanche, si ce n'est pas le cas, la machine examinera le revenu du bénéficiaire, qu'il soit bas, modéré ou important.

Les classifieurs sont influencés par la diversité des entrées qui leur sont présentées pendant la phase d'apprentissage, ainsi que par la pertinence et la qualité de ces entrées. Il y a une multitude de

caractéristiques qui peuvent être extraites d'une donnée d'entrée. Étant donné l'importance de ces caractéristiques pour distinguer les données à analyser, il serait judicieux de tirer parti de la diversité des informations qu'elles fournissent afin d'éviter les facteurs subjectifs des différents classifieurs qui peuvent entraîner des résultats erronés. Il est donc essentiel pour le banquier de réaliser un diagnostic fiable et valide afin de repérer les clients pouvant être considérés comme crédibles.

Cette étude repose sur des données simulées tirées d'une expérience des spécialistes du domaine bancaire. L'augmentation de l'espace de données entraînera une amélioration de la qualité des résultats et conduira les banques à recourir aux systèmes intelligents afin de rationaliser leurs décisions concernant les prêts bancaires.

4. DISCUSSION

Combiner plusieurs classifieurs, en les faisant voter par exemple, permet souvent de dépasser la performance de chacun des classifieurs pris isolément (Artieres, 2008)

Lorsqu'on embarque dans un processus d'apprentissage automatique, il est important de garder à l'esprit qu'il se peut que les données ne permettent peut-être pas d'arriver à un meilleur résultat que celui qu'est déjà détenu avec les méthodes classiques en place. Mais cela révèle assurément des informations intéressantes lors du processus d'analyse de données (Zakia, 2020).

Un autre facteur à prendre en compte est le compromis entre l'interprétabilité et l'efficacité. Certains algorithmes, tels que les arbres de décision, permettent l'interprétation et fournissent des explications claires sur leurs prédictions. D'autres algorithmes, tels que les réseaux neuronaux, peuvent être plus performants, mais manquent d'interprétabilité (8 algorithmes d'apprentissage automatique pour la prédiction, 13.05.2024).

Dans cette étude, il a été proposé une approche innovante qui combine différents algorithmes de l'apprentissage supervisé, tels que les réseaux de neurones, l'arbre de décision, le séparateur à vaste marge (SVM), la forêt aléatoire, le K-Neighbors Classifier (KNN) et la régression logistique, afin d'analyser l'impact de la classification sur l'ensemble de ces classifieurs, tout en évaluant la marge d'erreur de différentes prédictions. L'utilisation de la phase de visualisation a permis de trouver de nouvelles informations à partir des ensembles de données initiales, ce qui pourrait contribuer à améliorer la précision de la classification supervisée.

5. CONCLUSION

L'application de post-traitements sur les résultats obtenus aura un impact significatif sur l'amélioration de la qualité de la prise de décision concernant les prêts bancaires. Effectivement, l'association de différents classifieurs automatiques revêt une grande importance.

L'objectif de cette étude est d'améliorer la qualité de la classification des dossiers de prêts dans des situations où il y a peu d'informations prédictives, déterministes et fiables sur l'état et la nature de la formation des dossiers de prêts à l'instant de leur dépôt.

Il a été démontré que la combinaison de classifieurs améliore considérablement les performances du système de classification automatique par rapport à chaque classifieur pris individuellement. Il est également crucial de choisir les classifieurs afin d'atteindre une meilleure performance avec un nombre minimal de classifieurs sélectionnés.

La méthode du vote majoritaire des classifieurs a été suggérée afin d'améliorer la classification des dossiers de prêt des clients. L'évidence est une théorie fondée sur le principe de combinaison parallèle des systèmes répartis. Elle est universelle et s'applique à tous les types de classifications.

Références

Alan T., 1936. On computable numbers, with an application to the entscheidungs problem. *Proceedings of the London Mathematical Society, s2-42, 12 novembre 1936*, pp. 230-265. DOI 10.1112/plms/s2-42.1.230.

Barger P., 2003. *Evaluation et Validation de La Fiabilité et de la disponibilité des Systèmes D'Automatisation à Intelligence Distribuée, en Phase Dynamique*. Thèse de Doctorat de l'UHP Nancy 1, France,

Bikmukhametov T.,2020. *Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models, Computers & Chemical Engineering*, 63 p.

Bousquet O., Boucheron S. & Lugosi G., 2004. Introduction to Statistical Learning Theory, Springer, Heidelberg, Germany . Lecture Notes in *Artificial Intelligence*, pp. 169–207.

Charles P.-D., 1967. *Le risque de crédit bancaire*, Edition scientifique Ribier, Paris, 18 p.

Dash T., 2022. *Review of some techniques for inclusion of domain-knowledge into deep neural networks*, Scientific Reports, 16 p.

Gérard B. et Philippe F., *Crédit à la consommation : protection du consommateur, Paris, Delmas Express*, 200 p.

Guo S., 2020. *An introduction to Surrogate modeling, Part I: fundamentals » , sur Towards Data Science*, 12 p.

Hoffman D. G., *Managing operational risk: 20 firmwide best practice strategies*, books, 2002, 526p.

Lasserre Capdevielle J. & Storck M., 2012. *Le crédit aspects juridiques et économiques*. Paris, Dalloz, 210 p.

Louis-Ferdinand C., 1985. *Mort à crédit*. Paris, Gallimard, coll. « Folio », 622 p.

Lyonnet P., 2006. *Ingénierie de la Fiabilité.* Edition Tec et Doc, Lavoisier, Paris, France, 71 p.

Maloof M. A., 2022. *Machine Learning and Data Mining for Computer Security.* Springer, 47 p.

Markoff J., 2011. On 'Jeopardy!' Watson Win Is All but Trivial. *The New York Times*, 16 février 2011.

Mitchell T., 1997. *Machine Learning.* Wiley-interscience, 72 p.

Mitchell T., 2005., *Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*, Draft Version, 83 p.

Montanez-Barrera, J. A., *Correlated-informed neural networks: A new machine learning framework to predict pressure drop in micro-channels*, International Journal of Heat and Mass Transfer, 2022.

Ng A.Y. & Jordan M.I., 2002. *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes*. in NIPS 14, 2002.

Nouy D., 1993. La rentabilité des banques françaises. *Revue d'économie financière*, 27, 465-486.

Richard O., Duda Peter E., Hart David G., 2001. *Stork, Pattern Classification, Wiley-interscience*, 39 p.

Sylvie-Nuria N., , 2018. *Donnez du sens à vos décisions : 7 clés pour discerner et faire les bons choix Broché*. Grand livre, Eyrolles, 240 p.

Thierry A. 2008. *Apprentissage Automatique.* Big Data et Data Science, ECM ML - M2 IAAA, 50 p.

Y. Ng and M. I. Jordan, 2002. *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes*. in NIPS 14, 2002.

Zakia M., 2020. *Les 3 étapes essentielles de l'apprentissage automatique (Machine Learning)*, 49 p.