École Centrale Marseille

Rapport de projet SISN

Analyse de données de l'approbation de crédit

Ningwei XIE, Jingwei CHEN, Yuanhang XIANG, Shuze HE

Tuteur : Mouhamadou Lamine DIONG

Semestre 8

Table des matières

1. Présentation du projet

1.1 Introduction de la problématique

La décision d'approuver une carte de crédit ou un prêt dépend principalement du contexte personnel et financier du demandeur. Précisément, l'âge, le sexe, le revenu, le statut d'emploi, les antécédents de crédit et d'autres attributs contribuent à la décision d'approbation. Il est important de gérer le risque de crédit et de relever les défis efficacement pour la décision de crédit. L'analyse de crédit implique la mesure statistique, quantitative et qualitative pour étudier la probabilité d'une personne de rembourser le prêt à la banque à temps et de prédire sa caractéristique par défaut. L'analyse met l'accent sur la comptabilisation, l'évaluation et la réduction des risques financiers ou autres qui pourraient entraîner des pertes pour l'entreprise pendant le prêt.

1.2 Objectifs du projet

Dans ce projet, nous nous intéressons à construire une méthodologie d'analyse qui **(a) estimer les variables importantes** dans l'ensemble de données pour mieux comprendre des politiques d'approbation de crédit, **(b) produire le meilleur modèle pour prédire** le résultat d'un demandeur.

Pour l'accomplir, nous avons utilisé plusieurs techniques de **visualisation**, de **traitement des données**, de **classification supervisée**. Dans la construction du modèle prédictif, différentes techniques de classification ont été comparées pour trouver celle qui convenait le mieux à l'ensemble de données. Pour les analystes du secteur financier, le modèle peut être intégré pour automatiser le processus d'approbation des demandes de crédit. Ensuite, nous avons utilisé des méthodes de sélection de variables afin de déterminer l'ordre d'importance des différents attributs. Ces résultats peuvent servir d'une référence d'information pour les consommateurs.

1.3 Techniques concernées

Dans ce projet, on s'intéressera à programmer des algorithmes de traitement de données et de l'apprentissage supervisé en utilisant l'environnement de programmation **IPython notebook**. Nous utilisons en particulier les modules numpy, matplotlib, pandas, keras, sklearn, etc.

Les différentes étapes du traitement de la base de données sont les suivantes :

Prétraitement de la base :

1. Transformation des données
2. Normalisation,
3. Remplissage des valeurs manquantes
4. Visualisation

Modèles de classification :

5. kNN (k-Nearest Neighbors)
6. Réseau de Neurones
7. Decision Tree
8. Discrimination linéaire : LDA
9. SVM (Support Vector Machine)

Analyse :

10. Comparaison de la performance des modèles
11. Forward feature selection

1.4 Organisation du projet

1.4.1 Définition des rôles

Pour effectuer le traitement des données, la répartition des tâches au sein du groupe fut la suivante :

XIE Ningwei :

- ✓ Manipulation du réseau de neurones
- ✓ Visualisation des données
- ✓ Analyse des modèles

CHEN Jingwei :

- ✓ Manipulation de la méthode kNN et du Decision Tree
- ✓ Pré-traitement de la base de données

XIANG Yuanhang :

- ✓ Manipulation de la discrimination linéaire et du SVM
- ✓ Analyse des modèles

HE Shuze :

- ✓ Secrétariat
- ✓ Manipulation du SVM

1.4.2 Planning

19 Février	Introduction du projet et distribution des tâches
19-23 Février	Recherche globale du domaine concerné
5-9 Mars	Déterminer des techniques à pratiquer Préciser des détails de fonctionnement des techniques
7-14 Mars	Construire un plan général de manipulation
15-23 Mars	Choisir la base de données à analyser Etudier le contexte et des contraintes de la base Expliciter les difficultés à résoudre
23 Mars-3 Avril	Préparer la présentation intermédiaire
3-8 Avril	Recueillir tous les codes Python des méthodes d'apprentissage
9-13 Avril	Pré-traitement de la base de données et visualisation
14-23 Avril	Implémenter les méthodes d'apprentissage Recueillir les résultats sur la base d'apprentissage et la base de généralisation
24 Avril-30 Avril	Optimiser les méthodes et ajuster les hyperparamètres
1 Mai-20 Mai	Analyser les résultats et comparer les méthodes Implémenter la réduction de paramètres
20-25 Mai	Préparer le rapport final

2. Traitement de la base de données

2.1 Description de la base étudiée

La base des données concerne à l'approbation de crédit, extrait de l'archive du dépôt d'apprentissage automatique de l'Université de Californie à Irvine (UCI) [1]. Les données présentées sont les informations des individus soumettant la demande de crédit et les résultats de cette demande.

Dans sa forme initiale, non modifiée, l'ensemble de données mis à disposition par l'UCI contient **690 instances**, représentant 690 individus, et **16 attributs** nommées A1 - A16. La 16ème variable contient le résultat de la demande (l'attribut de classe), soit positive (représentée par "+") signifie accorder, ou négatif (représenté par "-") signifiant rejeter. Cet ensemble de données multivariées est un mélange d'attributs ayant des valeurs de données continues (entières, réelles), discrets et catégoriques ainsi que des valeurs manquantes spécifiquement dans A1, A2, A4, A5, A6, A7 et A14. Ces valeurs manquantes constituent 5% de l'ensemble de données.

