2022 – 2023

Le Rapport RCP209

Matar Doky DIOP

Sommaire

1 Objectif

La décision d'accorder ou non un crédit à un individu est un enjeu crucial pour les banques. L'objectif de cette étude est de classer des individus en deux classes (bon ou mauvais crédit risqué) à partir d'un ensemble de variables décrivant les individus.

2 Analyse Exploratoire des données

Le dataset original préparé par le Dr. Hofmann contient 1000 instances avec 20 variables (13 catégorielles et 7 numériques) et se trouve dans le fichier german.data. Chaque instance représente un individu qui contracte un crédit auprès d'une banque. Chaque individu est classé comme bon ou mauvais risque de crédit selon l'ensemble des variables. Nous allons à présent traiter, visualiser, et analyser le dataset afin de mieux l'appréhender au maximum.

2.1 Etude de la target

Le classement de bon ou de mauvais risque de crédit de chaque individu selon l'ensemble des variables du dataset est appelé target. La figure 1 présente les modalités de la target.

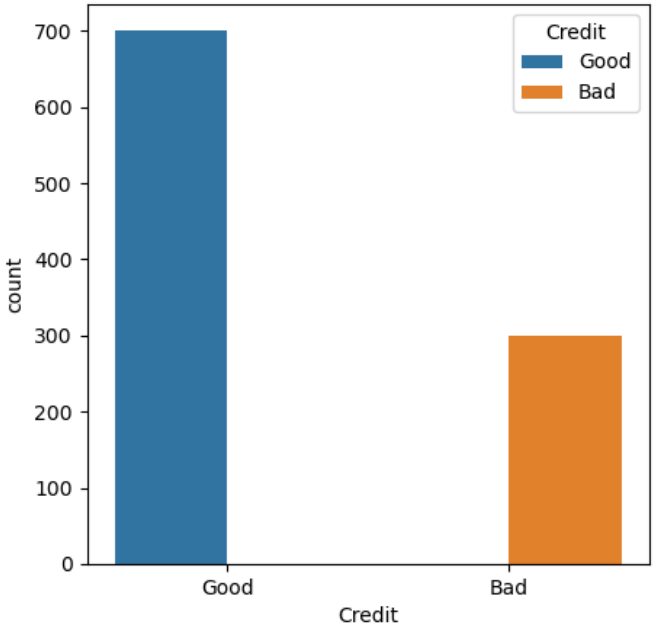

FIGURE 1 – Target

Nous avons observé un taux de 70% de bon risque de crédit et un taux de 30% de mauvais risque de crédit. Nous avons donc un déséquilibre entre les classes.

2.2 Etude des variables quantitatives

Les figures 2, 3, et 4 représentent les distributions de quelques variables quantitatives.

Nous avons observé que les distributions du montant d'un crédit et de l'âge des individus sont positivement asymétriques (portion de la boite à droite et la moustache droite sont plus longues qu'à gauche de la

