Revue du Contrôle de la Comptabilité et de l'Audit
ISSN: 2550-469X
Volume 5 : Numéro 4

L'utilisation de la méthode KPV émanant de l'intelligence artificielle pour la prédiction de la solvabilité des clients bancaires

Using the KNN method from artificial intelligence to the prediction of the solvency of banking customers

AMZILE Karim
Doctorant
Faculté des Sciences Juridiques, Economiques et Sociales-Agdal
Université Mohammed V Rabat
Département Sciences de Gestion
Maroc
karim.amzile@um5r.ac.ma

AMZILE Rajaa
Enseignante chercheuse
Faculté des Sciences Juridiques, Economiques et Sociales-Agdal
Université Mohammed V Rabat
Département Sciences de Gestion
Maroc
r.amzile@um5s.net.ma

Date de soumission : 07/11/2021

Date d'acceptation : 23/12/2021

Pour citer cet article :

AMZILE. K & AMZILE. R (2021) «L'utilisation de la méthode KPV émanant de l'intelligence artificielle pour la prédiction de la solvabilité des clients bancaires», Revue du contrôle, de la comptabilité et de l'audit « Volume 5 : Numéro 4» pp : 110 – 123

Résumé

La gestion du risque de crédit est un sujet important pour les banques et les établissements socio-économiques qui recueillent d'énormes quantités de données, dans l'intention de rendre obsolète la mauvaise décision. Dans cet article, nous avons étudié le comportement du classificateur KPV (K plus Proche Voisin) à la prédiction de la solvabilité des clients d'une banque. Ce classificateur serve à trouver la classe d'un nouveau client qui désire obtenir un crédit auprès d'une banque. À cet effet nous avons utilisé une base de données des clients d'une banque qui comporte des clients solvables et non-solvable.

Étant donné que la méthode utilisée relevant des techniques de l'intelligence artificielle, nous avons utilisé le langage Python comme outil de modélisation, par conséquent, nous avons commencé notre processus de modélisation par un prétraitement des données, par la suite nous avons exploré les différents résultats obtenus par les différentes distances, afin que nous puissions choisir la meilleure valeur de K, ensuite nous avons évalué et comparé les différents modèles de prédiction obtenus. Au terme du processus suivi, nous avons pu conclure que le modèle obtenu par la méthode $KPV - Manhattan$ donne des résultats satisfaisants en termes de niveau de précision et de prédictibilité, avec un niveau de précision qui dépasse les 93%.

Mots clés : KPV; Risque de crédit; Intelligence Artificielle; Data Mining; KNN.

Abstract

Credit risk management is an important topic for banks and socio-economic institutions that collect huge amounts of data, with the intention of making the wrong decision obsolete. In this article, we studied the behaviour of the KNN (K Nearest Neighbor) classifier to the prediction of the solvency of a bank's customers. This classifier is used to find the class of a new customer who wishes to obtain credit from a bank. For this purpose, we have used a database of the customers of a bank which includes creditworthy and non-financial customers.

Since the method used was based on artificial intelligence techniques, we used the Python language as a modeling tool, so we started our modelling process by pre-processing the data, later we explored the different results obtained by the different distances, so that we could choose the best value of K, then we evaluated and compared the different prediction models obtained. At the end of the process, we were able to conclude that the model obtained by the KPV-Manhattan method gives satisfactory results in terms of level of precision and predictability, with a level of precision that exceeds 93%.

Keywords: KNN ; Data Mining **;** Artificial Intelligence ; credit risk ; KPV

Introduction

Avec le développement rapide du marché de crédit, ainsi que la fréquence du risque de crédit augmente et les pertes ultérieures encourues par les banques augmentent. Cela a rendu l'évaluation du crédit des demandeurs primordial et nécessaire, cependant un bon modèle de prévision peut aider les banques à prendre les bonnes décisions afin de réduire ce risque de défaut ainsi de minimiser les pertes encourues.

Il existe de nombreuses méthodes améliorées émanant de l'intelligence artificielle ; ces méthodes se caractérisent par leur robustesse ainsi que leur niveau élevé de précision et d'efficacité dans la classification des données. Dans cet article nous décrivons la méthode K-plus proche voisin (KPV) ainsi que son niveau de performance dans la prédictibilité de solvabilité des clients bancaires, à cet effet, nous avons utilisé des données des clients d'une banque et on a utilisé le langage de programmation Python pour explorer les performances de la méthode KPV.