Attribut	Type original de valeur	Valeurs	Signification
A1	Binaire	a, b	Gender
A2	Continu	10-80	Age
A3	Continu	0-16	Debt
A4	Caractère	u, y, l, t	Married
A5	Caractère	g, p, gg	Bank Customer
A6	Caractère	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff	Education Level
A7	Caractère	v, h, bb, j, n, z, dd, ff, o	Ethnicity
A8	Continu	0-5	Years Employed
A9	Binaire	t, f	Prior Default
A10	Binaire	t, f	Employed
A11	Entier	0, 01, 02, …, 20	Credit Score
A12	Binaire	t, f	Drivers License
A13	Caractère	g, p, s	Citizen
A14	Entier	00000-00400	Zip Code
A15	Entier	0-13212	Income
A16	Binaire	+, -	Classes

2.2 Transformation des données

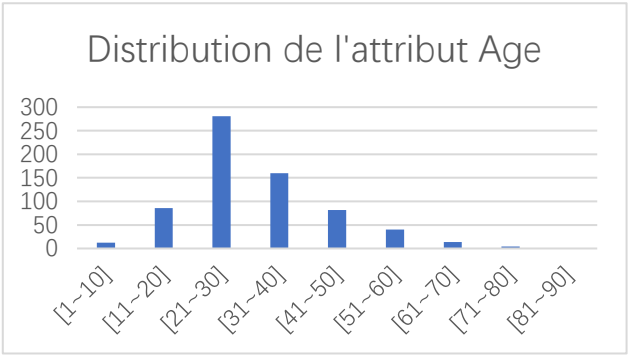
Comme mentionné, les données contiennent des valeurs catégorielles qui sont transformées en valeurs binaires ou en facteurs 1 et 0. Par exemple :

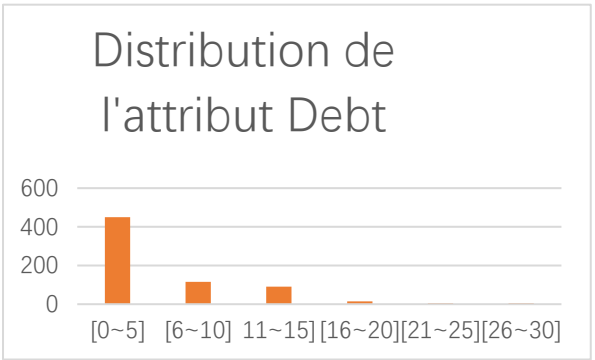
Attribut	Valeurs originales	Valeurs transformées
A1 : Gender	a, b	1, 0 (où '1' représente homme et '0' représente femme)
A9: Prior Default A10: Employed A12: Driver's License A16 : Result (Class)	t, f	1, 0 (où '1' est considéré comme vrai / oui / passe et '0' comme faux / non / échec.)

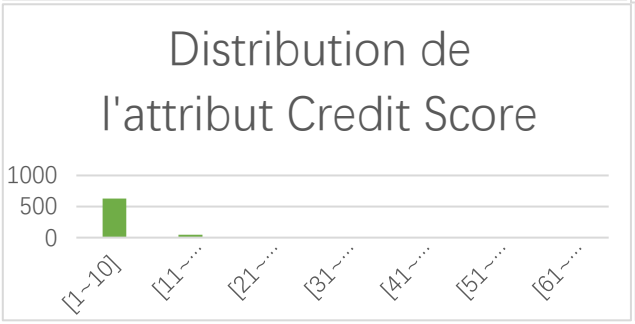
Nous manipulons de la même manière les attributs A4, A5, A6, A12 et A13 qui ont plusieurs états à transformer.

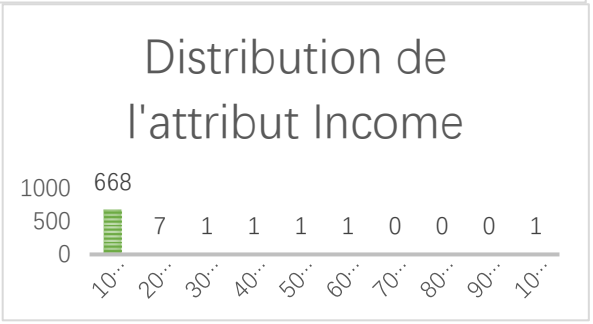
2.3 Normalisation

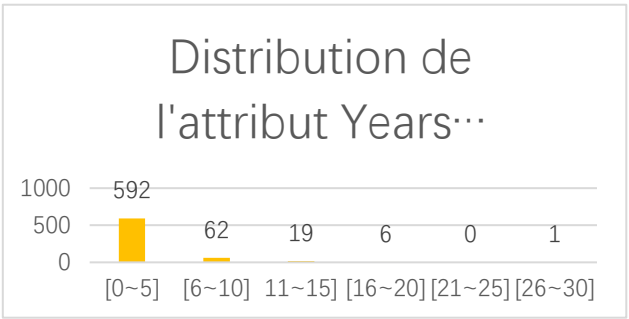
La distribution des 5 variables continues *Age, Debt, Credit score, ZipCode, Income and Years employed* a été observée initialement et affichée au-dessous :



