République Tunisienne

Ministère de l'Enseignement Supérieur

et de la Recherche Scientifique

Université de Tunis El Manar

Institut Supérieur d'Informatique d'El Manar

Rapport de Projet de Fin d'Études

Présenté en vue de l'obtention du

Diplôme National de licence en sciences informatiques

Spécialité : Génie Logiciel et Système d'Information (cs)

Par

Malek Bokri

mise en place d'une application bancaire pour la classification des très petites entreprises à base de machine learning

Encadrant professionnel : **Monsieur Nader Trigui**

Encadrant académique : **Monsieur Sahbi Bahroun**

Réalisé au sein de Banque Internationale Arabe de Tunisie

Année Universitaire 2022 - 2023

République Tunisienne

Ministère de l'Enseignement Supérieur

et de la Recherche Scientifique

Université de Tunis El Manar

Institut Supérieur d'Informatique d'El Manar

Rapport de Projet de Fin d'Études

Présenté en vue de l'obtention du

Diplôme National de licence en sciences informatiques

Spécialité : Génie Logiciel et Système d'Information (cs)

Par

Malek Bokri

mise en place d'une application bancaire pour la classification des très petites entreprises à base de machine learning

Encadrant professionnel : **Monsieur Nader Trigui**

Encadrant académique : **Monsieur Sahbi Bahroun**

Réalisé au sein de Banque Internationale Arabe de Tunisie

Année Universitaire 2022 - 2023

J'autorise l'étudiant à faire le dépôt de son rapport de stage en vue d'une soutenance.

Encadrant professionnel, **Monsieur Nader Trigui**

Signature et cachet

J'autorise l'étudiant à faire le dépôt de son rapport de stage en vue d'une soutenance.

Encadrant académique, **Monsieur Sahbi Bahroun**

Signature

Dédicaces

Je dédie ce travail à :

Monsieur **Monsieur Nader Trigui**, Monsieur **Monsieur Sahbi Bahroun** pour m'avoir encadré et fait de leurs mieux afin de m'aider.

etc.

Malek Bokri

Remerciements

Je remercie

Je suis reconnaissant

J'exprime ma gratitude

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

— **ACT_COURANT** = Actif courant

— **AUC** = area under curve

— **AUT_FIN** = Auto-financement

— **BFR** = Besoin de fonds de roulement

— **BIAT** = banque internationale arabe de tunisie

— **CA** = Chiffre d'affaires

— **CAP_PROP** = Capitaux propres

— **CFN** = Cash-flow net

— **COUV_DET** = Couverture

— **DS** = Dette structurée

— **ENN** = edited nearest neighbor

— **FPN** = Fonds propres net

— **FPR** = False positive rate

— **FR** = Fonds de roulement

— **LIQ** = Liquidité

— **LIQ_COUR** = Liquidité courante

— **LIQ_IMMEDIATE** = Liquidité immédiate

— **MLP** = multilayer perceptron

— **PASS_COURANT** = Passif courant

— **RBF** = Radial basis function

— **RENN** = repeated edited nearest neighbor

— **RN** = Résultat net

— **RNA** = réseau de neurones artificiel

— **ROC** = receiver operating characteristic

— **ROE** = Return on equity

— **ROT_ACTIF** = Rotation des actifs

— **ROT_CAP_PROP** = Rotation des capitaux propres

— **SVM** = support vector machine

— **TOT_BL** = Bilan total

— **TPE** = très petite entreprise

— **TPR** = true positive rate

Introduction générale

Application du machine learning : Les applications de machine learning s'étendent à de nombreux domaines. C'est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données et de s'adapter à différentes situations. Les algorithmes d'apprentissage automatique peuvent analyser de grandes quantités de données et détecter des modèles, des tendances et des relations invisibles à l'œil nu. Cette capacité à gérer des quantités massives de données a de profondes implications pour de nombreux secteurs. Par exemple, les entreprises peuvent utiliser l'apprentissage automatique pour améliorer l'efficacité opérationnelle, prévoir les tendances du marché, optimiser les processus etc ... En résumé, le machine learning est une technologie qui ouvre de nombreuses perspectives d'amélioration des processus et de la prise de décision dans de nombreux domaines.

Application du machine learning pour les domaine financiérs : Dans un contexte économique, le machine learning a le potentiel de transformer le secteur bancaire et financier en fournissant des outils pour l'analyse de données à grande échelle et des prévisions plus précises. Les banques ont déjà migré vers des solutions de machine learning pour plusieurs problématiques. On peut citer la détection de fraudes avec des modèles de machine learning utilisés par "JPMorgan" et "Bank of America", qui les utilisent également pour évaluer les risques de crédit et améliorer la prévision des flux de trésorerie. Dans les marchés boursiers, on peut prendre comme exemple "Citadel", une société de gestion d'actifs et de fonds spéculatifs qui utilise le machine learning pour améliorer ses stratégies de commerce.

Problématique et principales contributions : Dans le cas ou les très petites entreprises sont souvent considérées comme des risques de crédit plus élevés que les grandes entreprises, les banques doivent être en mesure d'identifier à l'avance les clients susceptibles de ne pas rembourser leurs dettes.notre objectif dans ce projet est de réaliser un modèle de machine learning capable de prédire les défauts bancaires des TPE à partir de données historiques de clients et de transactions bancaires. Pour cela, nous allons utiliser une approche basée sur l'apprentissage supervisé, qui consiste à entraîner le modèle à partir d'un ensemble de données annotées, c'est-à-dire des données pour lesquelles on connaît déjà la réponse (défaut ou non défaut).

Structure du manuscrit : Ce rapport sera divisé en quatre chapitres, chacun se concentrant sur une étape spécifique de notre projet de prédiction de défaut bancaire des très petites entreprises.Le premier chapitre introduira le cadre général du projet, les besoins fonctionnel et non fonctionnel et les termes financiers utilisés. Le deuxieme chapitre sera consacré à l'exploration et au prétraitement des données financières que nous avons collectées.Dans le troisième chapitre, nous discuterons des différents modèles de machine learning que nous avons évalués pour notre projet ainsi que la validation de notre modèle de machine learning le plus performant, en utilisant des techniques de test et de validation.Enfin, dans le quatriéme et dernier chapitre, nous présenterons la mise en place d'un site web de prédiction de défaut bancaire des petites entreprises, expliquant en détail le processus de développement et de déploiement.

Chapitre 1

Contexte général

Plan

Introduction

Dans ce chapitre on cherche à présenter le cadre du projet tel que l'organisme d'accueil et son organigramme, faire une étude et critique de l'existant, poser la problématique, expliquer le travail demandé ainsi que le besoin fonctionnel et non fonctionnel.On vise aussi a definir les termes du domain financier utilisés dans ce projet.

1.1 Organisme d'accueil

Figure 1.1 : Banque Internationale Arabe de Tunisie

BIAT [1], Banque Universelle Créée en 1976, BIAT - Banque Internationale Arabe de Tunisie est aujourd'hui l'une des premières banques du pays, se classant au premier rang parmi de nombreuses banques sur de nombreux indicateurs. BIAT - Universal Bank développe toutes les opérations bancaires et Groupe bancaire avec des filiales dans les domaines de l'assurance et de la gestion de patrimoine, Capital investissement ou courtage boursier. Partant de la proximité et de l'engagement social, la BIAT mise sur son développement compétence et solidité au service des clients et de l'économie tunisienne.la figures 1.2 l'organigramme du BIAT

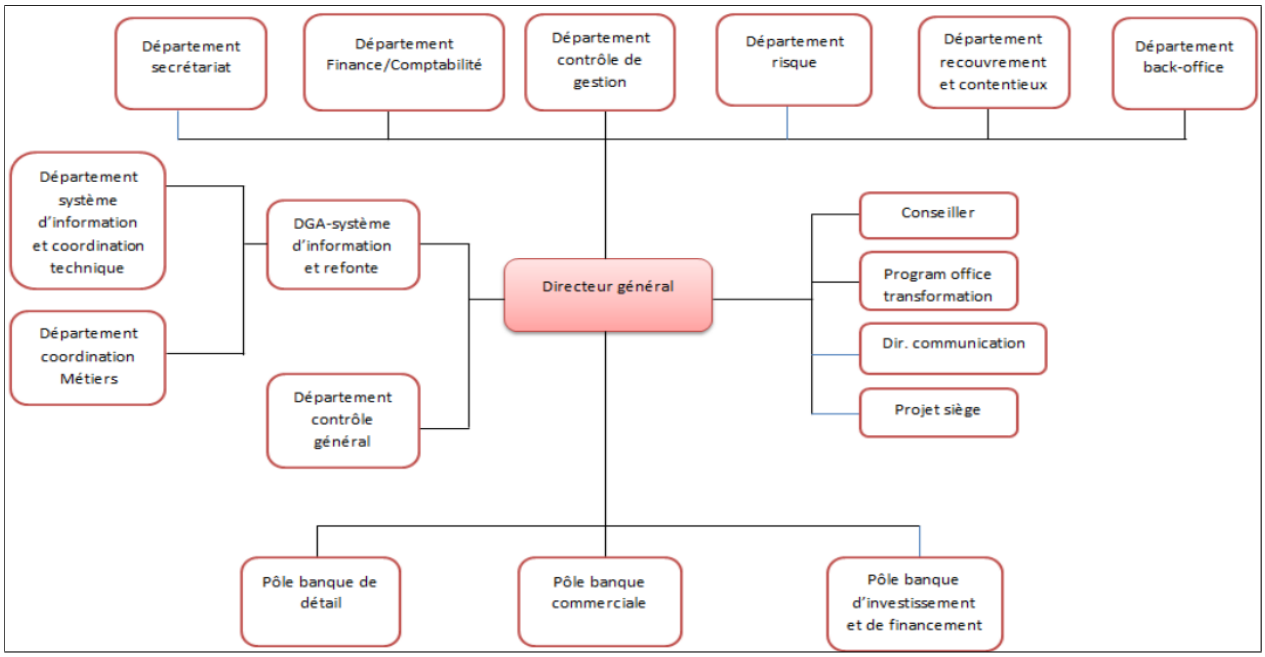

Figure 1.2 : organigramme BIAT

1.2 Étude et critique de l'éxistant

1.2.1 étude de l'existant : approche de jugement

Les prêts sont évalués de cette façon dans plusieurs banques à travers le monde,Cette approche est subjectif basé sur l'expérience des agents de crédit. En général,des points/pondérations sont attribués aux emprunteurs en fonction d'attributs spécifiés.Ceux-ci sont pondérés et convertis en un prix ou un score déterminant. La décision finale est prise par l'agent de crédit sur la base de ses informations, expérience, bon sens. En général, l'approche d'évaluation est basée sur des critères appelés 5 C :

- Caractère : antécédents et réputation de l'emprunteur
- Capitaux propres : la contribution de l'emprunteur à l'investissement
- Collatéral : une garantie pour soutenir le prêt en cas de défaut
- Capacité : capacité de l'emprunteur à rembourser le prêt
- Conditions : Rentabilité globale de l'Emprunteur

1.2.2 critique de l'existant

Bien que l'approche de jugement soit utiliser dans le scoring et la prédiction des défaut bancaire elle a ses inconvénient tel que :

— Subjectivité : L'approche de jugement avec les 5C repose sur des décideurs humains tels que les analystes de crédit. Ce qui peut affecter la décision énormément et peut entraîner des des incohérences dans les évaluations .

— méthode traditionnelle : L'approche de jugement avec les 5C est souvent basée sur des méthodes traditionnelles d'évaluation de crédit, ne profitant pas des avance technologiqueCela peut limiter sa capacité à prendre en compte de manière exhaustive tous les facteurs pertinents.

— Temps et coûts : L'évaluation basée sur les 5C peut être une approche longue qui nécessitant des ressources humaines et financières importantes. De plus, elle peut être moins adaptée pour évaluer rapidement les très petites entreprises qui ont des données financières limitées ou difficiles à obtenir.

1.3 Problématique

Comment concevoir et mettre en œuvre un puissant modèle d'apprentissage supervisé pour prédire les défauts bancaires des très petites entreprises (TPE) à partir de données économiques et financières ? Quelles sont les décisions de conception nécessaires pour obtenir les meilleurs résultats ? Comment l'efficacité du modèle va-t-elle être évaluée et comment un utilisateur pourra exploiter les prédictions de ce modèle ?

1.4 Solution proposée et objectifs globaux du projet

L'objectif du projet est de passer par toutes les étapes nécessaires pour construire un modèle fiable qui peut prédire un défaut bancaire d'une TPE. ces étapes sont :

— collecte de données fiables et pertinentes.

— exploration et analyse de ces données.

— nettoyage et préparation des données.