Par conséquent notre méthodologie consiste à utiliser le processus des techniques du Data Mining, en commençant par un pré-traitement et une analyse exploratoire des données, ensuite la construction d'un algorithme d'apprentissage automatique du KPV et enfin l'évaluation du modèle à l'aide des diverses métriques de validation ainsi que les ratios d'erreur.

Dans le présent article on a essayé de répondre à la question :

Est-ce que les méthodes émanant de l'intelligence artificielle peuvent être une meilleure alternative des méthodes classiques dans la gestion du risque de crédit ?

Cependant, pour répondre à cette question, nous avons choisi la méthode KPV pour pouvoir y répondre. De ce fait nous avons commencé notre article par une revue de littérature, qui explore les différents propos et résultats obtenus par des chercheurs scientifiques, par la suit nous avons défini la méthode KPV et sa formulation mathématique, ensuite nous avons parcouru les données historiques des clients bancaires utilisées ainsi que leur structure, et nous avons analysé les différentes variables explicatives afin de garder que les variables significatives. Au terme de notre article nous avons exposé les résultats atteints ainsi le niveau de prédictibilité des modèles obtenus.

1. Revue de littérature

La classification permet de prévoir les classes des données de test après avoir entrainé le modèle par une telle méthode de classification en utilisant des données d'entrainement. Au

cours des dernières décennies, un grand nombre de méthodes de classification ont été mises au point dans des applications réelles (Luo et al, 2016) (Li et al,2016), parmi lesquelles la classification KPV (kNN) a été considérée comme l'une des 10 principales méthodes de classification les plus utilisées (Wu et al, 2008), en raison de sa simplicité et de son efficacité. Ainsi, la méthode kNN a été développée avec succès dans des applications du Data Mining, telles que la classification, la régression et le remplissage des valeurs manquantes. L'idée principale du KPV est de prédire l'étiquette d'un individu émanant de données de test par la règle de vote, c'est-à-dire que l'étiquette de l'individu est prédite via les k individus de données d'entrainement les plus similaires (Cheng et al, 2015).

Pour (Shichao Zhang et al, 2017) [1]les résultats expérimentaux d'utilisation du KPV ont montré que la méthode est plus précise et plus efficace, ainsi que, en plus des tâches de classification, la méthode KPV on peut l'utiliser pour la régression et l'imputation des données manquantes.

2. Présentation du modèle KPV

La méthode KPV (K-plus proche Voisin) ou KNN (K- nearest neighbors) figure parmi les plus simples algorithmes d'apprentissage automatique (relevant de l'intelligence artificielle). Par conséquent, les KNN sont aussi connus comme des apprenants paresseux (Cunningham et Delany, 2007). Toutefois, les KNN ont réussi à régler un grand nombre de problèmes commerciaux (Jiang et al, 2012) et (Mccord et al, 2011).

Dans un contexte de classification d'une nouvelle observation $\boldsymbol{x}$, l'idée fondatrice simple est de faire voter les plus proches voisins de cette observation.

Formellement, soit $\boldsymbol{L}$ l'ensemble de données à disposition ou échantillon d'apprentissage :

$$L = \{(y_i, x_i), i = 1, \dots \dots, n_L\}$$

Avec :

- y_i (Variable à expliquer / dépendante) il reflète la classe de l'individu i pour le cas de classification et il représente un réel $y_i \in \mathbb{R}$ dans le cas de la régression
- $x_{ij} = (x_{i1}, \dots \dots \dots . x_{np})$ $i\epsilon\{1, n\}$ Représente l'indice de l'individu $j\epsilon\{1, p\}$ Représente l'indice des variables prédictives (variables explicatives) de l'individu i.

Pour effectuer une prédiction, l'algorithme KNN va se baser sur le jeu de données en entier. En effet, pour une observation, qui ne fait pas parti du jeu de données, qu'on souhaite prédire, l'algorithme va chercher les K observations du jeu de données les plus proches de notre

observation. Ensuite pour ces K voisins, l'algorithme se basera sur leurs variables dépendantes (variable à expliquer) y pour calculer la valeur de la variable $\hat{y}$ de l'observation qu'on souhaite prédire.