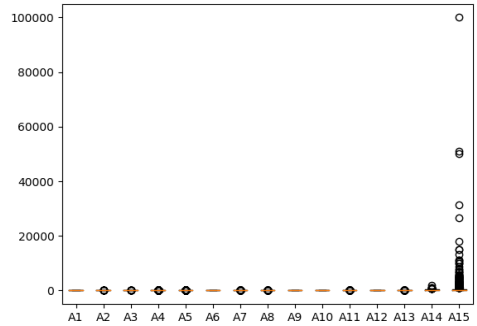


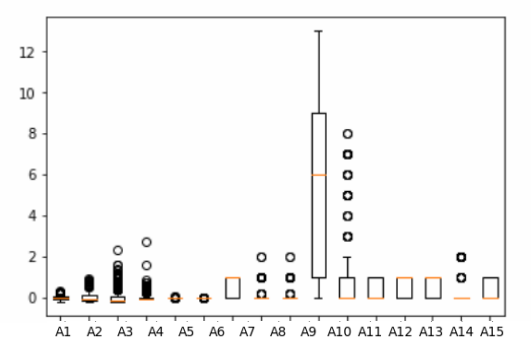






Ces diagrammes initiaux montrent que les attributs ont des distributions inclinées vers la gauche indiquant que les données ne sont pas bien réparties sur la moyenne. Cela pourrait être attribué à la population provenant d'un seul secteur économique. La normalisation a été appliquée à ces attributs pour réduire l'asymétrie. Pour une meilleure représentation graphique, les variables continues, *Years Employed*, *Credit Score*, *Age*, *Debt* et ZipCode, *Income* ont été normalisées à l'échelle uniforme. Les figures ci-dessous sont les diagrammes en boîte des données avant et après la normalisation.





La normalisation contient les étapes suivantes :

1) Soustraire l'espérance pour centraliser des données

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{N_X} \end{bmatrix} = \frac{1}{m}\sum_{i=1}^{m} X^{(i)}$$

μ est le vecteur contenant les espérances des attributs de dimension $(N_X, 1)$.

$$X := X - \mu$$

2) Normaliser la variance pour homogénéiser des données

$$\sigma^2 = \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_{N_X}^2 \end{bmatrix} = \frac{1}{m}\sum_{i=1}^{m} X^{(i)} \text{ ** } 2$$

Où **2 représente de calculer la puissance carrée pour chaque élément.

σ^2 est le vecteur contenant les variances des attributs de dimension $(N_X, 1)$.

$$X := X/\sigma^2$$

2.4 Remplissage des valeurs manquantes

La base contient des valeurs manquantes sur 7 des 16 variables (*Age, Sex, Citizen, Bank Customer, Education Level, Ethnicity et Zip Code*) qui se trouvent dans 37 instances (5% de l'ensemble). Initialement, elles ont été remplies par le mot « NA ». Parmi eux, « Age » est une variable continue.

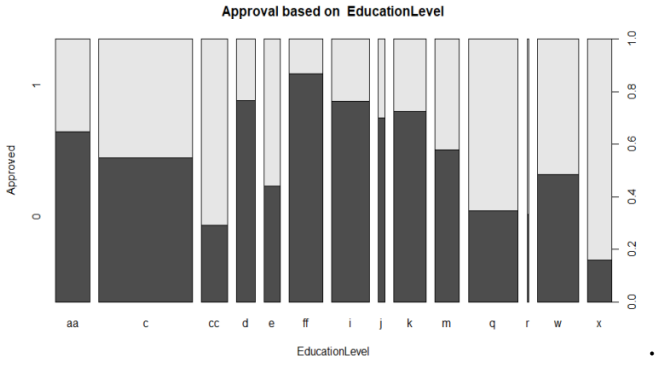
Nous programmons une fonction en utilisant la méthode Random Forest Regressor qui permet de prédire les valeurs manquantes à partir d'autres valeurs.

2.5 Visualisation des données

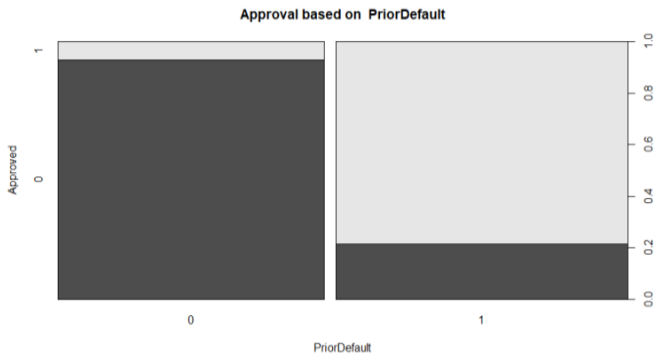
Une analyse visualisation effectuée sur l'ensemble de données peut donner une idée des relations possibles entre les attributs et observer tout effet visible de chaque attribut sur le résultat (si une demande est finalement approuvée ou non).

