المعهد الوطني للبريد والمواصلات
ⵓⵔⴰⵔⵓ ⵏⵓⵎⵓⵔⵓ ⵏ ⵜⵓⵍⵓⵔⵜ ⴰ ⵜⵎⵔⵓⵍⵓⵎ
Institut National des Postes et Télécommunications

AMOA - INE2

Rapport de Projet de l'Analyse de Données Avancée et Scoring

Prédiction des difficultés financières des clients dans les deux années à venir

---

Sous l'encadrement de : Pr.Zineb El Akkaoui

Année universitaire: 2021/2022

Realisé par:
- EL ALAOUI Naoufal
- Ouberkni Hicham
- Bourquouquou Haitam

Contents

1 Introduction

Le crédit scoring est généralement considéré comme une méthode d'évaluation utilisée par les organismes bancaires pour estimer le risque de défaut et mesurer la solvabilité de chaque entreprise et lui classer soit comme une entreprise saine ou une entreprise défaillante. Il impose d'utiliser les différentes techniques statistiques en vue d'obtenir un modèle de scoring basé sur les caractéristiques de l'emprunteur.

Dans ce projet, l'objectif est de construire des modèles de prévision de score, permettant de prévoir si un client aurait des difficultés financières dans les deux années à venir. Les modèles à construire doivent utiliser les informations comme le salaire moyen, l'âge, le taux d'intérêt, etc...

Les données existent dans les fichiers .csv joints avec ce document : ScoringTraining, ScoringTest et SampleScoring et la description des variables prédictives introduites dans le fichier Data Dictionary. Ces données permettront d'apprendre et de valider les modèles de prévision de score construits, ainsi que d'en évaluer la performance

2 Préparation des données

2.1 Description des variables

- **SeriousDlqin2yrs :** La variable **cible** qui est une variable binaire pour savoir si la personne concernée a connu un retard crédit de 90 jours ou pire
- **RevolvingUtilizationOfUnsecuredLines :** Solde total sur les cartes de crédit et les marges de crédit personnelles sauf les dettes immobilières et aucune dette à tempérament comme les prêts automobiles divisé par la somme des limites de crédit
- **age :** Âge de l'emprunteur en années
- **NumberOfTime30-59DaysPastDueNotWorse :** Nombre de fois où l'emprunteur a été en souffrance depuis 30 à 59 jours, mais pas pire au cours des 2 dernières années.
- **DebtRatio :** Paiements mensuels de la dette, pension alimentaire, frais de subsistance divisés par le revenu brut mensuel
- **MonthlyIncome :** Revenu mensuel
- **NumberOfOpenCreditLinesAndLoans :** Nombre de prêts ouverts (à tempérament comme un prêt automobile ou une hypothèque) et de marges de crédit (par exemple, des cartes de crédit)
- **NumberOfTimes90DaysLate :** Nombre de fois où l'emprunteur est en souffrance depuis 90 jours ou plus.
- **NumberRealEstateLoansOrLines :** Nombre de prêts hypothécaires et immobiliers, y compris les marges de crédit sur valeur domiciliaire

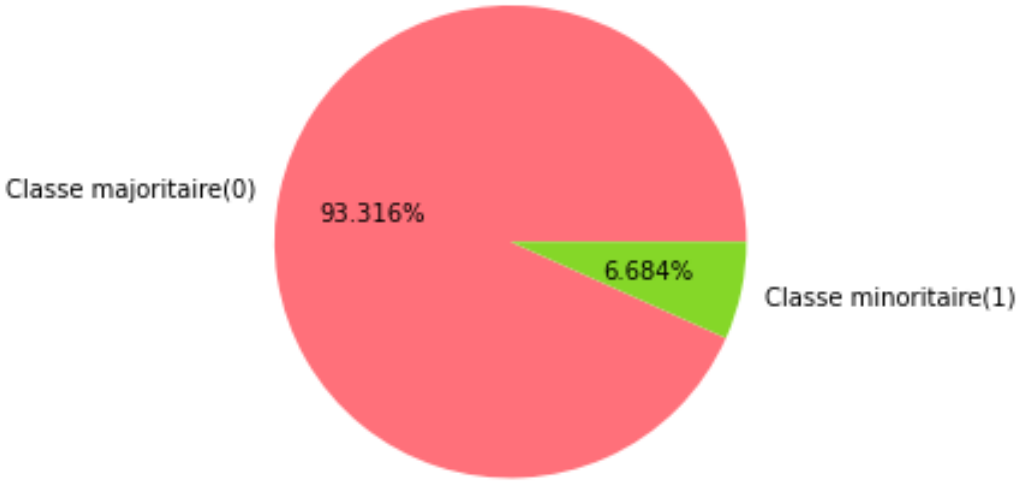
- **NumberOfTime60-89DaysPastDueNotWorse :** Le nombre de fois où l'emprunteur a été en souffrance depuis 60 à 89 jours, mais pas pire au cours des 2 dernières années.
- **NumberOfDependents :** Le nombre de personnes à charge dans la famille hors eux-mêmes (conjoint, enfants, etc.)