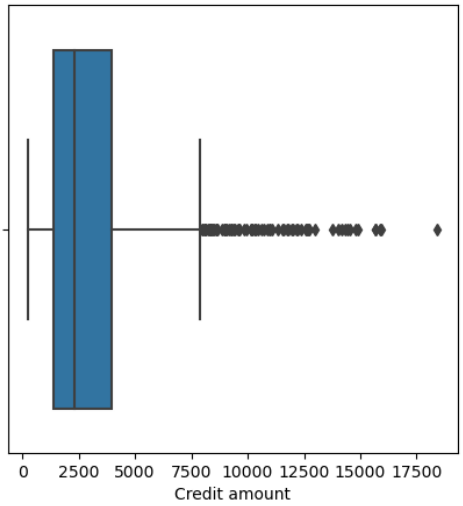

FIGURE 2 – Montant du crédit

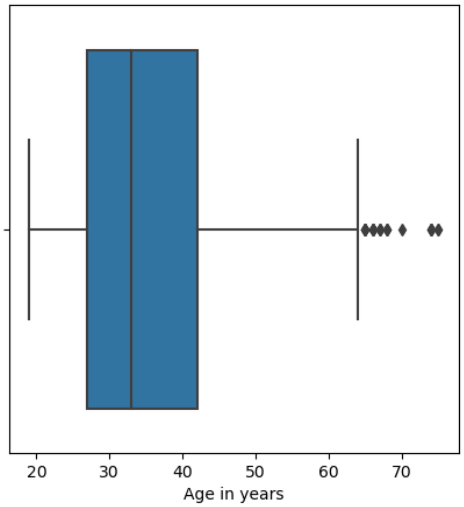

FIGURE 3 – Age en mois

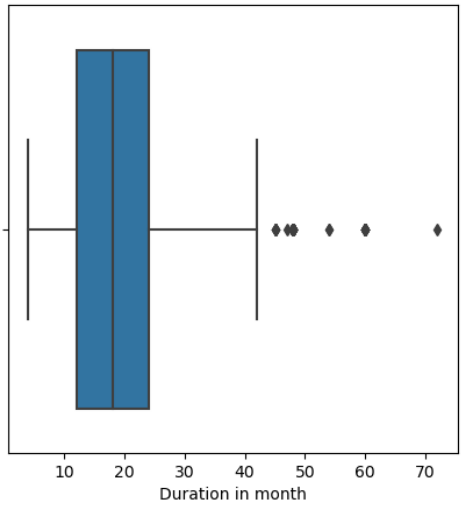

FIGURE 4 – Durée en mois

médiane), le montant moyen d'un crédit emprunté est de 3271 DM , avec une médiane de 2320 DM, la moyenne d'âge des individus est de 36 ans, avec une médiane de 33 ans, la durée moyenne d'un crédit s'établit à 20 mois avec une médiane de 18 mois.

Nous allons maintenant analyser la liaison entre ces variables quantitatives et la target à partir des figures 5, 6, et 7.