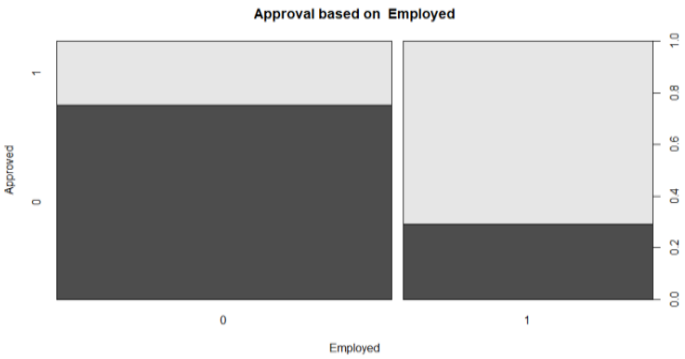
2.5.1 Effet du seul attribut sur l'approbation

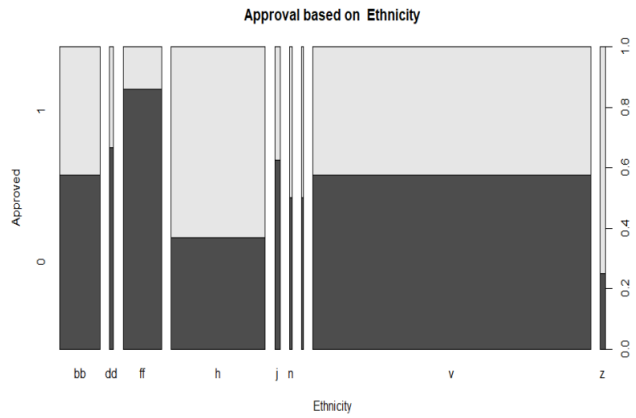
Les diagrammes des attributs discrets ont été effectués pour inspecter visuellement si la variable a influencé l'approbation de crédit. Ces diagrammes affichent le pourcentage d'approbation du crédit pour chaque catégorie.
Les attributs *Prior default* et *Employment status* semblent avoir l'effet le plus significatif sur l'approbation du crédit. Les personnes ayant un défaut antérieur sont rejetées plus de 90% du temps contre 20% pour celles qui n'en ont pas. Evidemment, les personnes employées sont acceptées environs 60% des cas



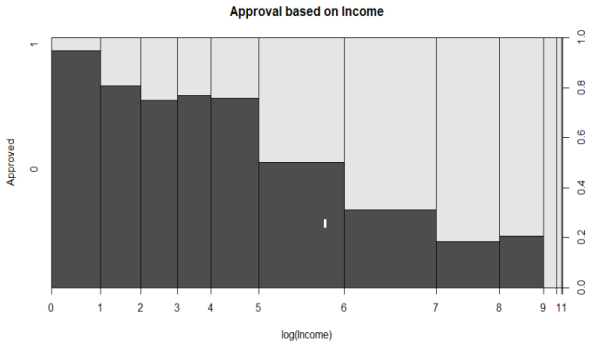
c'est prévisible, niveau d'éducation élevé = salaire élevé en moyenne. Les personnes ayant le niveau d'éducation « x » ont 85% de chances d'être approuvées tandis que celles qui ont suivi un enseignement « ff » sont rejetées dans environ 85% des cas.

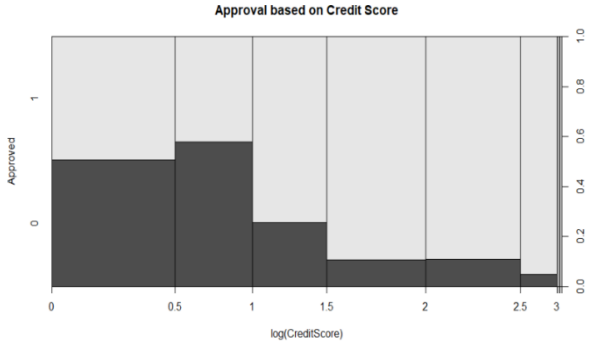






L'attribut *Ethnicity*, montre un taux d'approbation d'environ 75% pour les personnes identifiées par le groupe « z » et inversement un taux de rejet de près de 90% pour les personnes du groupe « ff ». Il serait utile d'étudier la corrélation de l'attribut avec des autres pour trouver la cause.





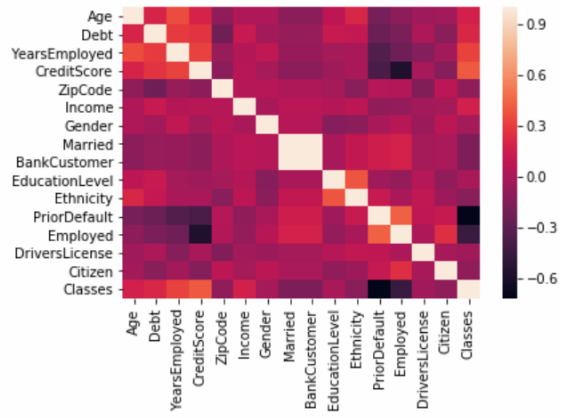
Parmi les variables continues, l'attribut *Income* et *Credit Score* semblent avoir un effet important sur le résultat :

Un pointage de crédit élevé entraîne l'approbation du crédit environ 90% du temps et les candidats ayant un revenu plus élevé ont un taux d'approbation supérieur à la moyenne.