— choisir des modèles à entraîner et tester

— tester les modèles

— déploiement du modèle de prédiction finale

1.5 Besoins fonctionnel et non fonctionnel

1.5.1 Besoin fonctionnel

Le besoin fonctionnel de ce projet est de développer un modèle de machine learning capable de prédire les défauts bancaires des TPE, en utilisant des données financières comme entrée. et l'utiliser dans une application web. avec laquelle un utilisateur peut :

Utilisateur	
s'authentifier	Un utilisateur peut s'authentifier en entrant son mot de passe et nom d'utilisateur.
prédir	Un utilisateur peut prédir la probabilité de défaut bancaire d'une TPE en entrant ses données financieres.
consulter historique	Un utilisateur peut consulter ses propres prédictions précédentes
administrateur	
s'authentifier	Un administrateur peut s'authentifier en entrant son mot de passe et nom d'utilisateur.
prédir	Un administrateur peut prédir la probabilité de défaut bancaire d'une TPE en entrant ses données financieres.
consulter historique	Un administrateur peut consulter toutes les prédictions précédentes
gérer comptes utilisateur	Un administrateur peut consulter, ajouter ou supprimer un utilisateur

Tableau 1.1 : besoins fonctionnels

la figure 1.3 explique ces besoin avec un diagramme de cas d'utilisation globale

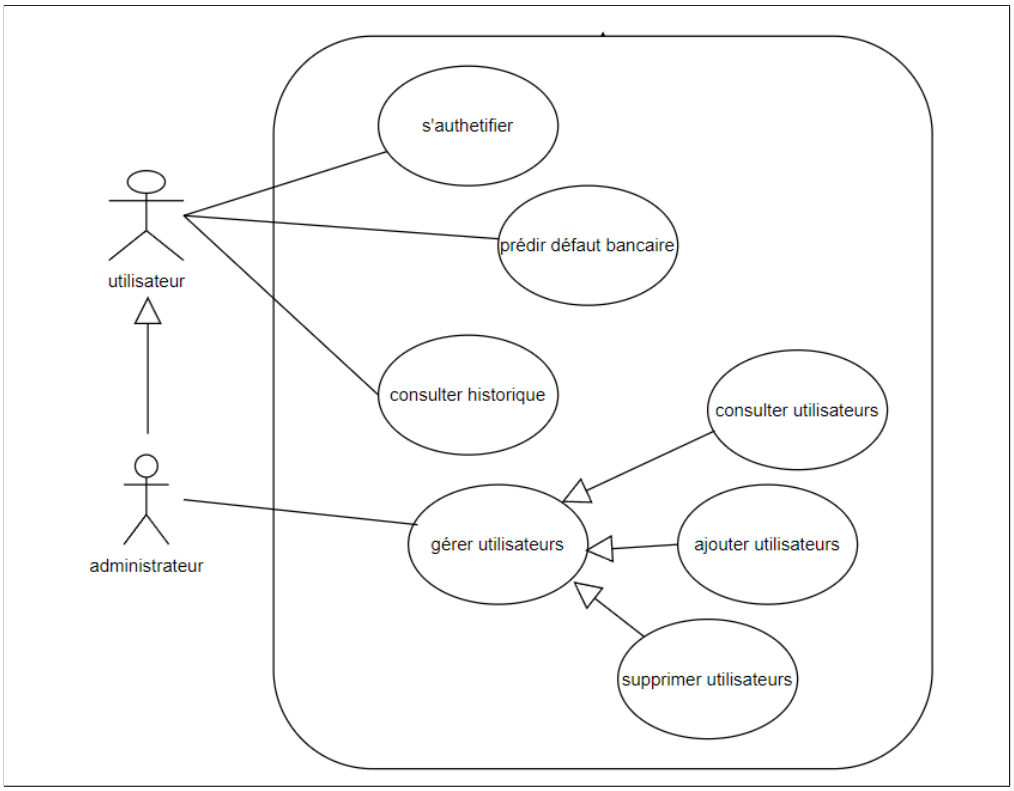

Figure 1.3 : diagramme de cas d'utilisation global

1.5.2 Besoins non fonctionnel

Notre application de prediction doit respecter des besoins non fonctionnel :

besoins non fonctionnel	
Sécurité	Un utilisateur peut s'authentifier en entrant son mot de passe et nom d'utilisateur.
fiabilité	le modéle de prediction doit étre performant et etourne une probabilité de défaut bancaire fiable

Tableau 1.2 : besoins non fonctionnels

1.6 Explication financière

1.6.1 Défaut bancaire

En général, le défaut bancaire peut être défini comme le non-paiement ou le défaut de paiement d'un prêt bancaire ou le non-respect des conditions de l'accord de prêt. Les faillites bancaires peuvent prendre de nombreuses formes, y compris des retards de paiement, des paiements

partiels ou incomplets, et même une incapacité totale à rembourser les prêts. Ces défauts peuvent avoir de graves conséquences financières pour l'emprunteur, notamment des frais supplémentaires, une détérioration du crédit, des recours juridiques et, plus grave encore, la confiscation ou la saisie de la garantie. dans le cadre de ce projet on définit le défaut bancaire comme retard de paiement de 90 jours au plus

1.6.2 Très petite entreprise

Une très petite entreprise (TPE) [2] est généralement une entreprise qui emploie moins de 20 salariés et dont le chiffre d'affaires annuel ne dépasse pas les 2 millions d'euros. Il s'agit généralement d'une entreprise individuelle, également appelée micro-entreprise. Ces TPE regroupent les indépendants tels que les travailleurs libéraux, les commerçants et les artisans. Cette structure se caractérise par un budget de départ qui ne nécessite pas de fonds d'amorçage ni de main-d'œuvre importants.

1.6.3 Les variables financières

Voici les définitions des termes financiers que nous allons utiliser pour notre application [3] :

— chiffre d'affaires : c'est la somme des ventes effectuées par celle-ci hors taxes passif courant : dette à court terme généralement moins de 12 mois actif courant :Ils comprennent le stock , l'ensemble des biens appartenant à l'entreprise.

— capitaux propres :Ils comprennent les fonds apportés ou mis à disposition de l'entreprise de façon durable.

— Résultat net : revenus des associés.

— liquidité : capacité à rembourser les dettes à court terme a partir des actifs courants

— fonds de roulement :différence entre ressources stables et emplois stables.

— besoin de fonds de roulement :la différence entre les besoins entre le cycle de financement des stocks et des créances et dettes à court terme.

— bilan total : le montant total des éléments composant l'actif ou le passif

— dette structuré : dette à long terme généralement plus de 12 mois

— fonds propres net : c'est le capital apporté par les actionnaires, ainsi des bénéfices laissées à la disposition de l'entreprise

— cash-flow net : mesure le flux de trésorerie d'une entrepris

1.6.4 Les ratios financièrs

nom du ratio	formule	interpretation
Solvabilité	CAP_PROP/TOT_BIL	La capacité de l'entreprise à rembourser ses dettes à long terme
Équilibre	FR/BFR	La capacité d'équilibrer ses revenus et ses dépenses
Liquidité immédiate	LIQ/PASS	La capacité d'une entreprise à rembourser ses dettes à court terme
Rotation des capitaux propres	CA/CAP_PROP	La capacité de l'entreprise à générer des bénéfices à partir des capitaux propres investis
Return on equity	RN/CAP_PROP	Le rendement généré par l'entreprise par rapport aux capitaux propres investis
Rotation des actifs	CA/TOT_BIL	La capacité de l'entreprise à générer des revenus à partir de ses actifs
Marge net	RN/CA	La rentabilité de l'entreprise en termes de bénéfice net par dinars de ventes
Liquidité courante	ACT/PASS	La capacité de l'entreprise à rembourser ses dettes à court terme à l'aide de ses actifs

Auto-financement	CAP_PROP/(CAP_PROP+DS)	La capacité de l'entreprise à générer suffisamment de liquidités en interne pour financer ses investissements
Couverture	DS/CFN	La capacité de l'entreprise à rembourser ses dettes à long terme
Leverage financier	PASS/CAP_PROP	La proportion de la dette dans la structure financière de l'entreprise
Endettement	DS/FPN	La portion d'endettement de l'entreprise

Tableau 1.3 : ratios financiers

Conclusion

Dans ce chapitre on a déterminé le cadre du projet tel que l'organisme d'acceuil, une etude de l'existant et le travail demandé .On a aussi défini les besoins fonctionnels et non fonctionnels du projet et défini les termes du champ lexical financier. Dans le chapitre suivant nous allons expliquer les termes relative au domaines de finance nécessaire pour ce projet

Chapitre 2

Exploration et prétraitement des données

Plan

Introduction

Les données à utiliser ont été fournies par l'établissement d'accueil et une exploration de ces derniers est importante pour passer aux étapes suivantes . Dans ce chapitre on va explorer et avoir une idée générale sur la nature des données, calculer les ratios financiers et prétraiter les données avant de les utiliser dans les modéles de machine learning.

2.1 Exploration generale

les données fournis sont sous la forme d'un fichier csv qui contient 8378 lignes (clients) avec 14 colonne :

- NUM_CLT
- CAP_PROP
- TOT_BIL
- LIQ
- PASS_COURANT
- FR
- BFR
- CA
- RN
- ACT_COURANT
- DS
- CFN
- FPN
- DEFAUT

Toutes les variables sont des variables sont continues à part la variable cible qui est binaire : 1 (défaut bancaire) ou 0 (pas de défaut bancaire) et la variable 'NUM_CLT' de type chaîne de caractères qui représente ID d'une entreprise. La figure 2.1 montre un aperçue de l'échantillon obtenue0

	NUM_CLT	CAP_PROP	TOT_BL	LIQ	PASS_COURANT	FR	BFR	CA	RN	ACT_COURANT	DS	CFN	FPN	DEFAUT
0	CLT1400	-570.818329	263.420337	5.620334	657.990689	-479.870894	-373.014838	824.667248	5.640226	178.119795	176.247977	24.943667	-570.818329	0
3	CLT364	109.000000	136.000000	25.000000	27.000000	102.000000	77.000000	369.000000	88.000000	129.000000	0.000000	90.000000	109.000000	0
4	CLT3687	61.390000	327.724000	48.131000	172.195000	58.287000	15.043000	257.303000	12.199000	230.482000	94.139000	36.296000	61.390000	0
5	CLT3446	154.311000	1441.542000	12.808000	787.101000	-549.719000	-116.983000	947.202000	42.823000	237.382000	500.130000	696.797000	154.311000	0
6	CLT1410	194.183161	383.119959	4.547186	188.936798	178.183037	173.635851	169.189850	-272.554747	367.119835	0.000000	-256.309231	194.183161	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8373	CLT2006	452.168000	1032.705000	50.655000	474.057000	526.406000	545.798000	954.838000	52.086000	1000.463000	106.480000	66.909000	452.168000	0
8374	CLT2544	-159.746445	30.623658	1.709029	35.673377	-13.865799	-15.574828	122.924498	30.377149	21.807578	154.696726	50.689857	-159.746445	0
8375	CLT2456	266.315000	648.941000	25.455000	306.088000	136.346000	188.161000	325.175000	28.712000	442.434000	76.538000	84.557000	266.315000	0
8376	CLT3636	140.000000	328.000000	0.000000	188.000000	140.000000	170.000000	543.000000	99.000000	328.000000	0.000000	99.000000	140.000000	1
8377	CLT1326	82.000000	224.000000	10.000000	73.000000	-2.000000	-12.000000	452.000000	72.000000	71.000000	69.000000	81.000000	82.000000	0

Figure 2.1 : aperçus d'échantillon fournis

On remarque dans la figure : 2.2 un déséquilibre énorme dans la distribution de la variable cible (défaut) avec 4.8% de l'échantillon est de classe 1 (client a commis un défaut bancaire)

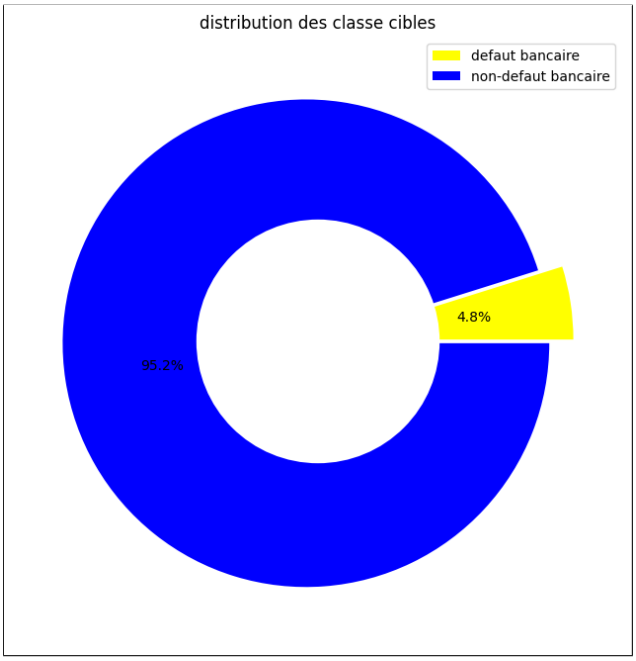

Figure 2.2 : distribution des classes cibles

2.2 Valeurs manquantes

l'existence des valeurs manquantes peut être due a plusieur raison, comme les mal saisi lors de l'échantillonnage ou leur non-existence au moment de la collecte de données. Afin d'avoir un meilleur modèle de prédiction il est très important de traiter ces valeurs avec une des démarches possibles.

— D'après la figure : 2.3 on peut dire que le nombre de lignes avec des valeurs manquantes est minuscule (3.1% des lignes) donc la méthode choisie était de supprimer ces lignes.