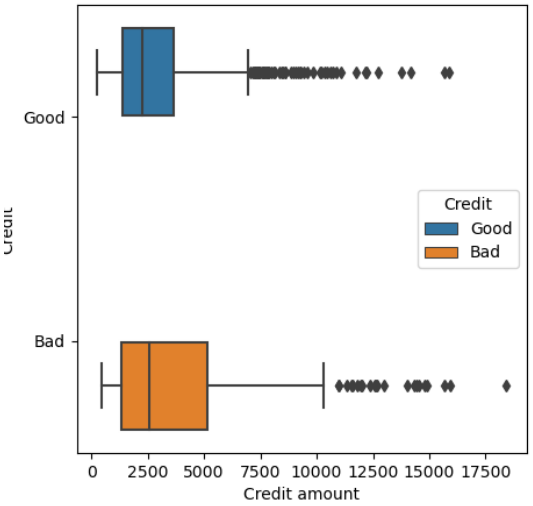

FIGURE 5 – Montant du crédit

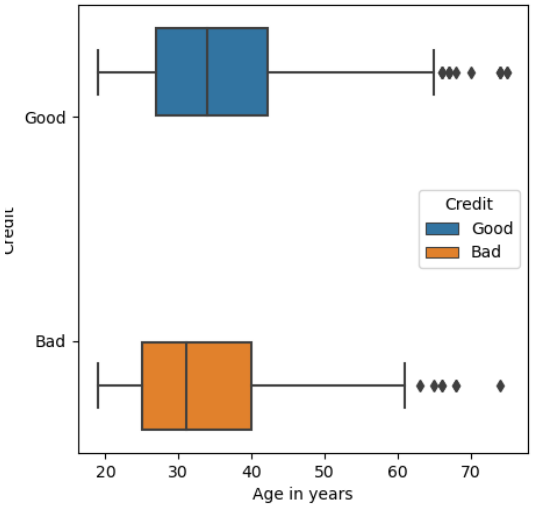

FIGURE 6 – Age en mois

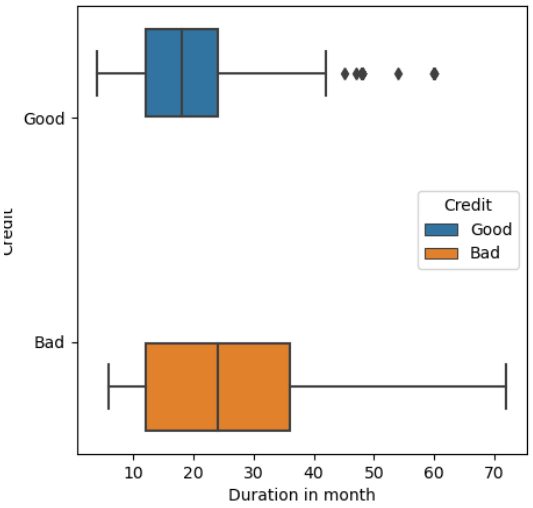

FIGURE 7 – Durée en mois

Nous avons observé une différence significative sur la durée moyenne d'un crédit entre les deux classes (bon ou mauvais risque de crédit). La durée moyenne d'un crédit est plus importante chez les individus présentant un mauvais risque de crédit. On peut admettre que la durée d'un crédit augmente avec les individus présentant un mauvais risque de crédit. Par contre, le montant moyen d'un crédit emprunté et la moyenne d'âge des individus ne varient pas significativement d'une classe à l'autre.