Par ailleurs :

- Si KNN est utilisé pour la régression, c'est la **moyenne** (ou la médiane) des variables y des K plus proches voisins observations qui servira pour la prédiction
- Si KNN est utilisé pour la classification, c'est le **mode** des variables y des K plus proches voisins observations qui servira pour la prédiction

3. Les fonctions du calcul de la distance

3.1. La distance *euclidienne*

Parmi les fonctions distance types, la distance euclidienne entre deux observations x_i, x_j est définie comme suit :

$$D_{euc}(x_i, x_j) = \left(\sum_{S=1}^{P}(x_{is} - x_{js})^2\right)^{\frac{1}{2}} = \sqrt{\sum_{S=1}^{P}(x_{is} - x_{js})^2}$$

$$avec : P\ représente\ le\ nombre\ de\ variable\ explicative$$

3.2. La distance *Minkowski*

Généralement on peut écrire distance de **Minkowski** :

$$D_{Mink}(x_i, x_j) = \left(\sum_{S=1}^{P}|x_{iS} - x_{jS}|^q\right)^{\frac{1}{q}} \quad avec\ q > 0$$

Quand $q = 2$ la distance de Minkowski est égale à la distance euclidienne et quand $q = 1$ il est égal à la distance de Manhattan.

3.3. La distance Manhattan

On peut écrire la fonction Distance de Manhattan comme suit :

$$D_{Manh}(x_i, x_j) = \sum_{S=1}^{P}|x_{iS} - x_{jS}|$$

Les mesures de distance les plus courantes sont les mesures de distance euclidienne et de Manhattan, les deux mesures mesurent la distance entre les observations pour tous les s variables explicatives.

4. Quelques règles sur le choix de k

Le paramètre k doit être déterminé par l'utilisateur : $K \in N$. En classification binaire, il est utile de choisir k impair pour éviter les votes égalitaires. Le meilleur choix de k dépend du jeu de donnée. En général, les grandes valeurs de k réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, mais rendent les frontières entre classes moins distinctes. Un bon k peut être sélectionné par diverses techniques heuristiques, par exemple, de validation-croisée. Nous choisirons la valeur de k qui minime l'erreur de classification.

5. Application du modèle KPV à la prédiction de la solvabilité des clients d'une banque

Pour les données utilisées, on a utilisé un échantillon qui comporte 30000 lignes de clients d'une banque, chaque ligne comporte une multitude de variables de type quantitatives et qualitatives. Nous avons essayé de nettoyer les données en éliminant toutes variables aberrantes et toute ligne et colonne qui comporte des valeurs manquantes.

Après ce pré-traitement des données, nous avons obtenu une base de données qui comporte 4000 lignes et 14 colonnes, parmi ces variables utilisées nous avons 13 variables explicatives et une seule variable à expliquer qui représente la solvabilité des clients.

Cependant, pour assurer une meilleure modélisation nous avons pensé à la codification des variables qualitative nominale. Le tableau ci-après (Tableau 1) illustre les valeurs de chaque variable explicative après la codification :

5.1. La Codification des variables

Tableau 1 : La codification des variables utilisées

	NAME_CONTRACT_TYPE
Cash loans	1
Revolving loans	0
	CODE_GENDER
F	1
M	0
	FLAG_OWN_CAR
Y	1
N	0
	NAME_INCOME_TYPE
State servant	1

Working	2
Commercial associate	3
Pensioner	4
	NAME_FAMILY_STATUS
Married	1
Single / not married	2
Civil marriage	3
Separated	4
Widow	5
	NAME_HOUSING_TYPE
House / apartment	1
With parents	2
Municipal apartment	3
Office apartment	4
Co-op apartment	5
Rented apartment	6
	NAME_EDUCATION_TYPE
Higher education	1
Incomplete higher	2
Secondary / secondary special	3
Lower secondary	4