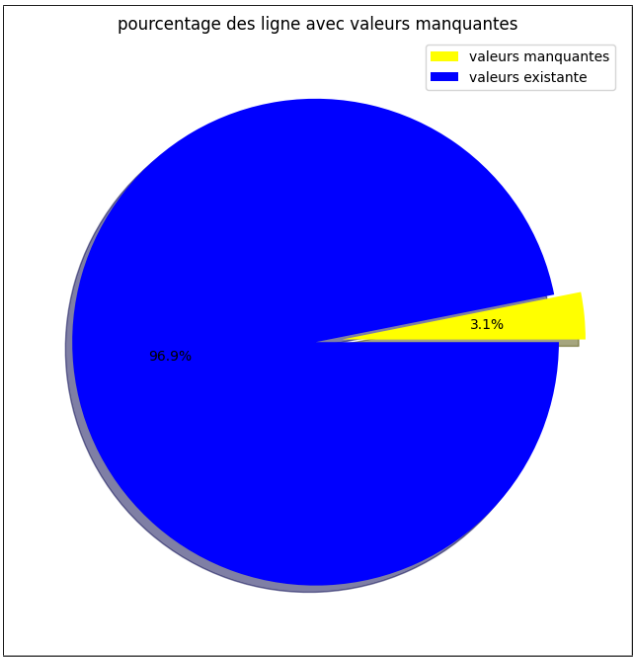

FIGURE 2.3 : pourcentages des valeurs manquantes

2.3 Valeurs aberrantes

[4] une valeur aberrante est une observation dont la valeur est beaucoup éloignée de la distribution de cette variable.l'un des moyen les plus répandu pour détecter des valeurs aberrantes est la visualisation des données avec 'boxplot' comme illustré par la figure 2.4.

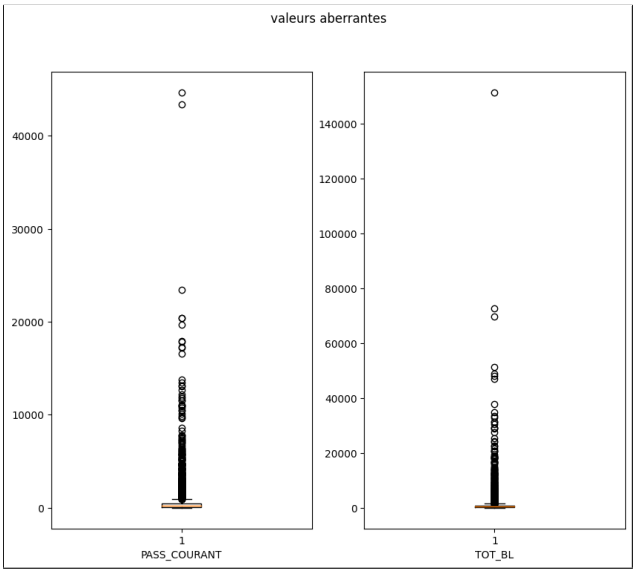

FIGURE 2.4 : visualisation de valeurs aberrantes

Après une visualisation de tout l'échantillon et detection des valeurs aberrantes, le pourcentage des valeurs aberrantes n'a pas dépassé les 0.5%, Comme on peut le voir dans la figure 2.5. Donc la

demarche choisie était de les supprimer pour ne pas affecter les modèles d'apprentissage.

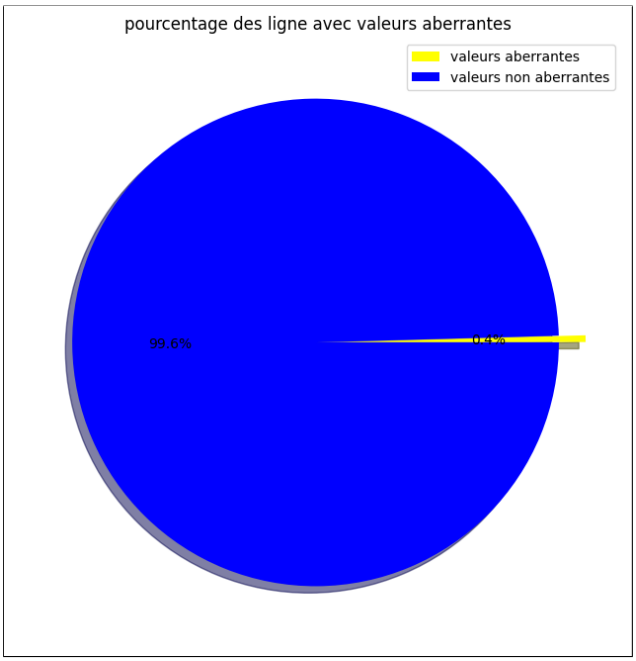

Figure 2.5 : pourcentages des valeurs aberrantes

2.4 Equilibrer les classe cibles avec RENN

2.4.1 L'algorithme ENN (Edited Nearest Neighbor)

[5] L'algorithme ENN développé par Dennis Wilson en 1972 consiste à trouver d'abord les k-voisins les plus proches de chaque observation, puis à vérifier si la classe majoritaire parmi les k-voisins est la même que celle de l'observation. Si la classe majoritaire des k-voisins et la classe de l'observation sont différentes, alors l'observation et ses k-voisins sont supprimés du jeu de données. L'algorithme ENN peut être décrit comme suit : En partant d'un jeu de données contenant N observations, déterminer K comme le nombre de voisins les plus proches. Si K n'est pas défini, il doit être déterminé. Trouver les k-voisins les plus proches de chaque observation parmi les autres observations du jeu de données, puis renvoyer la classe majoritaire parmi les k-voisins. Si la classe de l'observation et la classe majoritaire parmi les k-voisins sont différentes, alors l'observation et ses k-voisins sont supprimés du jeu de données. Répéter les étapes 2 et 3 jusqu'à ce que la proportion souhaitée de chaque classe soit atteinte .L'effet de ENN peut être visualisé dans la figure 2.6.

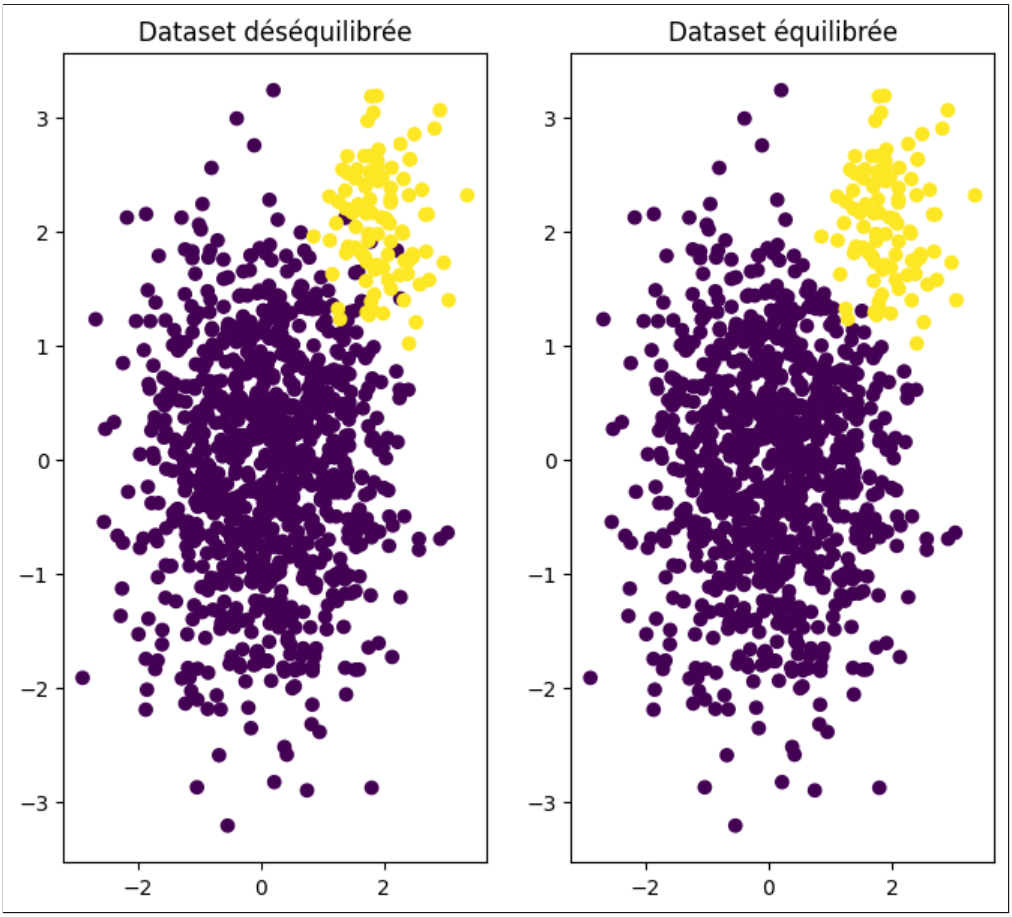

Figure 2.6 : exemple de l'algorithme ENN

2.4.2 L'algorithme RENN (repeated Edited Nearest Neighbor)

[6]L'algorithme Repeated Edited Nearest Neighbor est une technique de sous-échantillonnage qui utilise l'algorithme ENN plusieurs fois successive jusqu'à ce qu'il ne peut plus supprimer d'autres observations. après utilisation de cet algorithme le pourcentage de défaut bancaire a passé de 4.8% à 31.3% et de 7730 observation à 1146 observations pertinentes comme il est présenté dans la figure 2.7.

(Les observations supprimées seront utilisées ultérieurement dans la validation des modèles.)

Après ces étapes de nettoyage de données on procède à calculer les ratios financiers pour continuer le pretraitement des donnees

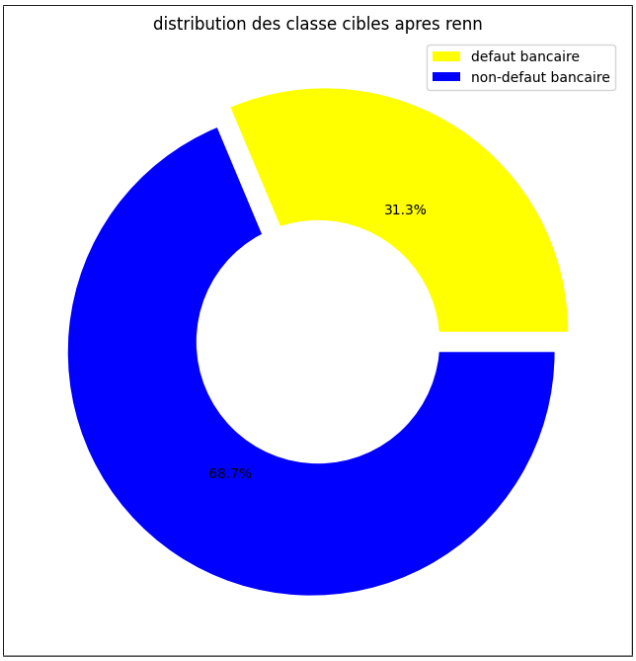

FIGURE 2.7 : resultat apres RENN

2.5 traitement des correlation

2.5.1 Test de pearson

definition [7]Le test de Pearson est un test de corrélation qui trouve une relation linéaire entre deux variables continues. ce coefficient varie entre 1 et -1 avec :

- r=1 : corrélation positive, si l'une des variables augmente l'autre augment et l'inverse est vrai
- r=-1 : corrélation négative , si l'une des variables augmente l'autre diminue et l'inverse est vrai.
- r=0 : pas de liaisons entre les deux variables.

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

- X_i : i-eme valeur des variables X
- $\bar{x}$: moyenne de X
- de même pour Y

pré-test pour avoir un résultat réel du test de Pearson il est impératif de traiter les valeurs extrêmes dans l'échantillon. on a opté alors pour la méthode la plus simple. Pour chaque variable

on a :

— diviser l'échantillon en 3 partie valeurs minimales (5%) , valeurs moyenne (90%) et valeurs maximales (5%).

— remplacer les valeurs minimales par la plus petite valeur des valeurs moyennes.

— remplacer les valeurs maximales par la plus grande valeur des valeurs moyennes.

— comme expliqué par la figure 2.8.