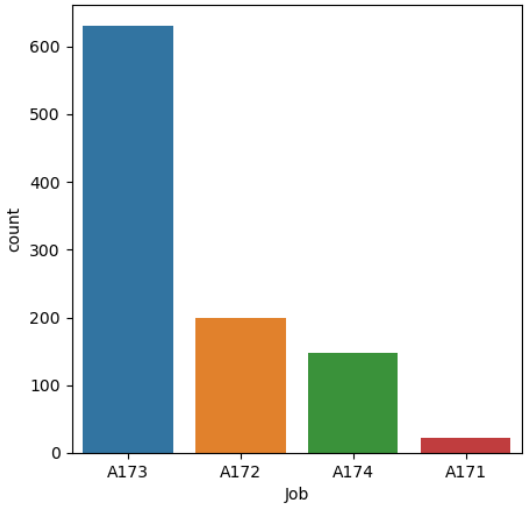

FIGURE 8 – Job

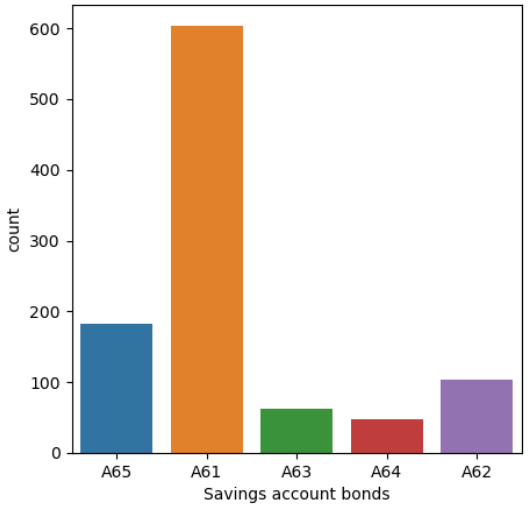

FIGURE 9 – Savings account

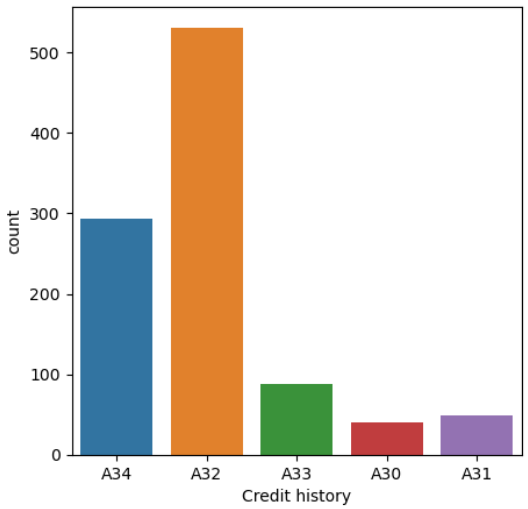

FIGURE 10 – Credit history

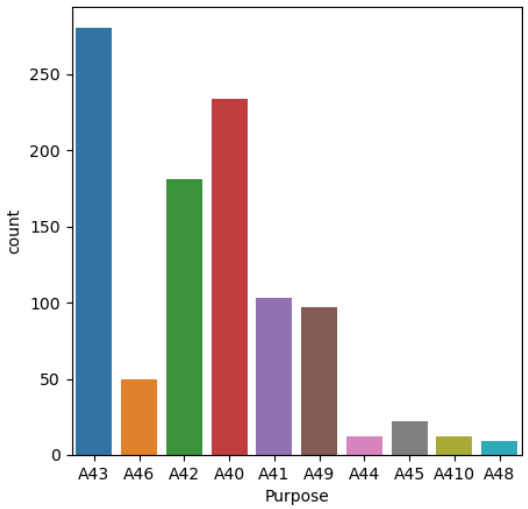

FIGURE 11 – Purpose

2.3 Etude des variables qualitatives

Les figures 8, 9, 10, et 11 représentent les modalités de quelques variables qualitatives. Nous avons parmi ces variables :

- Job : A171 : unemployed/ unskilled - non-resident, A172 : unskilled - resident, A173 : skilled employee / official, A174 : management/ self-employed/ highly qualified employee/ officer ,
- Savings account/bonds : A61 : ... < 100 DM, A62 : 100 <= ... < 500 DM, A63 : 500 <= ... < 1000 DM, A64 : .. >= 1000 DM, A65 : unknown/ no savings account,
- Credit history : A30 : no credits taken/all credits paid back duly, A31 : all credits at bank paid back duly, A32 : existing credits paid back duly till now, A33 : delay in paying off in the past, A34 : critical account/ other credits existing (not at this bank) ,
- Purpose : A40 : car (new), A41 : car (used), A42 : furniture/equipment, A43 : radio/television, A44 : domestic appliances, A45 : repairs, A46 : education, A47 : (vacation - does not exist ?)

Nous avons observé que 63% des individus sont des employés qualifiés, des fonctionnaires (A173) contre

20% de ceux qui sont non qualifiés, 60 % des individus ont un compte d'épargne (A61) dont le montant est inférieur à 100 DM contre 18% de ceux qui n'ont pas de comptes d'épargne, 53% des individus ont des crédits existants dûment remboursés jusqu'à maintenant (A32) contre 29% de ceux qui sont en difficultés de remboursement, 58% des individus sont des hommes célibataires contre 31% de ceux qui sont des femmes divorcées/mariées/séparées.

2.4 Analyse en composantes principales (ACP)

Afin de mettre en évidence les relations existantes entre les variables ou entre groupes de variables, les ressemblances entre les individus, de visualiser en faible dimension les données, on réalise une ACP. L'ACP n'étant valide que pour les variables quantitatives, les variables qualitatives sont considérées comme des variables supplémentaires pour les aides à l'interprétation.

Les figures 12, et 13 représentent la projection des individus sur le premier et deuxième plan factoriel.

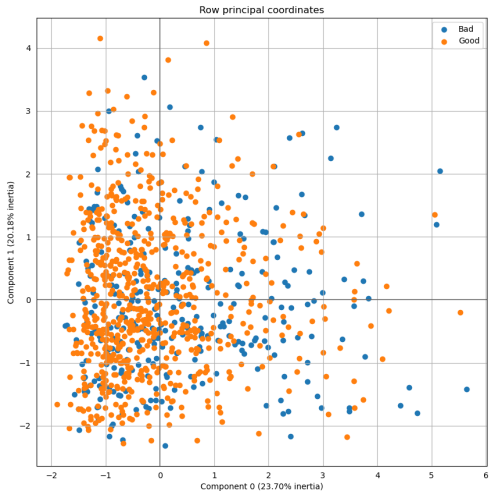

FIGURE 12 – Premier plan factoriel

FIGURE 13 – Deuxième plan factoriel

D'abord, nous avons constaté que le premier axe principal contribue à environ 23% de l'inertie totale (pourcentage d'information expliqué par un axe). Les deuxième, troisième et quatrième axes contribuent respectivement à 20,18%, 15,97% et 13,42% de l'inertie totale. Ainsi le plan factoriel formé par les deux premiers axes explique 43,88% de l'inertie totale contre 29,39% pour le plan factoriel formé par les deuxième et troisième axes.