2.5.2 Matrice de corrélation

Matrice de correlation

	Age	Debt	YearsEmployed	CreditScore	ZipCode	Income	Gender	Married	BankCustomer	EducationLevel	Ethnicity	PriorDefault	Employed	DriversLicense	Citizen	Classes
Age	1	0.2	0.36	0.19	-0.096	0.019	0.025	-0.11	-0.11	0.086	0.21	-0.2	-0.1	-0.041	-0.026	0.18
Debt	0.2	1	0.3	0.27	-0.22	0.12	-0.026	-0.069	-0.069	0.13	0.11	-0.25	-0.18	0.014	-0.12	0.21
YearsEmployed	0.36	0.3	1	0.32	-0.07	0.051	0.097	-0.066	-0.066	-0.013	0.0049	-0.35	-0.22	-0.14	-0.021	0.32
CreditScore	0.19	0.27	0.32	1	-0.11	0.064	-0.01	-0.12	-0.12	-0.041	0.0003	-0.38	-0.57	-0.0066	-0.14	0.41
ZipCode	-0.096	-0.22	-0.07	-0.11	1	0.059	0.06	0.02	0.02	-0.013	-0.12	0.055	0.045	-0.16	0.083	-0.1
Income	0.019	0.12	0.051	0.064	0.059	1	0.0031	0.069	0.069	0.052	0.084	-0.09	-0.078	-0.019	-0.012	0.18
Gender	0.025	-0.026	0.097	-0.01	0.06	0.0031	1	0.054	0.054	-0.14	-0.12	0.00038	0.061	-0.051	0.074	-0.012
Married	-0.11	-0.069	-0.066	-0.12	0.02	0.069	0.054	1	1	0.0022	0.1	0.15	0.18	-0.013	0.0083	-0.17
BankCustomer	-0.11	-0.069	-0.066	-0.12	0.02	0.069	0.054	1	1	0.0022	0.1	0.15	0.18	-0.013	0.0083	-0.17
EducationLevel	0.086	0.13	-0.013	-0.041	-0.013	0.052	-0.14	0.0022	0.0022	1	0.39	-0.038	-0.076	0.056	-0.09	0.003
Ethnicity	0.21	0.11	0.0049	0.0003	-0.12	0.084	-0.12	0.1	0.1	0.39	1	0.13	0.00088	0.097	-0.047	-0.13
PriorDefault	-0.2	-0.25	-0.35	-0.38	0.055	-0.09	0.00038	0.15	0.15	-0.038	0.13	1	0.43	0.093	0.11	-0.72
Employed	-0.1	-0.18	-0.22	-0.57	0.045	-0.078	0.061	0.18	0.18	-0.076	0.00088	0.43	1	0.019	0.24	-0.46
DriversLicense	-0.041	0.014	-0.14	-0.0066	-0.16	-0.019	-0.051	-0.013	-0.013	0.056	0.097	0.093	0.019	1	-0.0055	-0.033
Citizen	-0.026	-0.12	-0.021	-0.14	0.083	-0.012	0.074	0.0083	0.0083	-0.09	-0.047	0.11	0.24	-0.0055	1	-0.1
Classes	0.18	0.21	0.32	0.41	-0.1	0.18	-0.012	-0.17	-0.17	0.003	-0.13	-0.72	-0.46	-0.033	-0.1	1



A partir de l'analyse visuelle des attributs par paire et de la matrice de corrélation, les attributs *Prior default*, *Employed*, *Credit score* et *Years employed* ont la plus forte corrélation avec l'attribut de classe. Mais on constate aussi qu'ils sont très corrélés entre eux. Autrement dit l'information qu'ils contiennent est « similaire ». Ce phénomène influencera des résultats de la sélection de variables qu'on analysera au-dessous.

Pour être plus rigoureux, il vaut mieux utiliser la conception « correlation ration »[2] pour calculer les corrélations entre les variables quantitatives. Ici, on utilise seulement l'analyse de Fisher pour simuler et donner une idée générale.

3. Techniques de l'apprentissage supervisé

3.1 kNN (k-Nearest Neighbors)

La méthode des k plus proches voisins (kNN) est une méthode d'apprentissage supervisé basé sur les exemples, c'est une approximation locale et un apprentissage par cœur après avoir reporté tous les calculs à la classification. Il s'agit de classifier les données en calculant ses distances à partir d'une base connue. L'inconvénient de l'algorithme kNN est qu'il est très sensible à la structure locale des données.

3.2 Réseau de neurones

Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type probabiliste. Ils sont placés d'une part dans la famille des applications statistiques, qu'ils enrichissent avec un ensemble de paradigmes permettant de créer des classifications rapides, et d'autre part dans la famille des méthodes de l'intelligence artificielle auxquelles ils fournissent un mécanisme perceptif indépendant des idées propres de l'implémenteur, et fournissant des informations d'entrée au raisonnement logique formel.

Les réseaux de neurones artificiels ont besoin de massives de cas réels servant d'exemples pour leur apprentissage. Il y a des problèmes qui se traitent bien avec les réseaux de neurones, en particulier ceux de classification en domaines convexes.

3.3 Decision Tree

L'arbre de décision est un outil qui aide à la décision en représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches, et sont atteints en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés par exemple la sécurité, la fouille de données, la médecine, etc. Son avantage majeur est sa lisibilité et sa rapidité d'exécution. Il s'agit de plus des algorithmes d'apprentissage supervisé.

3.4 Discrimination linéaire : LDA

En statistique, l'analyse discriminante linéaire fait partie des techniques d'analyse discriminante prédictive. Il s'agit d'expliquer et de prédire l'appartenance d'un individu à une classe (groupe) prédéfinie à partir de ses caractéristiques mesurées à l'aide de variables prédictives.