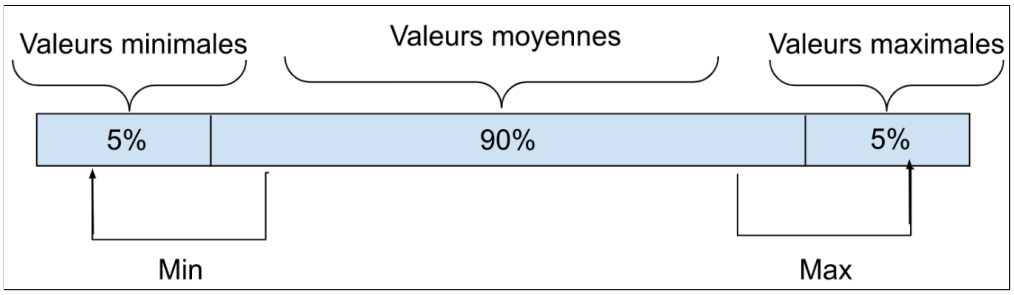

FIGURE 2.8 : traitement des valeurs extrêmes

realisation du test Nous allons considérer qu'il ya une relation linéaire entre X et Y si |r(X,Y)|>0,7.En examinant la figure 2.9 nous avons trouvé les résultat suivant :

— r('liq_cour' , 'liq_immediate')=0,9

— r('leverage_financier' , 'rot_cap_prop')=0,86

— r('endettement' , 'aut_fin')=-0,73

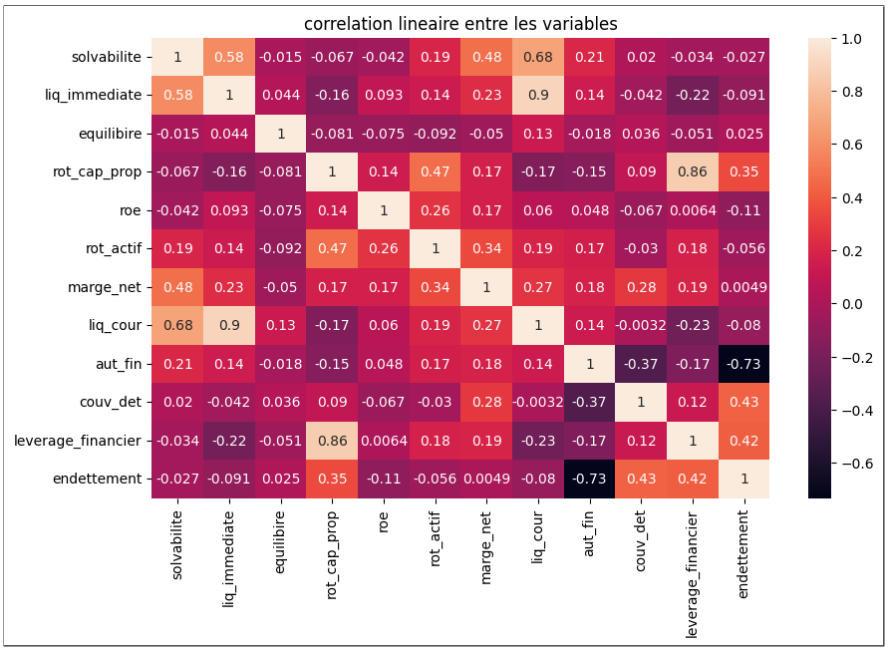

FIGURE 2.9 : visualisation du test de Pearson

2.5.2 Signification des correlations avec p-value

[7] La valeur p mesure la probabilité que la corrélation observée entre deux variables soit un hasard. Si la valeur p est inférieure au niveau de signification, on peut considérer que la corrélation est significative et que l'hypothèse nulle peut être rejetée et que la corrélation est significative. Le niveau de signification le plus commun est $\alpha = 0.05$

- p-value de la correlation entre ('liq_cour','liq_immediate') : 0.00 donc corrélation est significative
- p-value de la correlation entre('leverage_financier','rot_cap_prop') : 0.00 donc corrélation significative
- p-value de la correlation entre ('endettement','aut_fin') : 0.79 donc corrélation non significative

2.5.3 Sélection de caractéristiques discriminantes avec AUC.

[8]La courbe ROC (Receiver Operating Characteristic) est obtenue en tracant le taux de vrais positifs (TPR) par rapport au taux de faux positifs (FPR) pour différentes valeurs seuils de classification. Le TPR représente la probabilité qu'un échantillon positif soit correctement prédit comme positif, tandis que le FPR représente la probabilité qu'un échantillon négatif soit incorrectement prédit comme positif. En d'autres termes, la courbe ROC permet d'évaluer la performance d'un modèle de classification en mesurant sa capacité à discriminer les échantillons positifs des échantillons négatifs.

[8]L'AUC est une mesure numérique calculée à partir de cette courbe ROC qui indique de la performance globale de ce modèle en calculant l'air sous la courbe. .C'est une métrique utilisée pour validation de modèle mais elle est souvent utilisée pour évaluer l'importance des caractéristiques.

D'après notre analyse, on constate de la figure 2.10 :

—

- 'liq_cour' continent plus d'information sur la variable cible que 'liq_immediate'
- 'leverage_financier' continent plus d'information sur la variable cible que 'rot_cap_prop'

On supprime alors les ratios avec le moins de quantite d'information

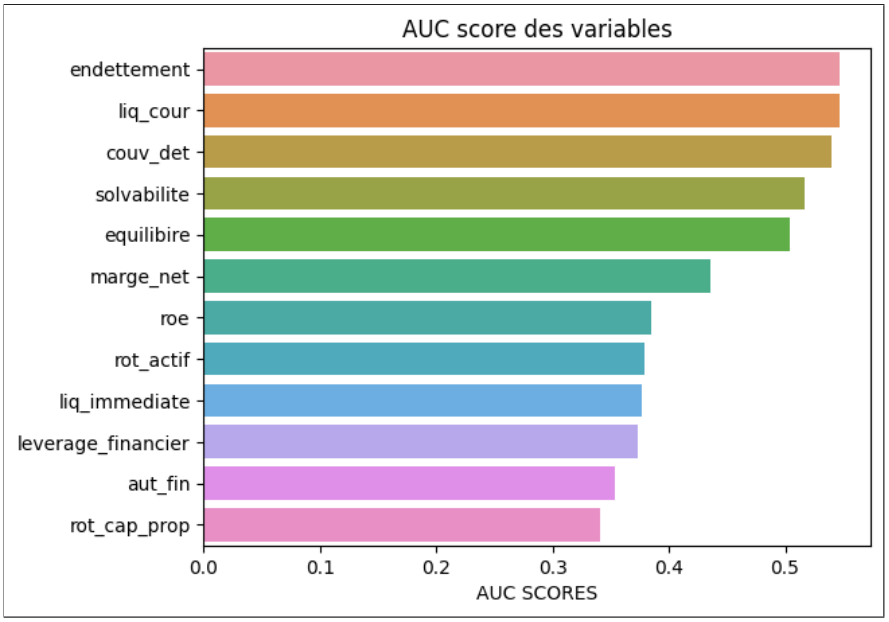

FIGURE 2.10 : quantite d'information des caracteristiques

2.6 Standardisation des variables

[9] Le standardisation des caractéristiques est une technique de prétraitement des données qui consiste à transformer les valeurs des variables d'un ensemble de données à des valeurs sur une échelle similaire. Cela est fait pour garantir que toutes les caractéristiques contribuent également au modèle et pour éviter que les fonctionnalités avec des valeurs plus importantes ne dominent le modèle.

Pour chaque valeur de chaque caractéristique, la standardisation suit la formule suivante :

$$z = \frac{x - \mu}{\sigma}$$

- x est la valeur originale
- μ est la moyenne de la variable
- σ est l'écart type de la variable : $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$ ou N nombre d'observation et X la variable en question.
- z est la valeur transformée .

dans la figure 2.11 on peut avoir un aperçue sur l'effet de la standardisation pour la variable "equilibre"

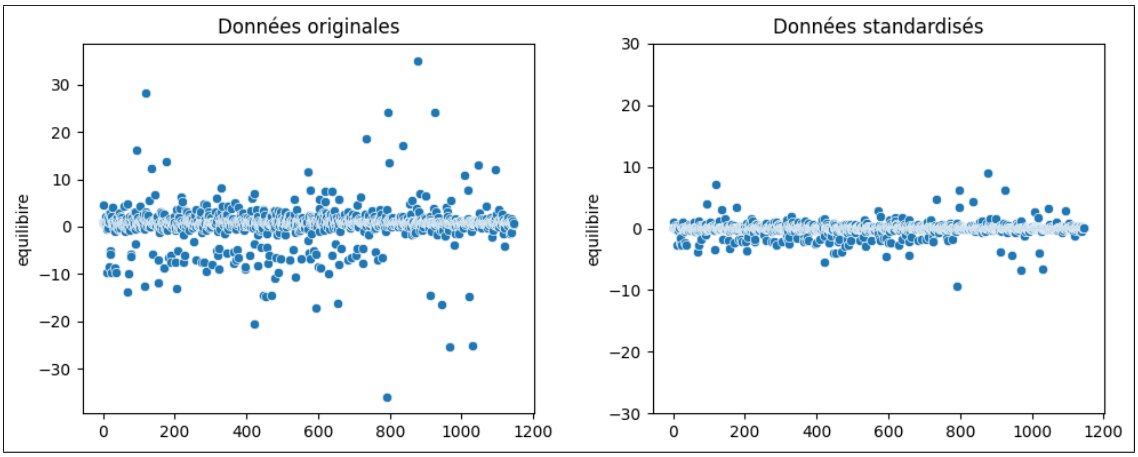

Figure 2.11 : effet de standardisation

Conclusion

Dans ce chapitre nous avons exploré et nettoyé les données pour les préparer à être assimiler par les modèles d'apprentissage. le chapitre suivant sera consacré à expliquer les modèle de machine learning testé pour ce projet, determiner les métriques utilisé et tester les modéles après apprentissage.

Chapitre 3

MODÈLES UTILISÉS ET TESTS DE VALIDATION

Plan

Introduction

Le choix du modèle pour prédire les défaut bancaire ,est crucial pour avoir une résultat de prédiction fiable .Dans ce chapitre nous allons expliquer l'algorithme de chaque modèle utilisé ,les métriques utilisés et les resultats des tests réalisés.

3.1 Regression logistique

[10] La régression logistique est un modèle statistique utilisé pour évaluer et caractériser la relation entre une variable cible, principalement binaire, notée 'y', et une ou plusieurs variables explicatives (qu'elles soient discrètes ou continues). Ce modèle permet également de prédire la probabilité qu'un événement se produise (y=1) ou non (y=0) en optimisant les coefficients de régression. Les résultats de la régression logistique sont toujours compris entre 0 et 1. Lorsque la valeur prédite dépasse un seuil donné, l'événement est considéré comme susceptible de se produire et s'il est inférieur à ce seuil, l'événement ne l'est pas.
Soit la fonction logit :

$$L = \beta X = \log\left(\frac{p}{1-p}\right) = \beta X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{3.1}$$

Avec β qui représente le vecteur des coefficients, $\beta = (\beta_0, \beta_1, \ldots, \beta_n)$ et X est le vecteur des variables indépendantes(X est une observation) $X = (X_0, X_1, \ldots, X_n)$.
Les probabilités que y = 0 est exprimées comme :

$$P(Y = 0|X) = \frac{e^{-\beta X}}{1 + e^{-\beta X}}$$

De même pour la probabilité que y=1 :

$$P(Y = 1|X) = \frac{1}{1 + e^{-\beta X}}$$

Les coefficients $\beta_0, \ldots, \beta_n$ sont estimés par la méthode du maximum de vraisemblance comme suit.
soit : X matrice des variables explicatives continues
y vecteur de la variable cible
α Taux d'apprentissage (paramètre d'apprentissage)

— 0 Initialiser les coefficients β avec des valeurs aléatoires ou zéro et la valeur de la perte (log-vraisemblance) à une valeur élevée initiale.

— 1 Pour chaque échantillon i dans les données d'entraînement X :

  — -Calculer z_i pour l'échantillon i : $z_i = \beta 0 + \beta 1 * X1_i + \beta 2 * X2_i + ... + \beta n * Xn_i$ et la probabilité prédite $p_i = 1/(1 + exp(-z_i))$

  — –Calculer l'erreur entre la probabilité prédite et la valeur réelle de la variable cible : $erreur_i = p_i - y_i$

  — -Mettre à jour les coefficients en utilisant la règle de mise à jour des coefficients de la régression logistique :

  $\beta_j = \beta_j - \alpha \cdot \left(\frac{1}{m}\right) \cdot \sum_{i=1}^{n} (\text{erreur}_i \cdot X_{ji})$

  — -Mettre à jour la perte en utilisant la fonction de perte (log-vraisemblance) :

  $\text{perte} = -\left(\frac{1}{m}\right) \sum_{i=1}^{m} (y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i))$

— 2 repeter 1 jusqu'on arrive à une différence entre les valeurs de perte actuelle et précédente inférieure à un seuil prédéfini ou nombre d'iteration maximal atteind.

— 3 Les coefficients $\beta 0$, $\beta 1$, ..., βn obtenus après convergence.

Pourquoi la regression logistique : La régression logistique est choisie car elle est l'approche statistique la plus couramment utilisée dans la notation de crédit.

3.2 Gradient Boosting

[11] Le “Gradient Boosting" est une méthode qui ajoute itérativement des modèles à un ensemble par cycles successifs. Le fonctionnement du Gradient Boosting peut être divisé en quatre étapes principales.

— 1-Initialisation de l'ensemble : la première étape consiste à initialiser l'ensemble avec un modèle d'arbre de décision simple, également appelé arbre de décision faible. Ce modèle d'arbre de décision est entraîné sur les données d'entraînement pour prédire les valeurs de sortie.

— 2-Calcul de la fonction de perte : après l'entraînement initial, l'algorithme calcule la fonction de perte pour le modèle d'arbre de décision. La fonction de perte est une mesure de la différence entre les prédictions du modèle et les valeurs réelles de sortie. L'objectif est de minimiser cette fonction de perte.

— 3-Entraînement de nouveaux modèles : après avoir calculé la fonction de perte, l'algorithme entraîne un nouveau modèle d'arbre de décision pour réduire la fonction de perte. Cela se fait en utilisant la descente de gradient pour ajuster les poids du modèle. Le nouvel arbre de décision est ajouté à l'ensemble.

— 4-Répétition : les étapes 2 et 3 sont répétées jusqu'à ce que la fonction de perte soit minimisée ou que le nombre d'arbres de décision soit atteint.

— ces etapes sont expliqués dans la figure 3.1.