Nous constatons aussi que les inerties des axes qui forment les plans factoriels ne sont pas significativement différentes, ce qui indique que le nuage des individus a une forme relativement sphérique. Ces plans factoriels n'ont donc pas de pertinence particulière pour décrire les ressemblances entre les individus. Ainsi, nous pouvons en déduire que la projection des variables sur ces deux plans factoriels n'aura pas de pertinence pour évaluer les relations entre les variables. Cependant, la figure 14 représente la matrice de corrélation entre les variables.
Nous n'avons pas observé une forte relation de corrélation entre les variables. Ainsi, nous pouvons donc considérer que toutes les variables sont importantes pour la mise en place d'un modèle décisionnel.

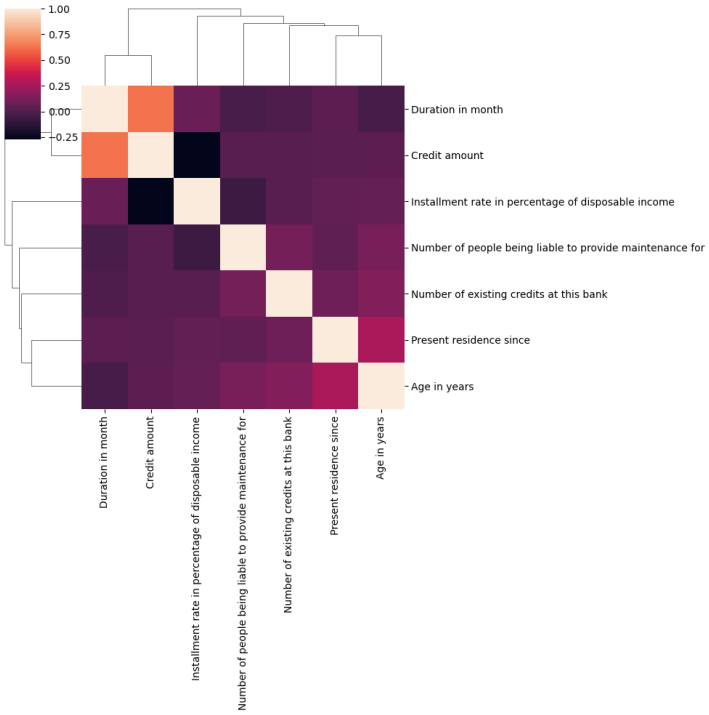

FIGURE 14 – Matrice de corrélation

3 Modèle décisionnel

3.1 Choix des modèles décisionnels

L'objectif d'un modèle décisionnel est de prédire si un individu présente un bon ou un mauvais risque de crédit.
L'analyse exploratoire a montré un déséquilibre entre les classes (bon ou mauvais risque de crédit). Ainsi, nous avons choisi les Support Vector Machines (SVM) et les forêts aléatoires qui sont des méthodes de modélisation moins sensibles au déséquilibre entre les classes.
En effet, les SVM cherchent à maximiser une marge entre l'hyperplan séparateur et les exemples d'apprentissages de chacune des classes. En conséquence, seuls comptent les vecteurs de support, ce qui permet de réduire l'impact du déséquilibre entre les classes. Concernant l'approche de la méthode des forêts aléatoires, elle est basée sur un ensemble de modèles et chaque modèle est construit sur un échantillon des observations, ce qui permet de réduire aussi l'impact du déséquilibre entre les classes.