Source : Auteurs

5.2. L'analyse univariée

Selon le résultat de l'exécution de l'analyse univarié sur SPSS, nous avons obtenu le tableau (Tableau 2) ci-dessous :

Tableau 2 : Tableau d'analyse univariée

Vi	*L'intitulé de la variable*	*B*	*E.S.*	*Wald*	*ddl*	*Sig*
V_2	NAME_CONTRACT_TYPE	,429	,132	10,522	1	,001
V_3	CODE_GENDER	-1,604	,165	94,361	1	,000
V_4	FLAG_OW0_CAR	-,424	,085	24,811	1	,000
V_5	CNT_CHILDREN	,341	,126	7,298	1	,007
V_6	AMT_INCOME_TOTAL	,000	,000	39,407	1	,000
V_7	AMT_CREDIT	,000	,000	34,029	1	,000
V_8	AMT_ANNUITY	,000	,000	,435	1	,510
V_9	AMT_GOODS_PRICE	,000	,000	44,091	1	,000
V_{10}	NAME_INCOME_TYPE	-,375	,063	35,415	1	,000
V_{11}	NAME_EDUCATION_TYPE	,746	,038	375,557	1	,000

V_{12}	NAME_FAMILY_STATUS	-,017	,043	,151	1	,697
V_{13}	NAME_HOUSING_TYPE	,108	,044	6,167	1	,013
V_{14}	CNT_FAM_MEMBERS	-,263	,111	5,650	1	,017
	Constante	1,497	,354	17,887	1	,000

Source : Auteurs

Selon l'analyse univariée on peut conclure que les variables $\boldsymbol{V_8}$ $\boldsymbol{(AMT_{ANNUITY})}$ $\boldsymbol{et}$ $\boldsymbol{V_{12}}$ **(NAME_FAMILY_STATUS)** ne sont pas significatives, et ce à cause de leurs valeurs de signification qui dépassent les 5%, cependant on doit les éliminer pour pouvoir construire un modèle puissant en termes de significativité en gardant seulement les variables qu'ont une probabilité significative.

5.3. La matrice de corrélation

Cependant on peut aussi vérifier pour le cas des variables explicatives qualitatives, s'il y a des corrélations entres les variables dépendantes, à cet effet on tracer la matrice de corrélation ci-après (Tableau 3) :

Tableau 3 : Matrice de corrélation RL

Correlations											
	V2	***V3***	***V4***	***V5***	***V6***	***V7***	***V9***	***V10***	***V11***	***V13***	***V14***
V2	*1*	*-,023*	*-,034**	*-,004*	*,000*	*,227***	*,188***	*-,027*	*,053***	*-,003*	*,007*
V3	*-,023*	*1*	*-,093***	*,059***	*-,026*	*,063***	*,062***	*,001*	*-,049***	*-,037**	*,079***
V4	*-,034**	*-,093***	*1*	*,083***	*,179***	*,108***	*,116***	*,024*	*-,146***	*-,053***	*,109***
V5	*-,004*	*,059***	*,083***	*1*	*-,056***	*-,033**	*-,042***	*-,013*	*-,014*	*,014*	*,888***
V6	*,000*	*-,026*	*,179***	*-,056***	*1*	*,379***	*,388***	*,196***	*-,273***	*-,058***	*-,067***
V7	*,227***	*,063***	*,108***	*-,033**	*,379***	*1*	*,986***	*,106***	*-,197***	*-,072***	*,012*
V9	*,188***	*,062***	*,116***	*-,042***	*,388***	*,986***	*1*	*,108***	*-,215***	*-,073***	*,005*
V10	*-,027*	*,001*	*,024*	*-,013*	*,196***	*,106***	*,108***	*1*	*-,104***	*-,005*	*-,028*
V11	*,053***	*-,049***	*-,146***	*-,014*	*-,273***	*-,197***	*-,215***	*-,104***	*1*	*,074***	*-,010*
V13	*-,003*	*-,037**	*-,053***	*,014*	*-,058***	*-,072***	*-,073***	*-,005*	*,074***	*1*	*-,015*
V14	*,007*	*,079***	*,109***	*,888***	*-,067***	*,012*	*,005*	*-,028*	*-,010*	*-,015*	*1*

Source : Auteurs

Selon (Tableau 3), on peut remarquer qu'il y a une forte corrélation entre V_9 et V_7 ainsi entre V_{14} et V_5, à cet effet on va procéder à élimination d'une seule variable dans chaque couple fortement corrélé, et ce pour assurer une meilleure prédictibilité du modèle. Pour cela on va éliminer $V_9 et\ V_{14}$.