La perte est calculer par : $L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$

où :

y_i est la vraie étiquette binaire de l'observation i (0 ou 1) $\hat{y}_i$ est la prédiction de probabilité de l'observation i d'être positive (entre 0 et 1) n est le nombre total d'observations

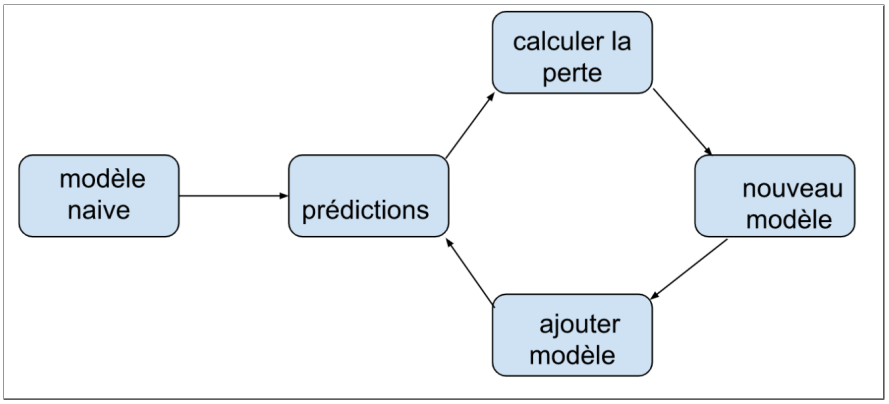

Figure 3.1 : fonctionnement du gardient boost

Pourquoi le gradient boost Gradient Boosting a était choisi particulièrement pour son performance lorsque les données présentent des caractéristiques non linéaires ou des interactions entre les variables explicatives, ce qui peut être le cas dans notre projet de prédiction de défaut bancaire.

3.3 forêt aléatoire

3.3.1 Arbre de decision

[12] Un arbre de décision est un modèle de prédiction qui utilise une structure en forme d'arbre pour représenter une série de décisions qui conduisent à une prédiction. Le processus de construction d'un arbre de décision se résume en ces 4 étapes :

— 1. Calculez le gain d'information de chaque variable avec l'indice de Gini

— 2. Sélectionnez la variable avec le mailleur gain pour diviser les données.

— 3. Sélectionnez la variable avec l'indice de Gini le plus faible pour diviser les données.

— 4. Répétez les étapes 1 à 3 pour chaque branche de l'arbre jusqu'à ce que toutes les feuilles atteignent un indice de gini nulle ou condition d'arrêt

Indice de gini C'est une mesure de la pureté d'un ensemble de données. le calcule se fait comme suit :

— 1. calculer Gini avant séparation : $Gini = 1 - ((p_0)^2 + (p_1)^2)$ avec p_0 probabilité qu'un echantillon soit de classe 0 et avec p_1 probabilité qu'un echantillon soit de classe 1

— 2. pour chaque variable X :

— a. Calculer $Gini(\text{X}) = \sum_{i=1}^{n} p(i) \cdot Gini(i)$
avec n nombre de valeurs que peut prendre X, p(i) probabilité qu'un echantillon a comme valeur la i éme valeur que X peut prendre.

— b. calculer le gain d'information de X
Information Gain(variable) = Gini(avant partition) - Gini(variable)

— 3. Choisir la variable avec l'information gain la plus élevée comme nœud

3.3.2 Forêt aléatoire

[13] La forêt aléatoire est considérée comme une technique avancée d'AD,elle consiste a faire des prédiction a partire de plusieur arbres de décision.

Cet algorithme procède Conformément à la figure 3.2 et comme suit :

— diviser l'ensemble de données en plusieurs sous-ensembles

— construire pour chaque sous ensemble un arbre de décision

— pour une nouvelle observation chaque arbre donne un résultat indépendamment des autre

— La prédiction du modèle est la classe la plus prédite (moyenne des prédictions pour les problèmes de régression)

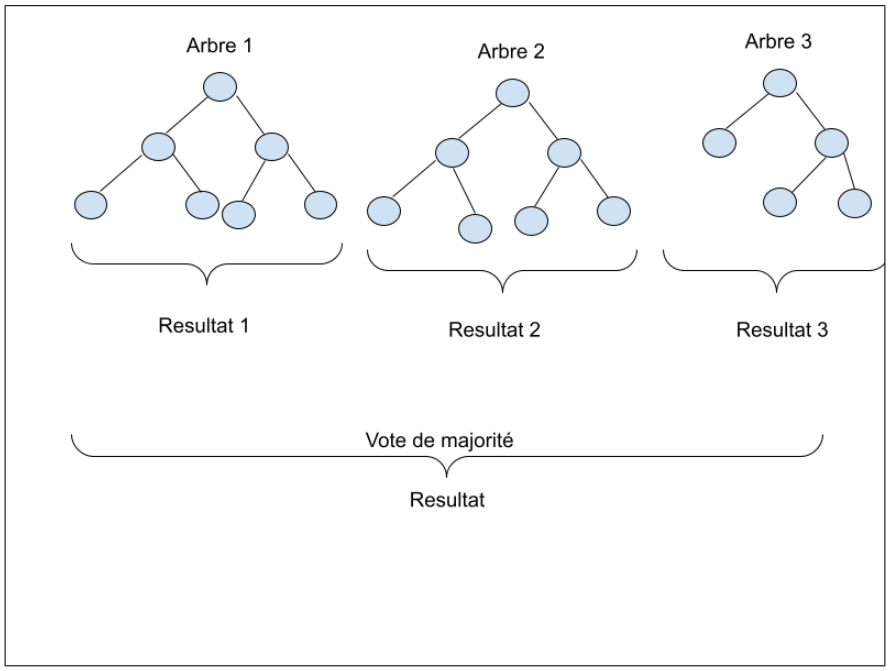

Figure 3.2 : fonctionnement du foret aleatoire

—

Pourquoi la forêt aléatoire La forêt aléatoire a été choisie pour la facilité d'interprétation des arbres de décision et leur capacité à capturer des relations non linéaires entre les variables e a capturer des relations non linéaires entre les variables. la forêt aléatoire permet d'exploiter les avantages des arbre de décision tout en évitant le sur-ajustement du modèle

3.4 machines à vecteurs de support

[14] Le Support à vecteur machine (SVM) est une technique de machine learning utilisée dans les problèmes de classification (ou de régression). Les SVM prennent un ensemble de deux classes d'entrées données et les prédisent afin de déterminer laquelle des deux classes probables à la sortie.Cette approche cherche le meilleur hyperplan pour que les classes à prédire soient sur différents côtés de la droite.L'hyperplan est une ligne droite qui divise l'espace en deux parties et qui est définie par l'équation :

$$w.x + b = 0 \tag{3.2}$$

où w est un vecteur normal à l'hyperplan et b est un scalaire qui permet de décaler l'hyperplan par rapport à l'origine.comme le montre la figure 3.3

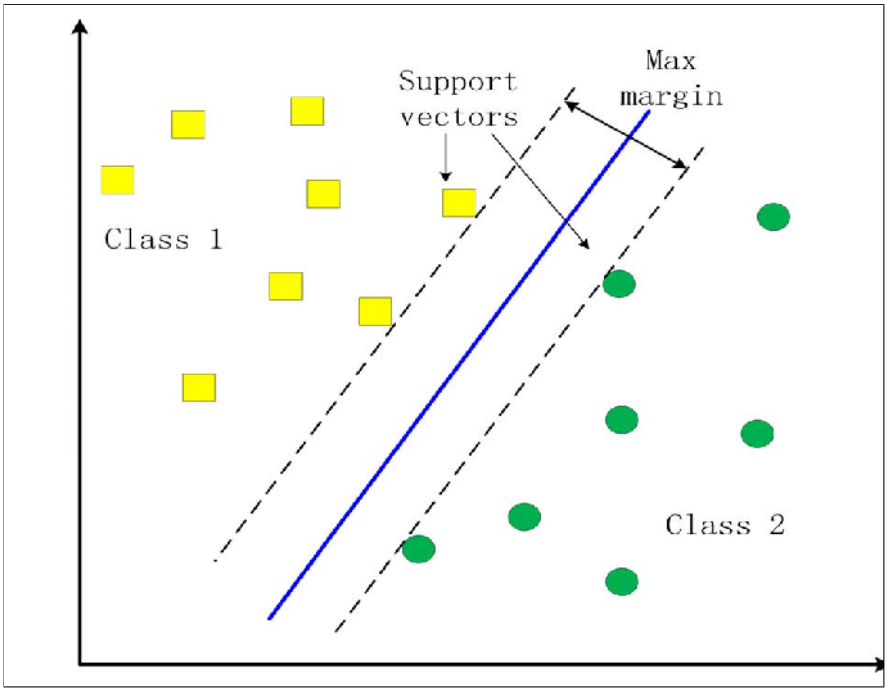

FIGURE 3.3 : fonctionnement du SVM

(Source :https ://www.researchgate.net/publication/261154218_A_gene_signature_for_breast_cancer_progno

)

Si les données ne sont pas linéairement séparables, l'approche SVM peut être améliorée en utilisant des techniques de transformation de données pour les projeter dans un espace de dimension supérieure où elles peuvent être séparées linéairement. il faut trouver une fonction de transformation qui mappe les données dans un espace de dimension supérieure où elles sont linéairement séparables. Cette fonction est appelée "Fonction de noyau". Le choix de la fonction de noyau dépend du type de données à traiter. Cette transformation peut etre expliqué par la figure 3.4 .

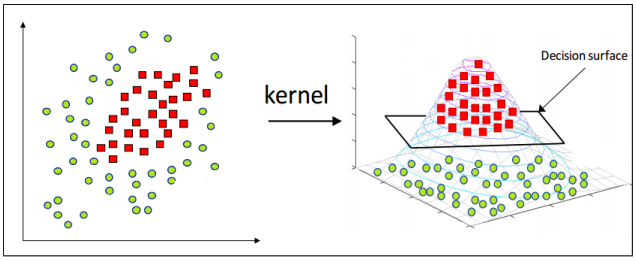

FIGURE 3.4 : transformation des données pour SVM

(Source :https ://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beg

Pourquoi le SVM Le modèle SVM a été considéré en raison de sa capacité à bien fonctionner avec des ensembles de données de petite à moyenne taille comme le notre et sa capacité à séparer

les données en deux classes (défaut ou non-défaut) en utilisant des frontières de décision optimales.

3.5 Reseau de neurones artificiel : perceptron multicouche

Reseau de neurones artificiel Les réseaux de neurones artificiels sont des systèmes de traitement de l'information dont le mécanisme est inspiré de la fonctionnalité des circuits neuronaux biologiques. Un réseau de neurones artificiels possède de nombreuses unités de traitement connectées les unes aux autres.

perceptron mmulticouche (Multi Layer Perceptron) [15] Un perceptron multicouche (MLP) est un modèle d'apprentissage supervisé de type de réseau de neurones artificiel capable de classer les données à l'aide d'une série de couches cachées avec des fonctions d'activation non linéaires. l'algorithme MLP s'execute comme suit :

- 1. initialiser des poids aléatoirement pour chaque entrée
- 2. calculer la somme pondérée de chaque neurone et le résultat est passe par une fonction d'activation 'f'

  $z_j = \sum_{i=1}^{n} w_i x_i + b_i$

  $a_j = f(z_j)$

  avec j la j-iéme neurone ,n nombre d'enntrées , w_i le i-iéme poids , x_i i-iéme entrée ei b_i i-iéme biais.
- 3. calculer l'erreur des prédictions par rapports au valeurs réelles
- 4. mettre à jours les pois avec rétropropagation du gradient
- 5. répéter de 2 à 4 jusqu'à atteindre une condition d'arrêt

Pourquoi le réseau de neurones artificiel Nous avons étudié les réseaux de neurones en raison de leur aptitude à comprendre les liens complexes et non linéaires entre les variables. . Les réseaux de neurones sont particulièrement utiles lorsque des caractéristiques cachées essentielles pour prédire les défauts bancaires sont présentes.

3.6 Tests des modéles et métriques utilisés

3.6.1 Matrice de confusion

[16] La matrice de confusion est une matrice carrée qui permet de visualiser les performances d'un modèle de classification en présentant le nombre de prédictions correctes et incorrectes effectuées par ce dernier pour chaque classe du problème. dans notre cas ou c'est une classification binaire , la matrice de confusion se présume dans le tableau suivant :

	Prédiction positive	**Prédiction négative**
Classe positive	TP	FN
Classe négative	FP	TN

Tableau 3.1 : Matrice de confusion

TP : vrai positive le nombre d'observations prédites vrai et sont réellement vrai

FN : faux négatif le nombre d'observations prédite faux est sont réellement vrai

FP : faux positive le nombre d'observation prédite vrai et sont réellement faux

TN : vrai négative le nombre d'observation prédite faux et sont réellement faux.

3.6.2 Rappel

[17]Le rappel mesure la proportion de vrais positifs parmi tous les exemples positifs dans l'ensemble de données. Plus le rappel est élevé, plus le modèle est capable d'identifier les exemples positifs.

$$Rappel = \frac{TP}{TP + FN} \tag{3.3}$$

3.6.3 Précision

[17] La précision mesure la proportion de vrais positifs parmi tous les exemples identifiés par le modèle comme positifs. Plus la précision est élevée, plus le modèle est capable d'identifier correctement les exemples positifs.