3.2 Pré-traitement

Afin d'évaluer la performance d'un modèle, il faut le soumettre à des données qu'il n'a jamais vues. Ainsi, nous avons décomposé notre dataset entre 80% de trainset et 20% de testset. Le trainset est utilisé pour entraîner les modèles et le testset permet de mesurer la performance des modèles.

Les algorithmes d'apprentissage automatique prennent des valeurs numériques en entrée. Ainsi, nous avons opté pour la technique ordinal encoding pour encoder les variables qualitatives.

Afin de réduire la complexité du modèle SVM, nous avons normalisé les données contrairement aux arbres de décision qui n'ont pas besoin de la normalisation des données.

3.3 Optimisation des hyperparamètres

Il existe plusieurs hyperparamètres dans les algorithmes d'apprentissage automatique. Nous avons opté pour l'optimiseur GridSearchCV qui permet de trouver le modèle avec les meilleurs hyperparamètres en comparant les différentes performances de chaque combinaison grâce à la technique de cross validation. La cross validation consiste à entraîner puis à valider un modèle sur plusieurs découpes possibles du trainset.

3.4 Transformation des données

L'analyse exploratoire à montré que le nuage des individus a une forme relativement sphérique, ce qui rend complexe la tâche de classification. Nous avons donc utilisé la classe PolynomialFeatures de Sklearn pour transformer les données, en ajoutant les carrées des variables aux variables existantes.

3.5 Sélection de variables

Notre dataset contient 20 variables dont 7 variables quantitatives et 13 variables qualitatives. Certaines de ces variables peuvent n'apporter aucune information prédictive et leur prise en compte peut ainsi agir comme du « bruit » dans le processus de modélisation. Ainsi, nous avons la méthode SelectKBest de Sklearn qui permet de sélectionner les K variables dont le score de test de dépendance avec la target est le plus élevé.

3.6 Evaluation

Afin d'évaluer la qualité des méthodes de classification, nous avons utilisé plusieurs mesures de performance :

- Recall : qui permet de réduire au maximum le nombre de faux négatifs
- Précision : qui permet de réduire au maximum le nombre le taux de faux positifs
- F1-score définit par le rapport entre la précision et le recall
- la matrice de confusion qui montre les erreurs de classement

3.7 Résultats

La table 1 présente les paramètres des modèles et la sélection des K variables dont le score de test de dépendance avec la target est le plus élevé. Ces valeurs sont obtenues par l'optimiseur GridSearchCV.

	Paramètres	SelectKBest
RandomForest	$max\_depth = 8$, $max\_features = 5$, $n\_estimators = 67$	11
SVM	–	10

TABLE 1 – params

Les figures 15, et 16 présentent les matrices de confusion des modèles. Nous avons obtenu respectivement pour les modèles RadomForest et SVM une erreur de classement de 18% et 17% avec la métrique Recall chez les individus présentant un mauvais risque de crédit.