3 Phase de prétraitement

La préparation du jeu de données qui représente convenablement le problème étudié est une étape cruciale avant la construction du modèle de scoring. L'efficacité du modèle est, en effet, fortement liée à l'échantillon de données utilisé.

3.1 Proportion de défauts

On va tracer le diagramme de camembert de la variable cible **SeriousDlqin2yrs** pour retrouver la proportion de défauts qui est la le pourcentage de la moyenne de cette variable.

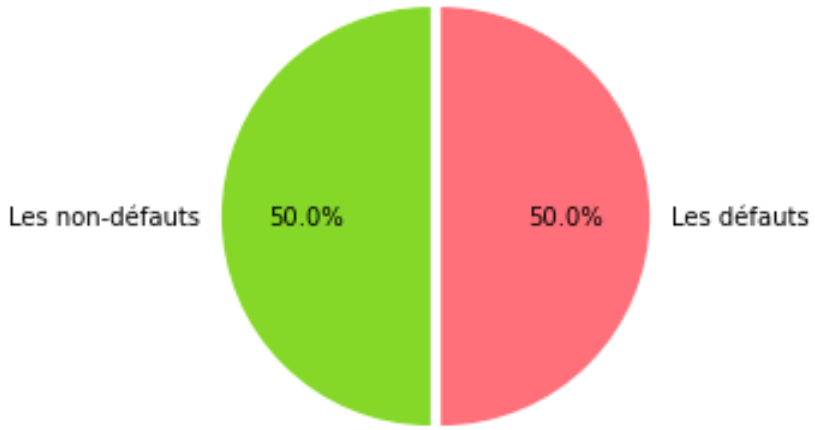


Ici, la proportion de défauts est égale à **6,684 %** .

• **Equilibrage des données d'apprentissage :**

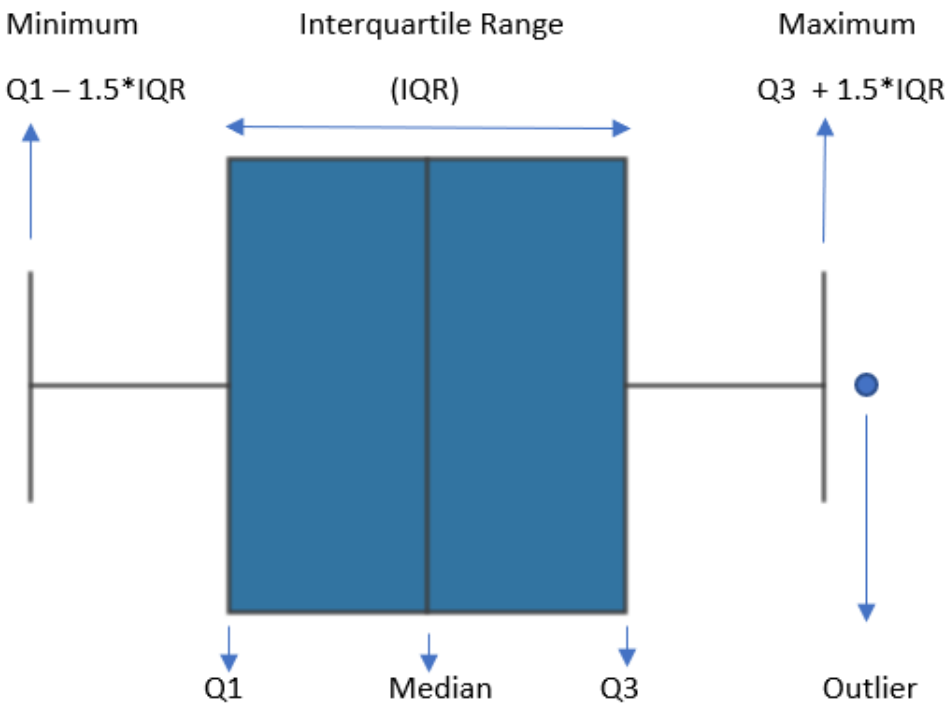
Afin d'equilibrer les données d'apprentissage on va utiliser une des méthodes de rééchantillonnage qui est **le sous-échantillonnage(Down-sampling)** qui fonctionne en diminuant le nombre d'observations de la classe majoritaire afin d'arriver à une classe majoritaire satisfaisante. Après le rééchantillonnage, le diagramme en camembert de la variable cible **SeriousDlqin2yrs** donne :