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

3.6.4 FPR(false positive rate)

[8]Le FPR mesure la proportion de faux positifs parmi tous les exemples negatifs dans l'ensemble de données. Plus le FPR est élevé, moins le modèle est précis dans la prédiction des échantillons négatifs.

$$FPR = \frac{FP}{TN + FP} \tag{3.5}$$

3.6.5 AUC et courbe ROC

[8] La courbe ROC est tracée en utilisant le taux de vrais positifs (qui est le rappel) sur l'axe des ordonnées et le taux de faux positifs (FPR) sur l'axe des abscisses pour différents seuils de classification. l'AUC représente le degré ou la mesure de séparabilité. Elle indique dans quelle mesure le modèle est capable de distinguer les classes.Donc plus la courbe ROC et etirré vers le coin gauche en haut du graphe plus AUC est important donc plus le modèle est performant.

3.6.6 F1-score

[17] Le F1-score est une mesure utile pour les cas où la précision et le rappel ont la même importance, et il permet de résumer la performance du modèle en une seule valeur en combinant le rappel et la précision .

$$f1 = 2\frac{\text{PrecisionRecall}}{\text{Precision} + \text{Recall}} \tag{3.6}$$

3.6.7 echantillonnage entraînement/validation

La validation des modèles est une étape cruciale dans le 'machine learning'. Elle permet d'évaluer la performance du modèle sur de nouvelles observations inconnues non utilisées pour l'entraînement.

A ce point nous avons un échantillon de 1146 observations avec comme variables des ratios calculés et sans corrélation significative entre eux, et ils sont standardisés , on peut voir dans un aperçu des données dans la figure 3.5 .

solvabilite	equilibire	roe	rot_actif	marge_net	liq_cour	aut_fin	couv_det	leverage_financier	endettement	DEFAUT
0.836446	0.515771	-0.086451	0.596413	0.049551	0.579250	-0.088775	0.041393	-0.373688	0.119824	0
0.579137	0.076750	0.372574	1.402240	0.061498	-0.109007	0.159080	-0.060642	-0.248729	-0.124976	1
0.564727	-3.863223	-0.049516	0.919009	0.048874	-0.136660	0.159080	-0.060642	-0.244705	-0.124976	1
-0.225480	-0.079230	0.010035	-0.451254	0.049906	-0.261845	-0.161897	-0.007380	0.472116	0.237347	0
-1.188715	0.109073	0.281298	-0.395880	0.010184	-0.327646	0.380618	-0.100539	-1.022634	-0.246359	0

Figure 3.5 : aperçu des données avant apprentissage

Puis, l'échantillon est divisé en ensembles de test et d'apprentissage, respectivement de taille 10% et 90%, avec la même proportion de défauts bancaires dans les deux parties comme expliqué par la figure 3.6.

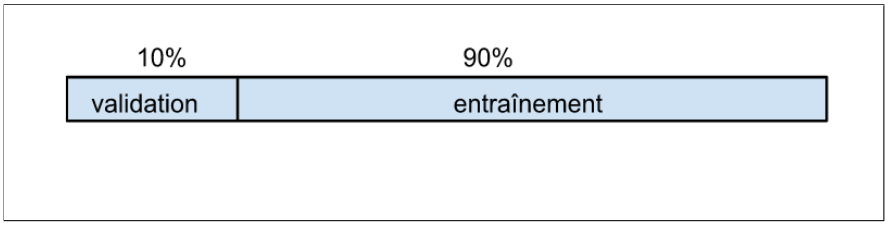

Figure 3.6 : echantillonnage validation/entraînement

3.6.8 Resultats des tests

Après avoir obtenu les résultats de prédiction pour l'échantillon de validation, on peut tracer la courbe ROC pour chaque modèle et comparer leur pouvoir de classification , voir figure 3.7. Il est important de noter que plus la courbe est étirée vers le haut à gauche, plus le modèle correspondant est fiable. Cela signifie qu'il est capable de mieux distinguer les observations positifs (risque de défaut bancaire) des observations négatifs (TPE saine), ce qui indique une meilleure capacité de classification.

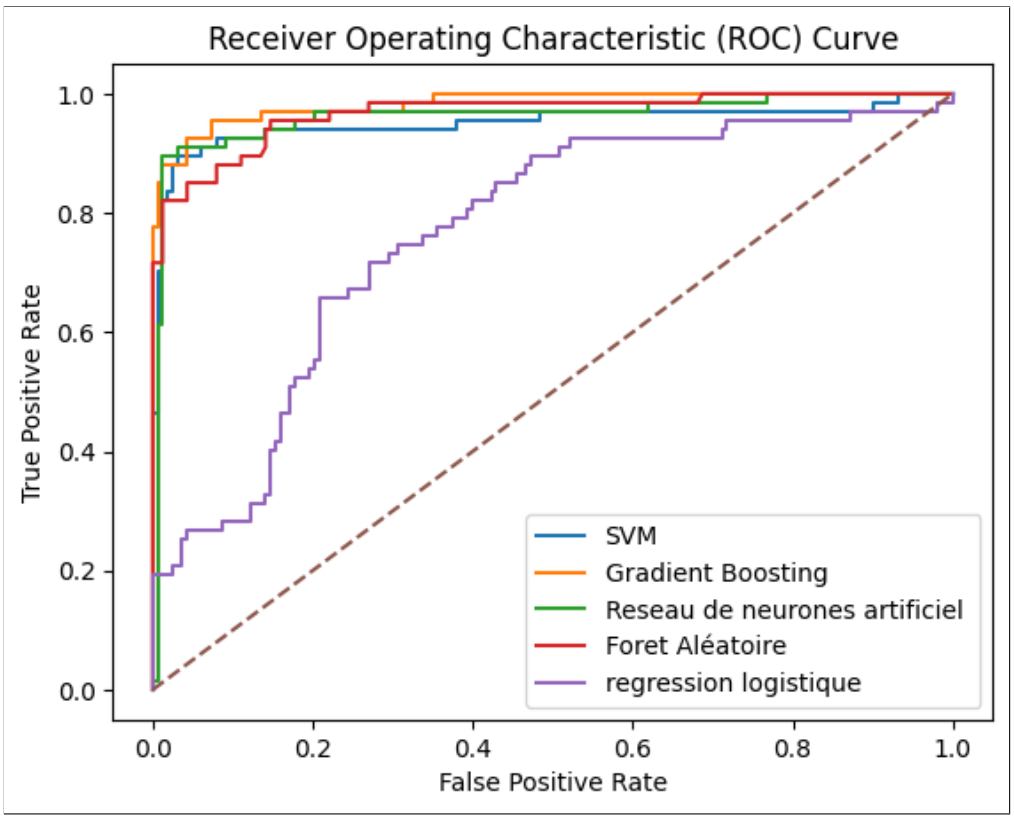

Figure 3.7 : Courbes ROC des modèles

En inspectant les courbes ROC de la figure 3.6, on peut constater que la régression logistique est de loin le modèle le moins performant. En revanche, le réseau de neurones artificiels et le gradient boosting présentent des performances similaires et élevées.

Et avec les même resultats on peut obtient les matrices de confusions de chaque modèle :

	Prédiction positive	**Prédiction négative**
Classe positive	60	7
Classe négative	8	155

Tableau 3.2 : Matrice de confusion : SVM

	Prédiction positive	**Prédiction négative**
Classe positive	59	8
Classe négative	4	159

Tableau 3.3 : Matrice de confusion : gradient boosting

	Prédiction positive	**Prédiction négative**
Classe positive	55	12
Classe négative	6	157

Tableau 3.4 : Matrice de confusion :foret aléatoire

	Prédiction positive	**Prédiction négative**
Classe positive	14	53
Classe négative	4	159

Tableau 3.5 : Matrice de confusion : regression logistique

	Prédiction positive	**Prédiction négative**
Classe positive	61	6
Classe négative	5	158

Tableau 3.6 : Matrice de confusion : reseau de neurones artificiel

Pour résumer les résultats obtenus de chaque modèle, nous avons calculé le rappel, la précision, le score F1 et l'AUC, que nous avons présentés dans le tableau 3.7.

Modèle	Recall score	Precision score	F1 score	AUC score
SVM	0.895522	0.882353	0.888889	0.950737
Gardiant boost	0.880597	0.936508	0.907692	0.983152
Forêt aléatoire	0.820896	0.901639	0.859375	0.967402
Régression logistique	0.208955	0.777778	0.329412	0.766505
RNA	0.910448	0.924242	0.917293	0.962549

Tableau 3.7 : scores des modèles

Pour s'assurer de la stabilité des modèles, pour chaque modèle on procède comme suit :

— 1-On prend un échantillon de base avant RENN (7730 observations).

— 2-On crée un sous-ensemble équilibré au hasard.

— 3-On le divise ensuite en ensembles de test et d'apprentissage.

— 4-On calcule l'AUC (Area Under Curve)

— 5-en répéte de 2 à 4 100 fois.

— 6-On calcule la variance de l'AUC des tests.

— D'après les résultats obtenus et présentés par la figure 3.8, on constate que la variance des modèles ne dépasse pas 0.001. On peut alors conclure que les modèles sont suffisamment stables.

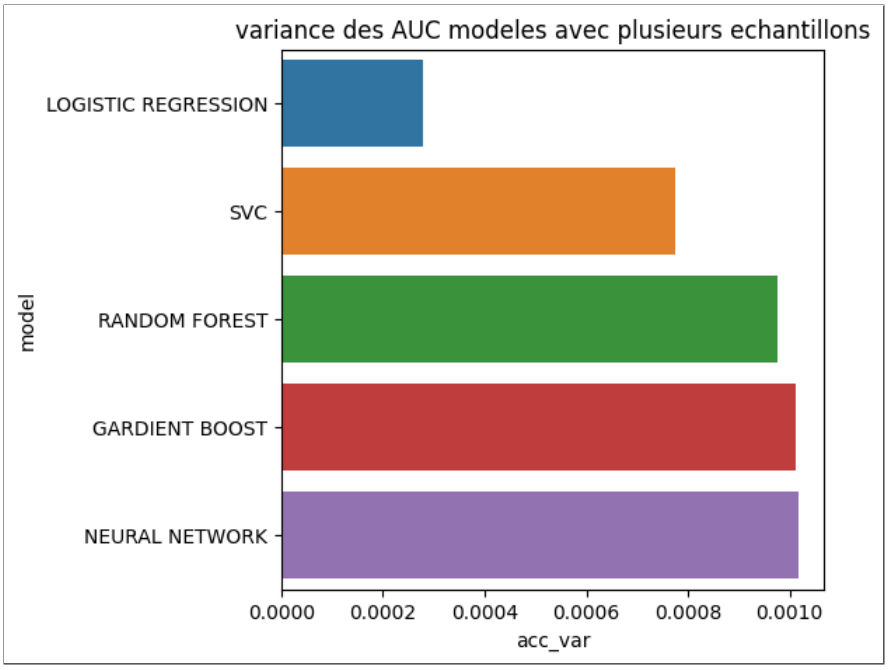

Figure 3.8 : la variance de l'AUC des modèles

Conclusion

Avec les données prétraitées, nous avons entraîné et testé la performance et la stabilité de cinq modèles d'apprentissage : le Gradient Boosting, la Forêt Aléatoire, le SVM, la Régression Logistique et le Perceptron Multicouche. Après avoir analysé les résultats obtenus, nous avons choisi le Gradient Boosting comme modèle final avec une AUC de 0.983152 et un F1-score de 0.907692. Étant donné que l'exploitation du modèle se fait sous forme d'obtention de probabilités, nous avons évité de choisir un seuil par nous-mêmes. Ce choix sera laissé à l'utilisateur (décideur). Dans le prochain chapitre, nous concevrons et réaliserons une application web pour exploiter les prédictions du modèle choisi.

Chapitre 4

DEPLOIMENT DU MODELE CHOISIT

Plan

Introduction

Pour une démonstration de l'utilisation du modèle de machine learning par un utilisateur nous avons opté pour le développement d'une simple application web avec laquelle un utilisateur peut faire des prédictions et consulter les prédictions précédentes.

4.1 choix technologique

4.1.1 scikit-learn

Pour le développement du modèle de machine learning nous avons choisi scikit-learn . c'est un module Python construit sur SciPy qui propose des solutions de traitement , échantillonnage de données ainsi qu'une variété de modèles d'apprentissage automatique qui peuvent être entraînés et déployés et des métriques de validation de ces derniers.

4.1.2 Flask

c'est une léger framework de développement web en python qui offre une certaine flexibilité qui permet aux développeurs de choisir les bibliothèques et les outils qu'ils souhaitent utiliser. Flask permet de séparer le frontend du backend (code Python). Cela rend plus facile la maintenance et la modification du code car on peut se concentrer sur un aspect de l'application à la fois.Ce framework est également extensible grâce à un grand nombre de bibliothèques tierces qui permettent d'ajouter rapidement et facilement de nouvelles fonctionnalités à l'application. c'est un bon choix pour développer une application web de petite taille comme la nôtre

4.1.3 Mongodb

MongoDB est un système de gestion de base de données NoSQL créé en 2009 qui utilise un modèle de données de document. Contrairement aux bases de données SQL, MongoDB stocke les données sous forme de documents JSON. Ce document a une structure flexible et peut être imbriqué les uns dans les autres. MongoDB est connu pour ses performances élevées en lecture et en écriture, ainsi que pour sa capacité à gérer de grandes quantités de données.L'un des atouts de Mongodb est sa communauté de développeurs qui dispose d'une documentation et d'un support en ligne important.