Donc on garde seulement 9 variables explicatives, cependant la valeur de p=9 alors les fonctions de distance s'écrivent comme suit :

$$D_{euc}(x_i, x_j) = \left(\sum_{S=1}^{9}(x_{is} - x_{js})^2\right)^{\frac{1}{2}}$$

Pour la méthode $Minkowski$ nous avons retenu la valeur de $q = 3$:

$$D_{Mink}(x_i, x_j) = \left(\sum_{S=1}^{9}|x_{iS} - x_{jS}|^3\right)^{\frac{1}{3}}$$

$$D_{Manh}(x_i, x_j) = \sum_{S=1}^{9}|x_{iS} - x_{jS}|$$

6. Résultats obtenus

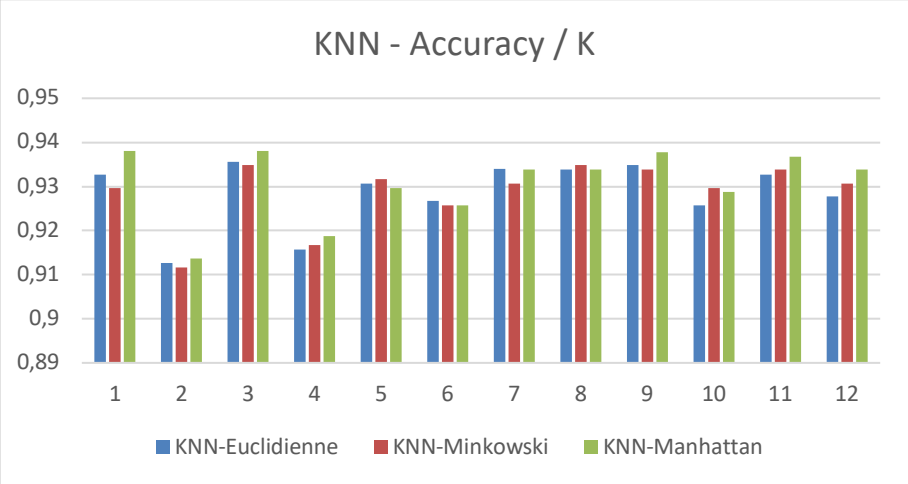
Pour choisir la valeur de k la plus pertinente nous avons calculé la valeur de $Accuracy$ pour chaque valeur de k qui varie entre $\{1\ et\ 12\}$, et ce pour les trois types de distance $(Manhattan, Euclidienne, Minkowski)$ et nous avons obtenu les résultats suivants (Tableau 4) :

Tableau 4 : Accuracy de chaque distance et pour chaque valeur de K

K=	1	2	3	4	5	6	7	8	9	10	11	12
KNN-Eucl	0,9327	0,9127	**0,936**	0,916	0,931	0,927	0,934	0,9338	0,935	0,9257	0,9327	0,9277
KNN-Mink	0,9297	0,9117	0,935	0,917	0,932	0,926	0,931	**0,9348**	0,934	0,9297	0,9338	0,9307
KNN-Manh	0,9381	0,9137	**0,938**	0,919	0,93	0,926	0,934	0,9338	0,938	0,9287	0,9368	0,9338

Source : Auteurs

Figure 1 : Présentation graphique des résultats obtenus



Source : Auteurs

Selon la présentation graphique (Figure 1) des résultats obtenus nous avons retenu les valeurs de k suivantes :

Tableau 5 : les valeurs de K retenues

Type de distance	K	Accuracy
KNN-Euclidienne	3	0,9356
KNN-Minkowski	8	0,9348
KNN-Manhattan	3	0,9381

Source : Auteurs

Parmi les trois méthodes de distance nous avons retenu celle de $Manhattan\ avec\ k = 3$ et une précision de 0,9381.