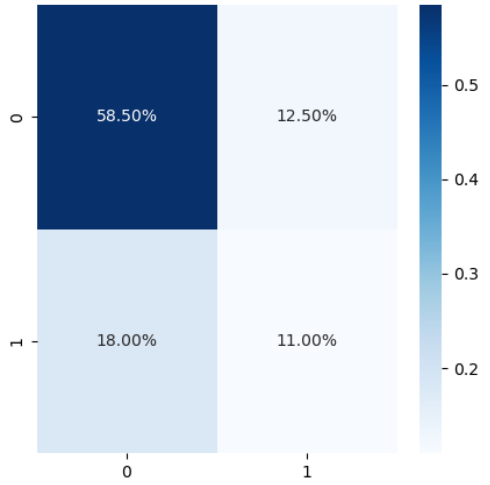

FIGURE 15 – Matrice de confusion RandomForest

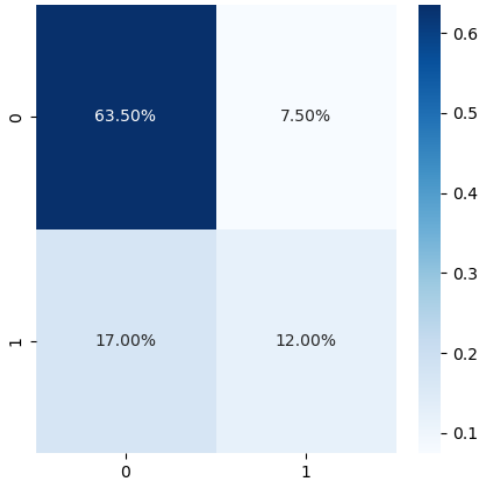

FIGURE 16 – Matrice de confusion SVM

Les tables 2, et 3 donnent les rapports de classification des modèles. Nous avons obtenu respectivement pour les modèles RadomForest et SVM un Recall de 34% et 36% chez les individus présentant un mauvais risque de crédit.

	Précision	Recall	F1-score	Support
bon risque	0.74	0.77	0.78	142
mauvais risque	0.38	0.34	0.36	58

TABLE 2 – Rapport de classification RandomForest

	Précision	Recall	F1-score	Support
bon risque	0.76	0.83	0.79	142
mauvais risque	0.47	0.36	0.41	58

TABLE 3 – Rapport de classification SVM

Les figures 17, et 18 présentent les courbes d'apprentissage des modèles qui montrent l'amélioration des performances des modèles en fonction de la quantité des données.

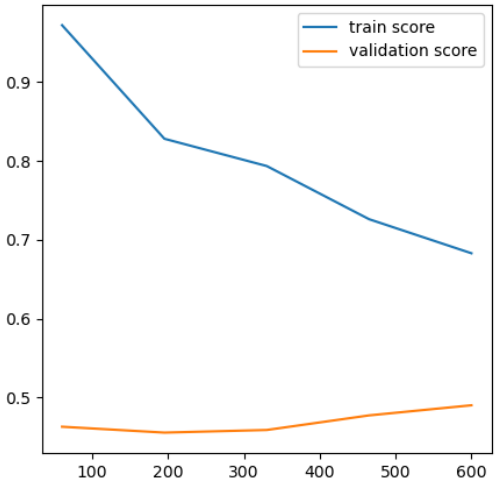

FIGURE 17 – Courbes d'apprentissage RandomForest

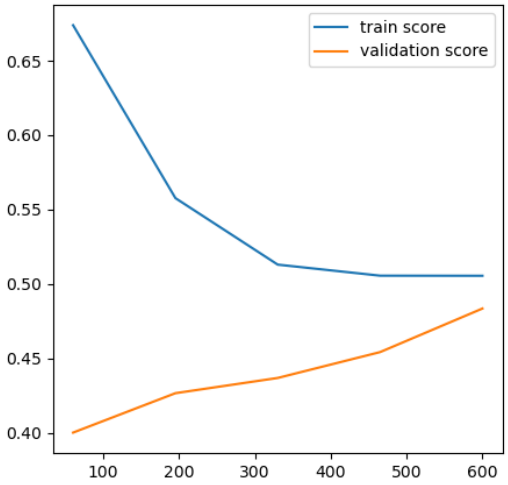

FIGURE 18 – Courbes d'apprentissage SVM

4 Conclusion

L'objectif de cette étude est de classer des individus selon un bon ou mauvais risque de crédit. L'analyse exploratoire des données a montré un déséquilibre entre les classes, ce qui nous a conduit à choisir les modèles décisionnels SVM et forêts aléatoires. Ces modèles permettent de réduire l'impact du déséquilibre entre les classes. Les résultats de ces modèles ont montré en moyenne des erreurs de classement de l'ordre de 40% avec la métrique Recall, ce qui est loin d'être optimal. Cependant, les courbes d'apprentissage ont mis en évidence les performances des modèles en fonction de la quantité des données. Plus nous avons de données, plus les performances des modèles augmentent. Ainsi, pour améliorer les résultats, nous pourrons envisager de générer des observations synthétiques pour la classe minoritaire afin d'équilibrer les proportions des classes. Les algorithmes de type SMOTE (Synthetic Minority Over-sampling TEchnique) permettent de générer de telles observations afin de rééquilibrer les données.