4.2 Concéption

4.2.1 Diagramme de séquence : Authentification

L'authentification est un processus crucial pour assurer la sécurité d'une application web. Un utilisateur doit s'identifier avant de pouvoir faire des prédictions de défaut bancaire ou consulter l'historique des prédictions précédentes.Comme expliqué dans la figure 4.1, l'utilisateur entre son nom d'utilisateur et son mot de passe. Si les données qu'il a entrées sont correctes, une session sera ouverte et l'utilisateur sera dirigé vers la page d'accueil.

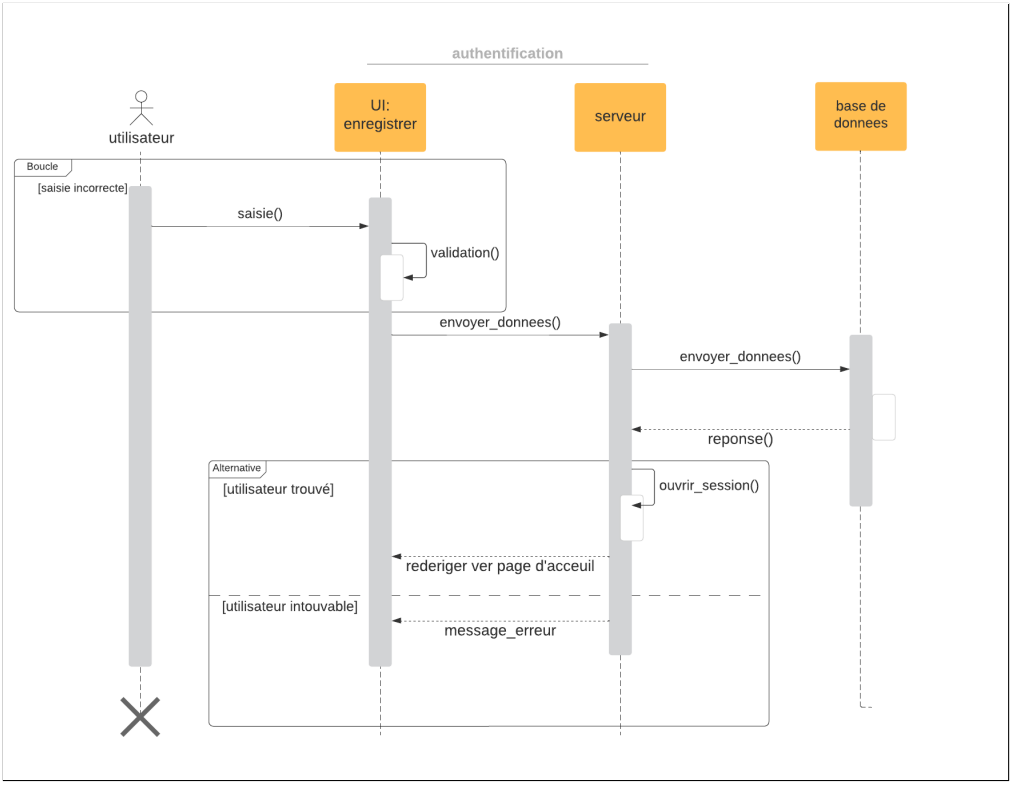

Figure 4.1 : diagramme de séquence : authentification

4.2.2 Diagramme de séquence : Prediction

Après authentification , l'utilisateur peut effectuer une prédiction de probabilité de défaut bancaire d'un client qui est une TPE en utilisant le modéle developpé précédemment, cette prédiction doit être sauvegardée dans la base de données avec les données de l'entreprise . La figure 4.2 est le diagramme de séquence qui décrit le processus de prédiction par détail.

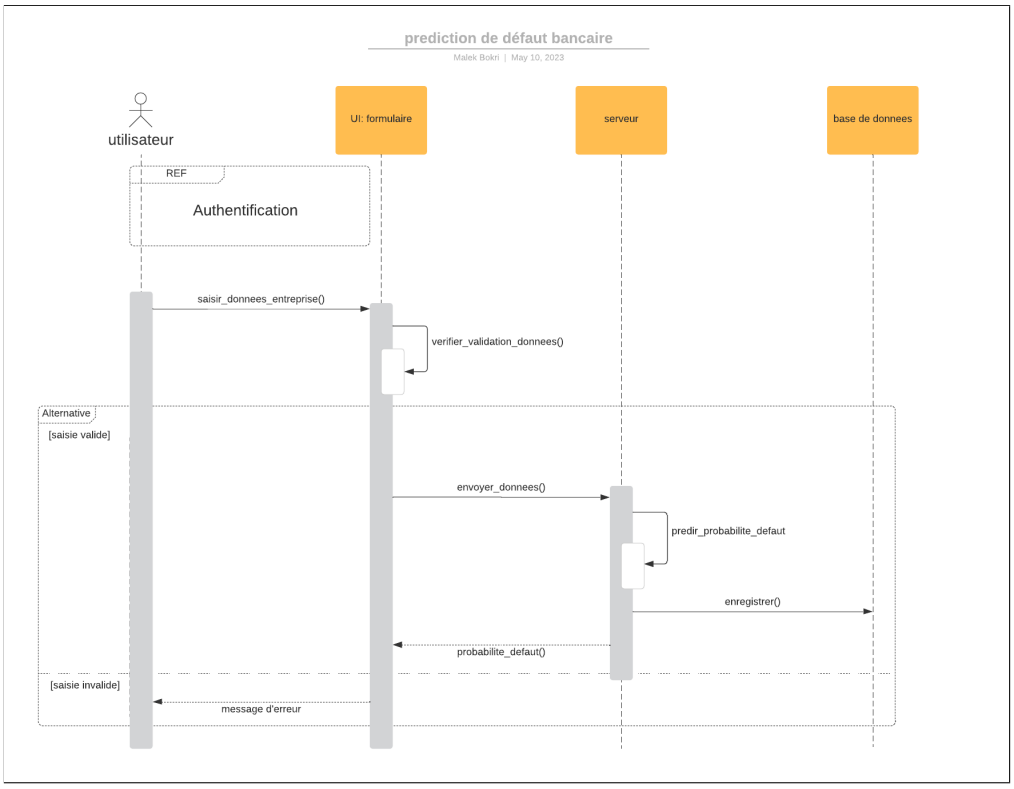

FIGURE 4.2 : diagramme de séquence : prédiction

4.2.3 Diagramme de séquence : Consultation d'historique

Un utilisateur authentifié peut consulter les prédiction précédentes qui sont déjà enregistrées dans la base de données.Il peut choisir l'ordre dans lequel l'historique est affiché comme prédiction les plus récentes d'abord ou les prédiction avec la plus grande probabilité de défaut bancaire d'abord etc . ce scénario est expliqué par le diagramme de séquence dans la figure 4.3 .

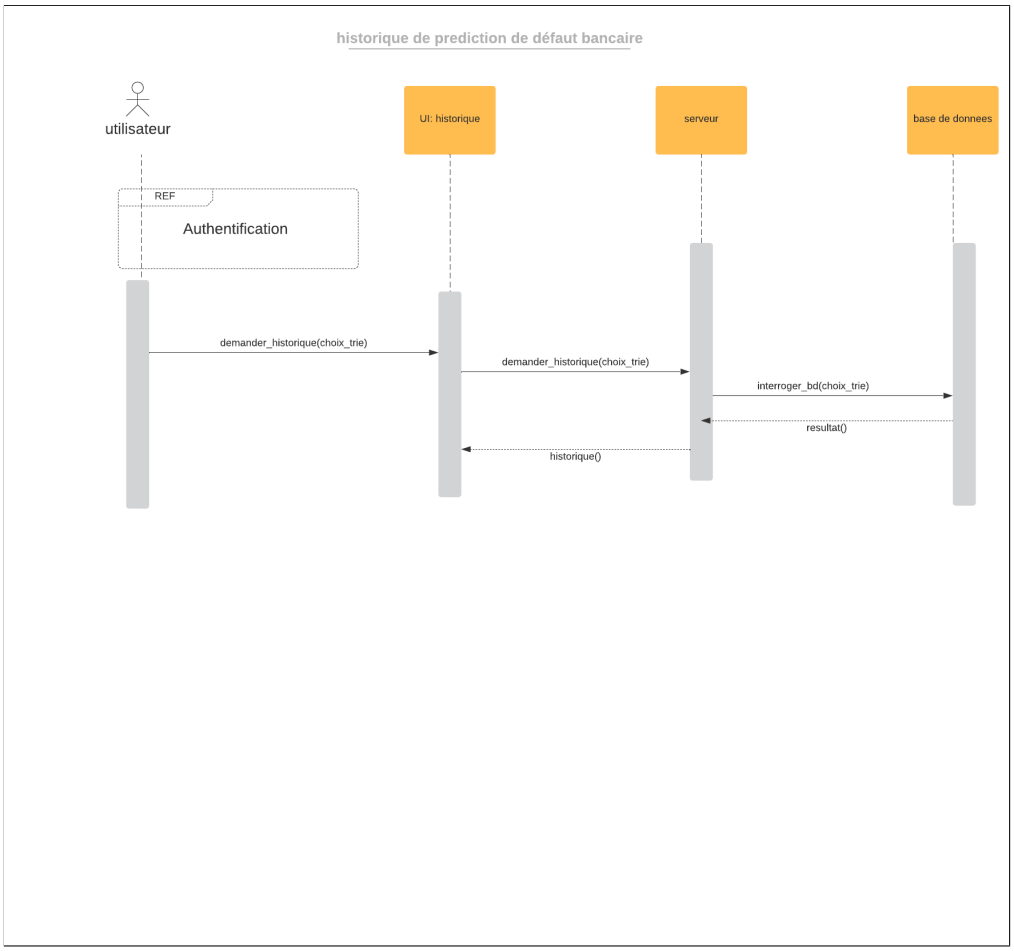

FIGURE 4.3 : diagramme de séquence : historique

4.2.4 Diagramme de séquence : ajout d'utilisateur

Après l'authentification si l'utilisateur est un administrateur il peut ajouter un utilisateur (son nom utilisateur et mot de passe) comme le montre le diagramme de séquence présenté par la figure 4.4.

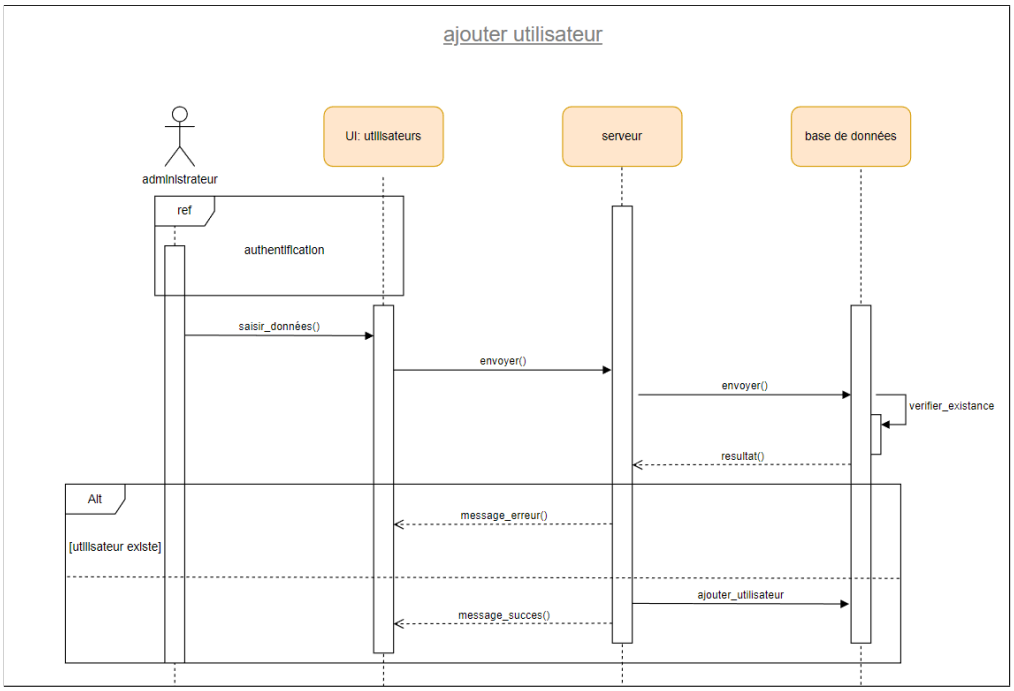

FIGURE 4.4 : diagramme de séquence :ajout d'utilisateur

4.3 Interface

4.3.1 Interface de connexion

Dans l'interface présentée dans la figure 4.5 l'utilisateur doit entrer le nom utilisateur et mot de passe et cliquer sur "connexion", si les données sont correcte il sera dirigé vers la page d'accueil sinon un message d'erreur sera affiché sous forme d'un 'alert'.