7. Évaluation du modèle obtenu

Pour évaluer les modèles relavant de l'intelligence artificielle, on utilise souvent la matrice de confusion qui nous permet de calculer les différentes métriques et types d'erreur, pour cela on doit dresser la matrice de confusion pour chaque méthode de distance.

Matrice Confusion Euclidienne		
	1	0
1	87	36
0	41	1033

Matrice Confusion Manhattan		
	1	0
1	95	28
0	46	1028

Matrice Confusion Minkowski		
	1	0
1	78	45
0	33	1041

Parmi les métriques utilisées, nous avons la sensibilité ($sensivity\ (SV)$), elle représente la proportion des clients **solvables** bien classés :

$$SV = \frac{VS}{VS + FS}$$

La **spécificité** ($specificity\ (SP)$) elle représente la proportion des clients **Insolvables** bien classés :

$$SP = \frac{VD}{VD + FD}$$

Et il y a aussi *l'Accuaracy* qui nous permet de mesurer la performance de notre modèle :

$$\boldsymbol{Accuracy} = \frac{Vrais\ négatifs + vrais\ Positifs}{vrais\ negatifs + Faux\ negatifs + Vrais\ Positifs + Faux\ positifs}$$

7.1. La mesure de taux des erreurs

On va utiliser trois ratios d'erreur :

$$e_1 = \frac{Nombre\ des\ observations(1)\ classeés\ (0)}{Nombre\ des\ obs(1)}$$

$$e_2 = \frac{Nombre\ des\ obs(0)\ classeés\ (1)}{Nombre\ des\ obs(0)}$$

$$e_3 = \frac{Nombre\ des\ obs(1)\ classeés\ (0) + Nombre\ des\ obs(0)\ classeés\ (1)}{Nombre\ des\ obs}$$

Après avoir calculer les différentes métriques et ratios d'erreur nous avons obtenu les résultats suivants (tableau 6) :

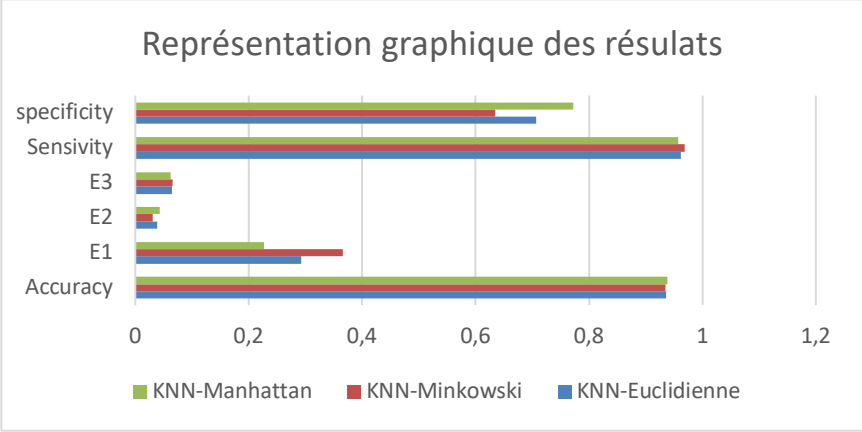
Tableau 6 : les valeurs d'erreurs pour chaque distance

	K	Accuracy	E1	E2	E3	Sensivity	Specificity
KNN-Euclidienne	3	0,9356	0,2926	0,0381	0,0643	0,9618	0,7073
KNN-Minkowski	8	0,9348	0,3658	0,0307	0,0651	0,9692	0,6341
KNN-Manhattan	3	0,9381	0,2276	0,0428	0,0618	0,9571	0,7723

Source : Auteurs

Après avoir calculer les différents indicateurs, nous avons conclu que la méthode Manhattan utilisée par KPV avec un k=3 est la meilleure en termes de niveau de prédictibilité de la solvabilité des clients bancaires, ainsi que son taux d'erreur le plus faible.