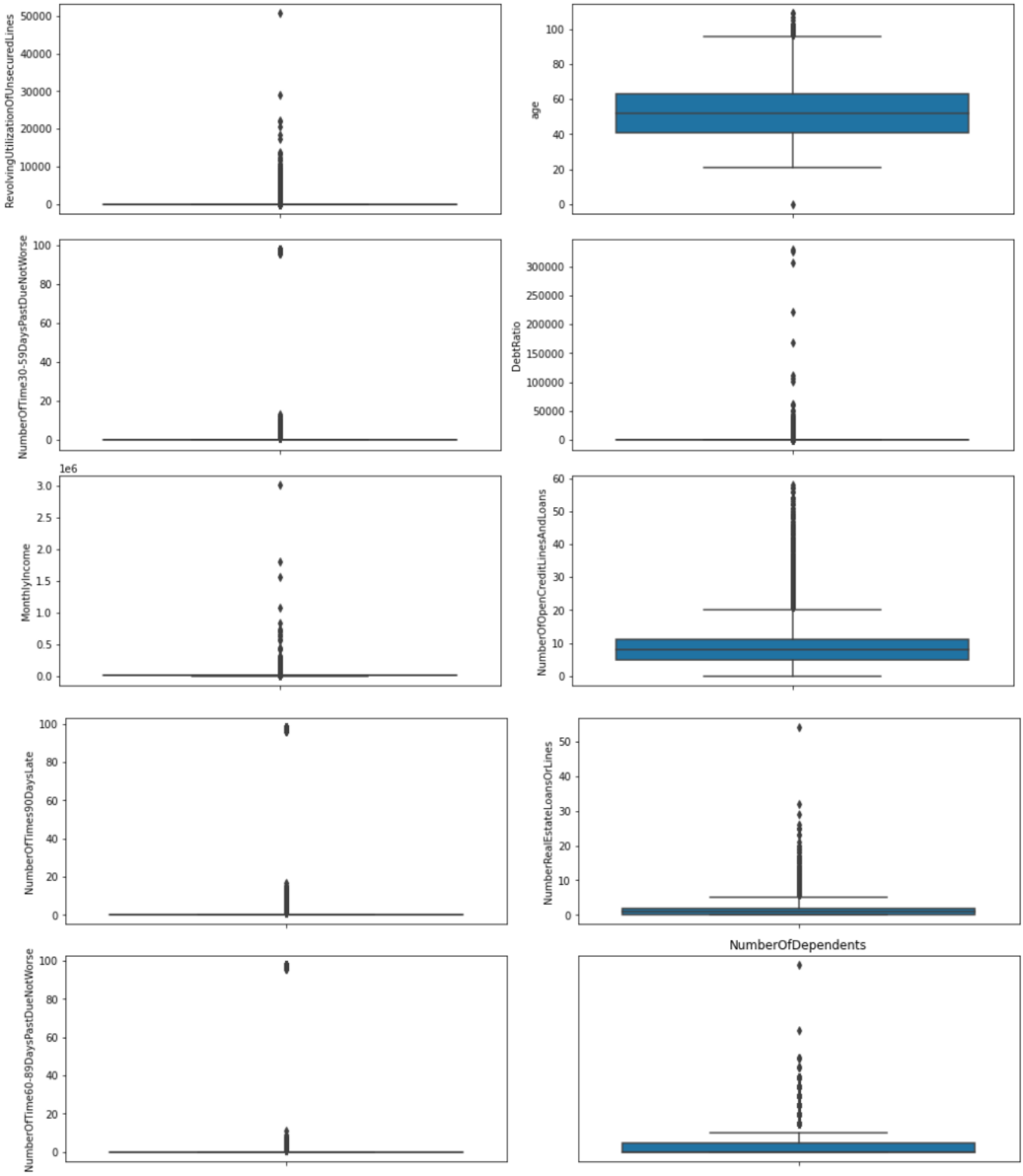
```
1    0.5
0    0.5
Name: SeriousDlqin2yrs, dtype: float64
```