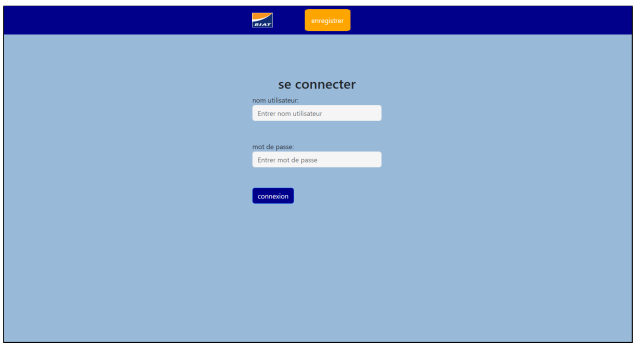

FIGURE 4.5 : interface de connexion

4.3.2 Interface page d'accueil

Comme on le constate dans la figure 4.6 c'est une interface simple pour naviguer entre les interface de prédiction et historique ou on peut deconnecter pour fermer la session.

FIGURE 4.6 : interface de page d'accueil

4.3.3 Interface de prédiction

cette interface montrée dans la figure 4.7 nous permet de consommer l API du modèle conçus pour prédire la probabilité de défaut bancaire , il suffit de remplir le formulaire et appuyer sur le bouton prédire . le résultat sera affiché sur cette même page en dessous du bouton.En appuyant sur le bouton ''réinitialiser on supprime toutes les entrée dans le formulaire .

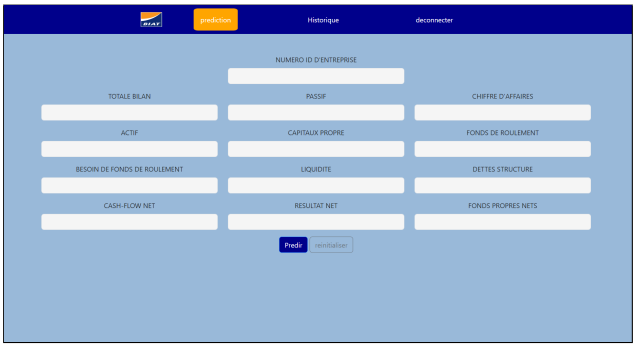

FIGURE 4.7 : interface de prédiction

4.3.4 Interface d'historique

L'interface de l'historique (figure 4.8) consiste d'un tableau dans lequel s'affiche chaque prédiction avec ses détail et la possibilité de la supprimer , ainsi qu'un champs pour rechercher une prédiction par id d'entreprise et un bouton pour choisir le trie avec lequel les prédiction doivent être affichés .

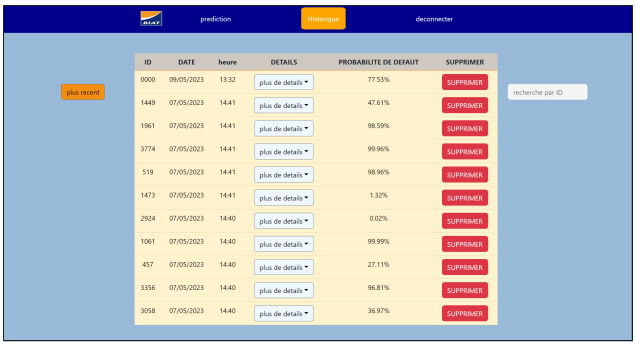

Figure 4.8 : interface de l'historique

4.3.5 Interface d'utilisateurs

L'interface d'utilisateurs affiche dans la figure 4.9 une interface ou un administrateur peut consulter les utilisateur existant et leur coordonnées ,chercher un utilisateur par nom d'utilisateur , ajouter un utilisateur en entrant un nom d'utilisateur et un mot de passe ou supprimer un existant en appuyant sur le bouton "supprimer".

Figure 4.9 : interface des utilisateurs

Conclusion

Dans ce chapitre, nous avons développé une application web pour utiliser le modèle de prédiction de défaut bancaire. Nous avons expliqué les technologies utilisées, présenté les diagrammes de séquence de notre application, et nous avons eu quelques aperçus des interfaces de l'application.

Conclusion générale

— Dans le cadre du projet PFE, nous avons réalisé, au sein de la Banque Internationale Arabe de Tunisie, un modèle de machine learning performant et fiable capable de prédire les défauts bancaires des très petites entreprises, et de les entraîner avec un ensemble de données annotées.

— Dans ce projet, nous avons défini et expliqué des termes financiers utilisés pour la prédiction, tels que le défaut bancaire et les ratios financiers. Nous avons exploré et prétraité les données pour avoir le plus d'informations possible à partir de ces dérniers. Ensuite, nous avons entraîné et testé les modèles avec des métriques bien précises. Pour finir, nous avons mis en place une application web simple pour effectuer et enregistrer les prédictions.

— Dans cette expérience, nous avons eu une idée sur le domaine de la science des données et du machine learning, les modèles, leurs avantages et les techniques de prétraitement. Nous avons aussi eu l'occasion d'en apprendre davantage sur le domaine des finances et le secteur bancaire. Ce stage m'a offert aussi l'occasion de travailler avec des technologies nouvelles pour moi, telles que le framework Flask de Python et un système de gestion de base de données non relationnel, à savoir MongoDB.

— Dans ce projet, nous avons consacré le plus de ressources, de temps et d'énergie à la partie apprentissage supervisé avec un ensemble de données relativement reduit alors l'application web développée peut être optimisée en :

  — Ajoutant des algorithmes d'apprentissage profond

  — utilisant une base de données d'apprentissage plus grande et plus détaillés pour renforcer le modéle

  — Améliorant le design des interfaces.

  — Ajoutant des méthodes de récupération de mots de passe.

  — Perfectionnant l'aspect sécurité de l'application.

Bibliographie

[1] « presentation du BIAT. » (), adresse : https://www.biat.com.tn/la-biat/presentation-generale.

[2] « TPE. » (), adresse : https://www.tanitjobs.com/blog/43/les-tpe-pme-crÃľatrices-d-emploi-en-tunisie.

[3] A. Nedjima et G. Wahiba, « Analyse financière d'une entreprise : cas de la STH-DRC de Bejaia, » Mémoire de maîtrise, Université A. Mira de Bejaia, 2017.

[4] A. Bhandari. « valeurs aberrantes. » (), adresse : https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/.

[5] R. A. A. Viadinugroho. « ENN. » (), adresse : https://towardsdatascience.com/imbalanced-classification-in-python-smote-enn-method-db5db06b8d50.

[6] « RENN. » (), adresse : https://imbalanced-learn.org/dev/references/generated/imblearn.under_sampling.RepeatedEditedNearestNeighbours.html.

[7] « correlation de Pearson et signification de corrélation. » (), adresse : http://www.biostat.ulg.ac.be/pages/Site_r/corr_pearson.html.

[8] « ROC AUC. » (), adresse : https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[9] A. Bhandari. « standardisation. » (), adresse : https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/.

[10] « regression logistique. » (), adresse : https://datatab.fr/tutorial/logistic-regression.

[11] « gradiant boosting. » (), adresse : https://www.kaggle.com/code/alexisbcook/xgboost.

[12] S. Bahroun, *Arbres de décision*, Cours de classe, 2022.

[13] « forêt aléatoire. » (), adresse : https://www.datacamp.com/tutorial/random-forests-classifier-python.

[14] « machine à vecteur de support. » (), adresse : https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

[15] « perceptron multicouche. » (), adresse : https://www.simplilearn.com/tutorials/deep-learning-tutorial/multilayer-perceptron.

[16] « matrice de confusion. » (), adresse : https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.

[17] « métriques utilisés. » (), adresse : https://inside-machinelearning.com/en/recall-precision-f1-score-simple-metric-explanation-machine-learning/.

Annexes

Détail du SVM

X l'ensemble des données d'entraînement avec n échantillons et m caractéristiques.
y vecteur de la variable cible
voici les étapes pour touver l'hyperplan d'un SVM

1 Calcul de la matrice de similarité : La matrice de similarité K est calculée à partir des données d'entraînement en utilisant la fonction noyau RBF. Pour chaque paire d'échantillons (i, j), la similarité est calculée selon la formule : $K(i, j) = \exp(-\gamma||X(i) - X(j)||^2)$, où γ est un paramètre du noyau.

2 Définition du problème d'optimisation : Il faut trouver les multiplicateurs de Lagrange (α) qui définissent l'hyperplan optimal de séparation.

Le problème d'optimisation peut être formulé comme suit :

Minimiser : $\frac{1}{2}\alpha^T K\alpha - \alpha^T y$
Sous contraintes : $y^T\alpha = 0$ $0 \leq \alpha \leq C$

où α est le vecteur des multiplicateurs de Lagrange,avec C est le paramètre de régularisation.

3 Résolution du problème d'optimisation : Le problème d'optimisation peut être résolu en utilisant des méthodes d'optimisation telles que la méthode des multiplicateurs de Lagrange . Une fois les multiplicateurs de Lagrange (α) trouvés, les vecteurs de support peuvent être trouvés. Les vecteurs de support sont les observations qui ont un multiplicateur de Lagrange non nul.

4 Calcul des coefficients de l'hyperplan : Les coefficients de l'hyperplan peuvent être calculés à partir des vecteurs de support et des multiplicateurs de Lagrange selon les formules suivantes :

$w = \sum(\alpha_i y_i X_i)$ $b = y_k - \sum(\alpha_i y_i K(X_i, X_k))$

où w est le vecteur de poids, b est le biais et k est l'indice d'un vecteur de support.

5 Resultat La classe prédite est déterminée en fonction du signe de $f(x) = \text{sign}\,(\sum(\alpha_i y_i K(x, X_i)) + b)$.

Detail du MLP

X l'ensemble des données d'entraînement avec n échantillons et m caractéristiques.
y vecteur de la variable cible
voici les étapes pour définir les coefficients d'un MLP

1 Initialisation des poids : Soit $W^{(1)}$ la matrice des poids entre la couche d'entrée et la couche cachée. Soit $W^{(2)}$ la matrice des poids entre la couche cachée et la couche de sortie. Les poids sont généralement initialisés aléatoirement.

2 Calcul de la propagation avant : Pour chaque échantillon d'entrée x_i : Calculez l'activation de la couche cachée $a^{(1)}$ selon la formule : $a^{(1)} = XW^{(1)}$ Appliquez une fonction d'activation non linéaire à $a^{(1)}$ pour obtenir les activations de la couche cachée $h^{(1)}$. Calculez l'activation de la couche de sortie $a^{(2)}$ selon la formule : $a^{(2)} = h^{(1)}W^{(2)}$ Appliquez une fonction d'activation appropriée à $a^{(2)}$ pour obtenir les activations de la couche de sortie $h^{(2)}$.

3 Calcul de la fonction de perte : Utilisez une fonction de perte appropriée pour mesurer l'écart entre les prédictions $h^{(2)}$ et les étiquettes de classe y.

4 Rétropropagation du gradient : Calculez le gradient de la fonction de perte par rapport aux poids $W^{(1)}$ et $W^{(2)}$ en utilisant la rétropropagation du gradient. Mettez à jour les poids en utilisant une méthode d'optimisation telle que la descente de gradient.

5 Répéter Répétez les étapes 2 à 4 jusqu'à ce que la convergence soit atteinte (c'est-à-dire que la fonction de perte soit minimisée).

6 Résultat Utilisez les poids entraînés pour effectuer une propagation avant sur de nouvelles instances et obtenir les prédictions de la classe.

ملخّص

في إطار مشروع التخرج وبغية الحصول على درجة الإجازة في علوم الحاسوب، قمنا بتطوير تطبيق ويب لتوقع العجز المصرفي للشركات الصغيرة جداً في البنك العربي الدولي لتونس . يتضمن المشروع تطوير نموذج للتعلم الآلي، وتدريبه واختباره باستخدام بيانات موسومة، ثم تطوير تطبيق ويب للتنبؤ بالعجز المصرفي باستخدام النموذج. تم استخدام لغة بايثون لتطوير الويب باستخدام إطار العمل فلاسك، وفي جانب التعلم الآلي باستخدام وحدة سايكت لارن.

كلمات مفاتيح : بايثون ، فلاسك ، سايكت لارن ، تطوير المواقع الشبكية ، التعلم الآلي

Résumé

Dans le cadre du projet de fin d'études et en vue de l'obtention de la licence en Science Informatique, nous avons réalisé une application web de prédiction des défauts bancaires pour les très petites entreprises au sein de la Banque Internationale Arabe de Tunisie (BIAT). Le projet consiste à développer un modèle d'apprentissage automatique, le former et le tester avec des données étiquetées, puis développer une application web de prédiction qui l'exploite. Pour ce projet, nous avons utilisé Python pour le développement web avec le framework Flask, et dans la partie d'apprentissage automatique avec le module Scikit-learn..

Mots clés : Python , Flask , scikit-learn , developpement web , apprentissage automatique

Abstract

As part of the final year project and with the aim of obtaining a Bachelor's degree in Computer Science, we have developed a web application for predicting banking defaults for micro-enterprises within the Arab International Bank of Tunisia (BIAT). The project involves developing a machine learning model, training and testing it with labeled data, and creating a web application for making predictions based on the model. Python was used for web development, using the Flask framework, and for machine learning, using the Scikit-learn module.

Keywords : Python , Flask, scikit-learn , web development,machine learning

2 نهج أبو الريحان الباروني 2080 أريانة الهاتف : 164 706 71 الفاكس : 698 706 71 البريد الالكتروني : isi@isim.rnu.tn
2, Abou Raihane Bayrouni 2080 l'Ariana **Tél** : 71 706 164 **Fax** : 71 706 698 **Email** : isi@isim.rnu.tn