Figure 2 : Présentation graphique des résultats

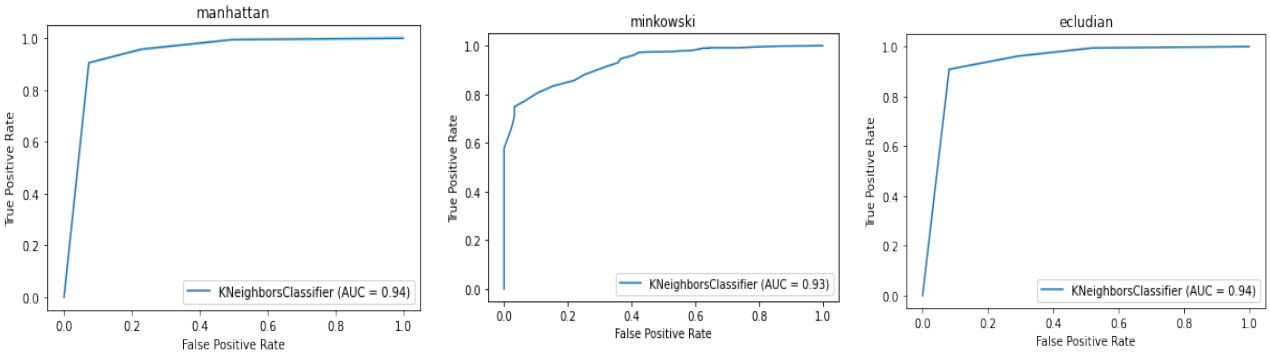


Source : Auteurs

7.2. La courbe ROC

On peut aussi mesurer ma performance par la présentation graphique de la courbe ROC (Figure 3), et aussi via la valeur AUC :

Figure 3 : La courbe ROC des 3 distances



Source: Auteurs

Selon la courbe $ROC_{\ Manhattan}$ ci-dessus on peut remarquer clairement la performance de notre modèle retenu, celui du Manhattan avec $AUC_{Manhattan} = 94\%$.

Conclusion

La gestion du risque de crédit a fait l'objet d'une plus grande attention au cours des dernières années. Dans cet article, nous avons pu dévoiler l'intérêt d'exploiter la performance des méthodes d'apprentissage automatiques à la prédiction de la solvabilité des clients bancaires. A cet effet nous avons utilisé la méthode KPV pour élaborer un modèle capable de prédire si un client bancaire est solvable ou non. Après avoir entrainer notre modèle KPV nous avons pu obtenir un modèle puissant, ce qui affirme aussi les propos de (Wu et al, 2008) dans leur papier scientifique.

Au terme de notre processus méthodologique, nous avons mesuré la performance de notre modèle par les différentes métriques. Cependant le modèle à base du KPV-Manhattan a été retenu, puisqu'il a donné des résultats qui reflètent une performance assez élevée de 94%, de ce fait, le modèle obtenu s'avère intéressant et il peut être recommander pour toute opération de gestion du risque de crédit.

Comme tout étude scientifique, le présent travail comporte certaines limites ; en termes des données utilisées et leurs disponibilités, par conséquent, nous pensons qu'il faut encore élargir le champ des variables indépendantes en vue d'avoir une vision plus claire et précise sur le comportement des clients bancaires.

Cependant, cette étude ouvre des nouvelles pistes de recherches dans lesquelles des études scientifiques ultérieures pourront être menées, de ce fait nous proposons une application de ces méthodes d'apprentissage automatique à la gestion des différents risques financiers.

BIBLIOGRAPHIE

(D. A. Anggoro, 2020) « The Implementation of Subspace Outlier Detection in K-Nearest Neighbors to Improve Accuracy in Bank Marketing Data », International Journal of Emerging Trends in Engineering Research.

(Delany, et al, 2007) "k-Nearest Neighbour Classifiers", Computer Science

(F.-L. Chen, et al, 2010) « Comparison of the Hybrid Credit Scoring Models Based on Various Classifiers », International Journal of Intelligent Information Technologies.

(H. Zhou, et al, 2013) « Application of the Hybrid SVM-KNN Model for Credit Scoring », in 2013 Ninth International Conference on Computational Intelligence and Security.

(J. Yao, et al, 2019) « Detecting Fraudulent Financial Statements for the Sustainable Development of the Socio-Economy in China: A Multi-Analytic Approach », Sustainability.

(S. Zhang, et al, 2017) « Learning k for kNN Classification », ACM Trans. Intell. Syst. Technol.