3.2 Identification et élimination des données extrêmes

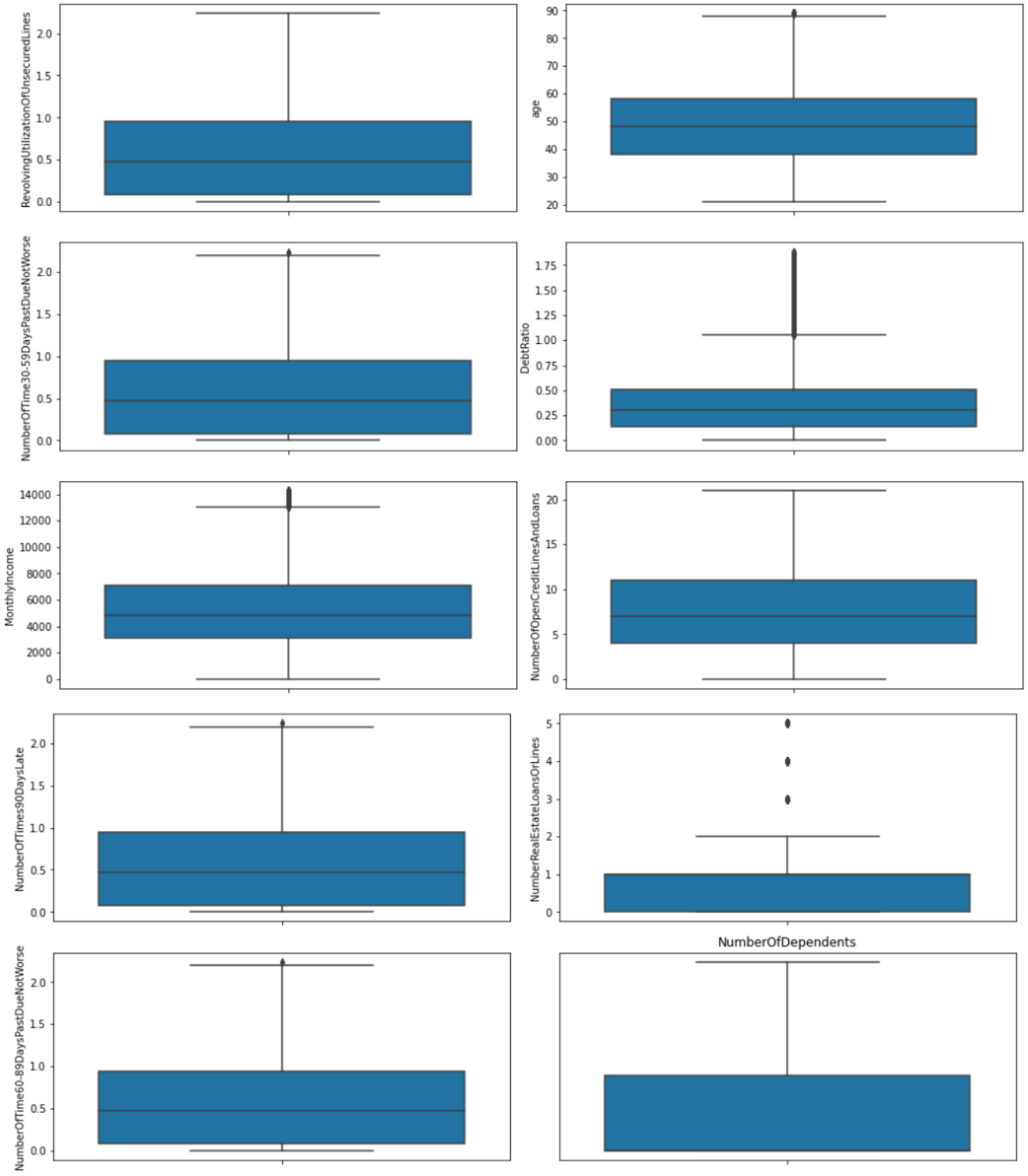
Une valeur aberrante est une valeur qui s'écarte fortement des valeurs des autres observations, anormalement faible ou élevée.