LDA a une très large gamme d'applications dans le domaine de la reconnaissance de formes (telles que la reconnaissance faciale, l'identification des navires, etc.), plus particulièrement, il est uniquement utilisé pour l'apprentissage supervise, et souvent utilisé pour des données dépendantes de la moyenne au lieu de la variance.

Il y a aussi une autre caractéristique : il peut être utilisé pour réduire des dimensions, il peut également être utilisé comme un classificateur, mais pour réduire des dimensions est le plus souvent.

3.5 SVM (Support Vector Machine)

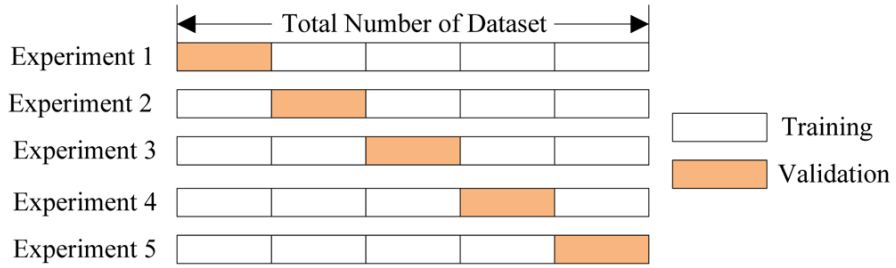
La machine à vecteur de support (SVM) est un modèle d'apprentissage supervisé qui peut résoudre des problèmes de discrimination et de régression. Il présente de nombreux avantages uniques dans la résolution de petits échantillons, de reconnaissance de formes non linéaires et de grandes dimensions. Il peut être appliqué à l'ajustement de fonction et d'autres problèmes d'apprentissage machine.

4. Analyse de la performance des techniques

Dans le processus de manipulation, on a utilisé le taux d'apprentissage et de généralisation avec leur barre d'erreur comme le critère de comparaison.

$$\tau_{\text{app}/gen} = \frac{nombre\ d'exemples\ bien\ classifié}{nombre\ total\ d'exemples}, \sigma_{\hat{\tau}}^2 = \frac{\tau(1-\tau)}{\text{m}}$$

Etant donné que le faible nombre d'instances pour l'apprentissage et pour le test, on applique **la méthode de validation croisée** pour obtenir des résultats plus fiables. Plus précisément, on divise l'échantillon original en k échantillons, puis on sélectionne un des k échantillons comme ensemble de validation et les k-1 autres échantillons constitueront l'ensemble d'apprentissage. On calcule comme dans la première méthode le score de performance, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les k-1 échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction. En pratique, on échantillonne l'ensemble en utilisant la fonction *Shufflesplit()* dans *sklearn*.



Les résultats de la prédiction des classes sont calculés par formule écrite en utilisant la fonction *predict ()* et démontrés en utilisant la matrice de confusion afin de déterminer la précision de la classification.

4.1 Résultats sans validation croisée

La base d'apprentissage : la base de généralisation = 1 : 4.

	kNN	Réseau de neurones	Decision tree	LDA	SVM
τ_{app}	0.894736	0.820326	1.0	0.695161	0.849364
$\sigma^2_{\widehat{\tau_{\text{app}}}}$	0.013074	0.016355	0.0	0.058463	0.015238
τ_{gen}	0.862319	0.833333	0.8188405	0.710144	0.876811
$\sigma^2_{\widehat{\tau_{gen}}}$	0.029331	0.031724	0.0327861	0.172718	0.027976

4.2 Résultats avec la validation croisée

La base d'apprentissage : la base de généralisation = 1 : 4.

	kNN	Réseau de neurones	Decision tree	LDA	SVM
τ_{app}	0.898729	0.831760	1.0	0.803709	0.854990
$\sigma^2_{\widehat{\tau_{\text{app}}}}$	0.012839	0.015926	0.0	0.050443	0.014994
τ_{gen}	0.805072	0.800724	0.8217391	0.797101	0.854347
$\sigma^2_{\widehat{\tau_{gen}}}$	0.033579	0.033737	0.0323713	0.153098	0.029854

Les taux de généralisation et les barres d'erreurs obtenues avec ou sans validation croisée sont assez proches pour le SVM et les arbres de décisions. Par contre, on note des différences significatives pour les autres méthodes. Dans la suite, nous utiliserons la validation croisée pour cette méthode.

4.3 Matrices de confusion (le test)

La matrice de confusion permet de visualiser la performance des méthodes pour chaque classe. La barre d'erreur des taux de précision est quasiment égale aux valeurs au-dessus.