• **Èlimination des valeurs aberrantes**

En se basant sur les 2 relations Q1-1,5 * IQR et Q3+1,5 * IQR, on a éliminé les valeurs aberrantes. Après l'élimination des valeurs aberrantes, les boites à moustaches figurent comme ceci :

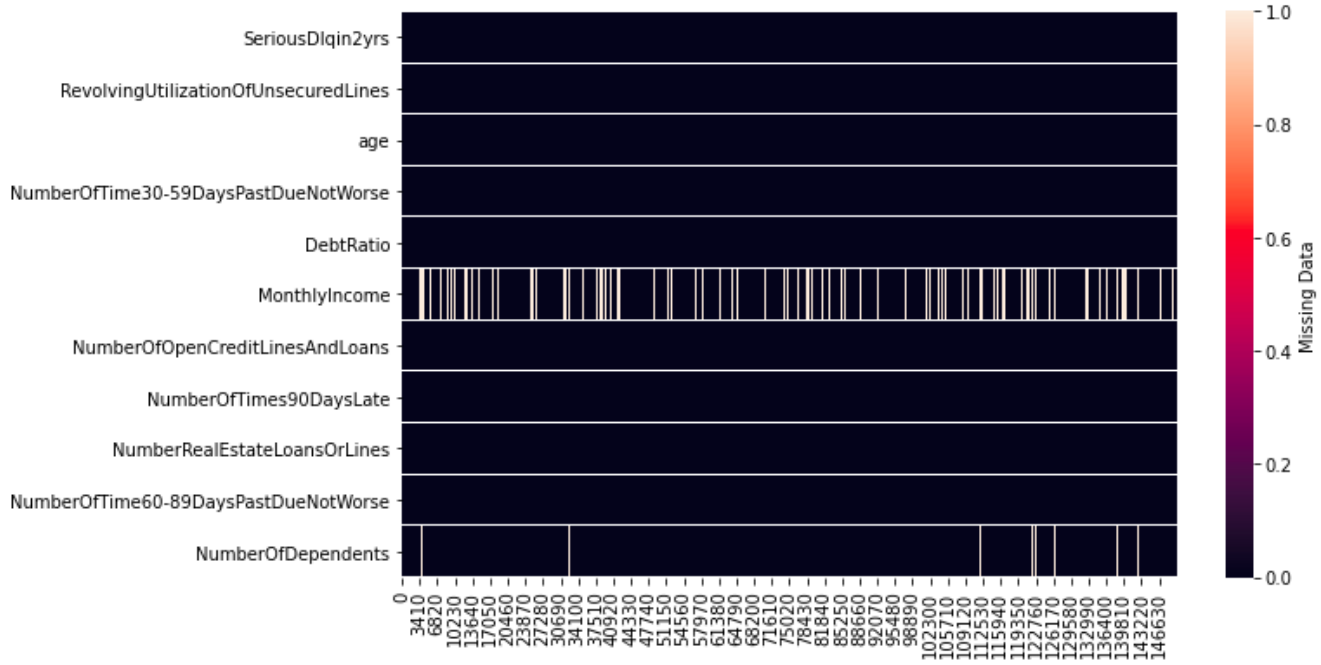

RevolvingUtilizationOfUnsecuredLines
age
NumberOfTime30-59DaysPastDueNotWorse
DebtRatio
MonthlyIncome
NumberOfOpenCreditLinesAndLoans
NumberOfTimes90DaysLate
NumberRealEstateLoansOrLines
NumberOfTime60-89DaysPastDueNotWorse
NumberOfDependents