Prédictions\Réel	+	-
+	0.7958	0.1397
-	0.2042	0.8603

Knn

Prédictions\Réel	+	-
+	0.7864	0.1044
-	0.2136	0.8956

Réseau de neurones

Prédictions\Réel	+	-
+	0.8517	0.1957
-	0.1483	0.8043

Decision Tree

Prédictions\Réel	+	-
+	0.7378	0.1070
-	0.2622	0.8929

LDA

Prédictions\Réel	+	-
+	0.9121	0.2141
-	0.0879	0.7859

SVM

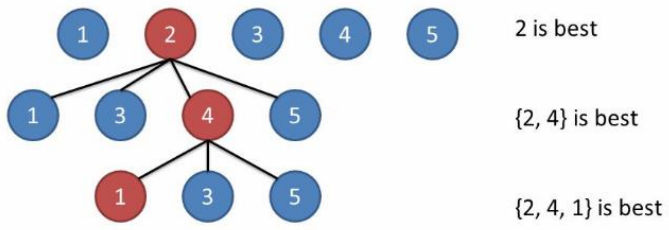
4.4 Conclusion

À partir des résultats obtenus, la méthode SVM surpasse tous les autres classificateurs avec une précision maximale de 85.43%. La decision tree, et kNN fonctionnent de façon satisfaisante, mais ils nécessitent un test persistant des codes pour trouver des valeurs optimales des hyperparamètres. D'un autre côté, le réseau de neurones et la LDA a une précision relativement faible.

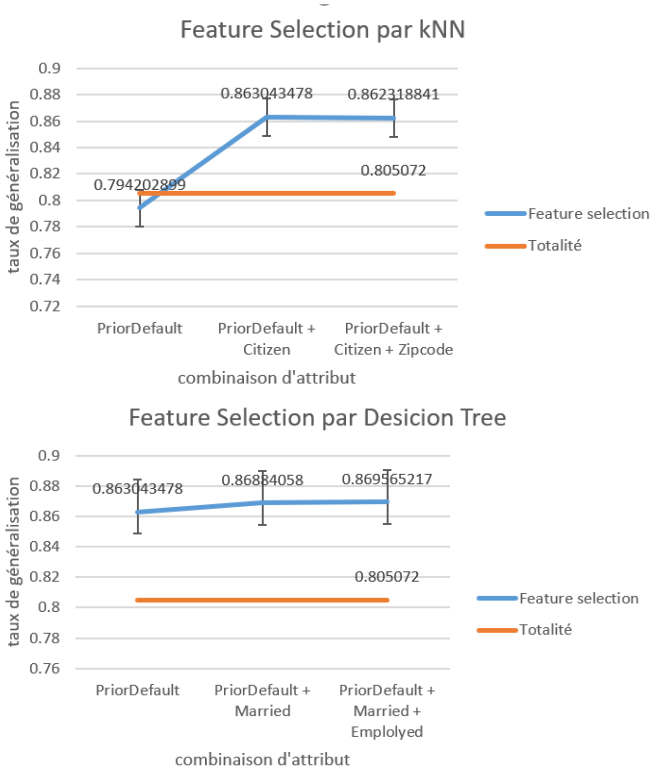
En analysant les matrices de confusion, on conclut que la plupart des algorithmes est quasiment autant performante pour classifier les deux classes, sauf SVM, qui classifie des de la classe positif (+) plus précisément. On analysera la cause du résultat après effecteur la feature selection avec SVM.

5. Approfondissement : Forward feature selection

La forward feature selection commence par l'évaluation de tous les sous-ensembles d'variables qui ne contiennent qu'une seule variable d'entrée. On garde l'attribut qui donne la meilleure performance. La sélection trouve ensuite le meilleur sous-ensemble constitué de deux attributs. Les sous-ensembles d'entrée avec trois, quatre et plus attributs d'entrée sont évalués de la même manière.

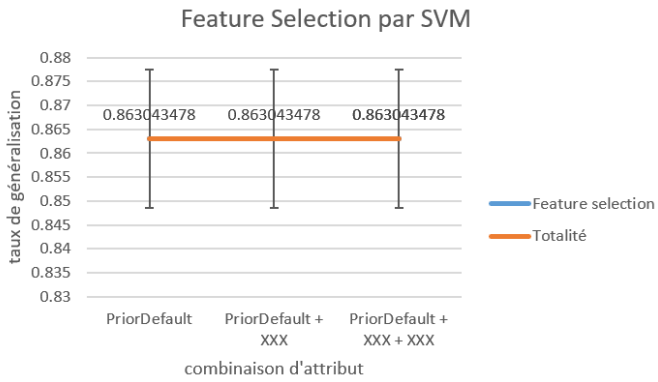


On effectue cette sélection avec plusieurs méthodes : kNN, l'arbre de décision, le réseau de neurones et SVM. L'évaluation de la performance consiste à calculer le taux de généralisation de la validation croisée avec 10 échantillon. Pour mieux comparer les résultats, on échantillonne l'ensemble en utilisant la fonction *Shufflesplit()* dans *sklearn* et fixe que *random_state =5*.



On a vu dans l'analyse de la matrice de corrélation que les attributs *Prior default, Employed, Credit score* et *Years employed* sont les plus corrélés avec la classification. Ils se présentent naturellement dans les choix d'attributs les plus performants. Mais ils contiennent des informations similaires qui pousse des algorithmes à choisir une ou deux variables entre eux et les couple avec une autre variable moins corrélée comme *Married, Citizen,* ou *Zip code* pour capturer plus d'information.

Lors de la manipulation avec SVM, on a constaté un phénomène intéressant que les taux de généralisation des différentes combinaisons d'attribut sont le même, qui est égale aussi le taux avec la totalité des attributs. C'est-à-dire que l'algorithme SVM utilise seulement l'attribut *Prior default* pour classifier. En fonction du résultat de l'analyse visuelle, l'attribut *Prior default* est plus « sensible » à la classification de la classe *positif* (+) car les de cette classe est moins dispersé dans les deux valeurs de cet attribut.



6. Conclusion

Parmi les modèles conçus pour prédire les résultats des demandes de crédit, SVM et l'arbre de décision ont fourni les résultats les plus précis. Pour améliorer, on pourra inclure la combinaison de deux techniques ou plus pour produire un modèle de classification avec un taux de précision plus élevé. Utilisé par les créanciers, un modèle amélioré peut réduire considérablement le risque d'octroi de crédit aux défaillants potentiels.

D'après notre analyse des attributs, nous sommes en mesure de conclure que ceux qui sont les plus importants pour déterminer le résultat d'une demande de crédit sont le revenu(*Income*), les années d'emploi (*Years employed*), la cote de crédit (*Credit score*) et si le demandeur a fait défaut sur un compte de crédit antérieur (*Prior default)*. Ce résultat était presque cohérent sur toutes les techniques appliquées : la visualisation, l'analyse de corrélation, l'apprentissage supervisé, et la forward feature selection. Corrélativement, les autres variables de l'ensemble de données : l'âge (*Age*), la dette(*Debt*), la citoyenneté (*Citizen*), l'état civil (*Married*), Sexe (*Gender*), le niveau de scolarité (*Education level*), l'ethnicité (*Ethnicity*), le code postal (*Zip code*), le permis de conduire (*Drivers licence*) et la situation du client bancaire (*Bank customer*) n'ont pas eu d'effet significatif sur l'approbation d'une demande.

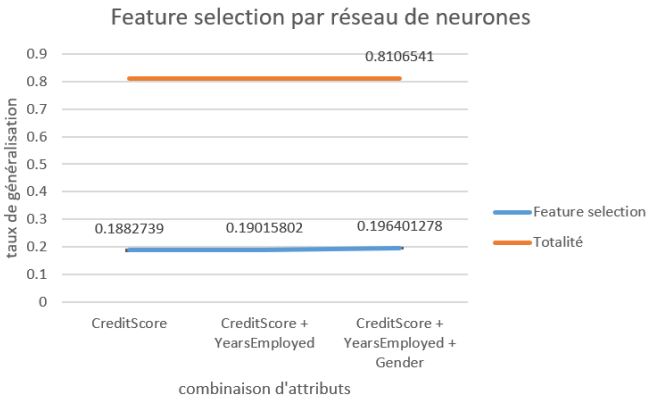
Un revenu plus élevé augmente les chances d'un demandeur d'être approuvé pour le crédit. Le niveau de scolarité, bien que n'étant pas un prédicteur conventionnel de l'approbation du crédit, montre une certaine corrélation avec le taux d'approbation. Les candidats ayant un niveau d'éducation ont des taux d'approbation beaucoup plus élevés que les candidats ayant d'autres niveaux d'éducation. Une explication possible pourrait être qu'avec un niveau d'éducation plus haut, une personne aura un revenu plus élevé, augmentant ainsi les chances d'approbation.

Même si l'appartenance ethnique est un attribut protégé et ne tient pas compte de la décision d'approbation de crédit, les candidats s'identifiant à une ethnie ont reçu un taux d'approbation considérablement élevé tandis que les autres ont eu un taux de refus plus élevé que la moyenne.

7. Parties de l'analyse à améliorer

Pendant le projet, on a rencontré plusieurs difficultés :

1. Il existe un mélange des attributs quantitatives et catégoriques qui pose des problèmes de calcul de la corrélation.
2. Dans le prétraitement de la base de données, notre moyen de la numérisation des attributs catégoriques peut entraîner de la mauvaise performance des méthodes qui consistent à classifier des instances par comparer leurs distances avec les deux classes, comme kNN et LDA.
3. Faute d'exemples, la précision du réseau de neurones est moins élevée que prévu.
4. Lorsqu'on effectue la sélection de variables avec le réseau de neurones, on a rencontré un problème insoluble : la performance des différentes combinaisons des attributs sont toutes anormales.



8. Remerciement

Nous tenons à remercier notre tuteur du projet, M. DIONG, pour nous' avoir accordé sa confiance et pour tout l'aide qui nous' a apporté tout au long du projet, afin de nous permettre de mieux accomplir le travail et l'étude.

9. Reference

[1] Credit Approval Data Set, submitted by quinlan '@' cs.su.oz.au, http://archive.ics.uci.edu/ml/datasets/Credit+Approval
[2] Correlation ratio, Wikipédia, https://en.wikipedia.org/wiki/Correlation_ratio
[3] An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
[4] The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition), Trevor Hastie, Robert Tibshirani, Jerome Friedman
[5] Credit Approval Analysis using R, Deepesh Khaneja, November 2017
[6] Réseau de neurones artificiels, Wikipédia, https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels#Utilit%C3%A9
[7] k-nearest neighbors algorithm, Wikipédia, https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[8] Decision tree, Wikipédia, https://en.wikipedia.org/wiki/Decision_tree
[9] Linear discriminant analysis, Wikipédia, https://en.wikipedia.org/wiki/Linear_discriminant_analysis
[10] Support vector machine, Wikipédia, https://en.wikipedia.org/wiki/Support_vector_machine