3.3 Traitement des valeurs manquantes

Le nombre total des valereus manquantes pour chaque colonne :

```
RevolvingUtilizationOfUnsecuredLines       0
age                                        0
DebtRatio                                  0
MonthlyIncome                            227
NumberOfOpenCreditLinesAndLoans            0
NumberRealEstateLoansOrLines               0
NumberOfDependents                        41
NumberOfTime30-59DaysPastDueNotWorse     127
NumberOfTimes90DaysLate                  127
NumberOfTime60-89DaysPastDueNotWorse     127
```

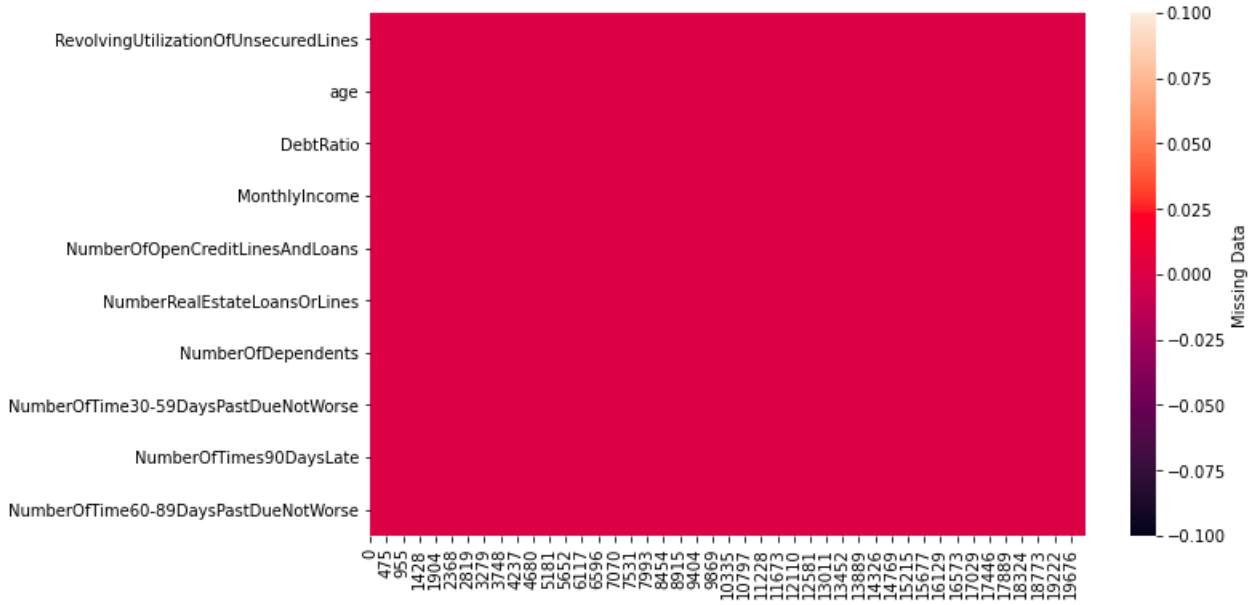
- Heatmap avant traitement des missing values :



```
Out[20]: RevolvingUtilizationOfUnsecuredLines     0.000000
         age                                      0.000000
         DebtRatio                                0.000000
         MonthlyIncome                            0.015842
         NumberOfOpenCreditLinesAndLoans          0.000000
         NumberRealEstateLoansOrLines             0.000000
         NumberOfDependents                       0.003075
         NumberOfTime30-59DaysPastDueNotWorse     0.008824
         NumberOfTimes90DaysLate                  0.008824
         NumberOfTime60-89DaysPastDueNotWorse     0.008824
         dtype: float64
```

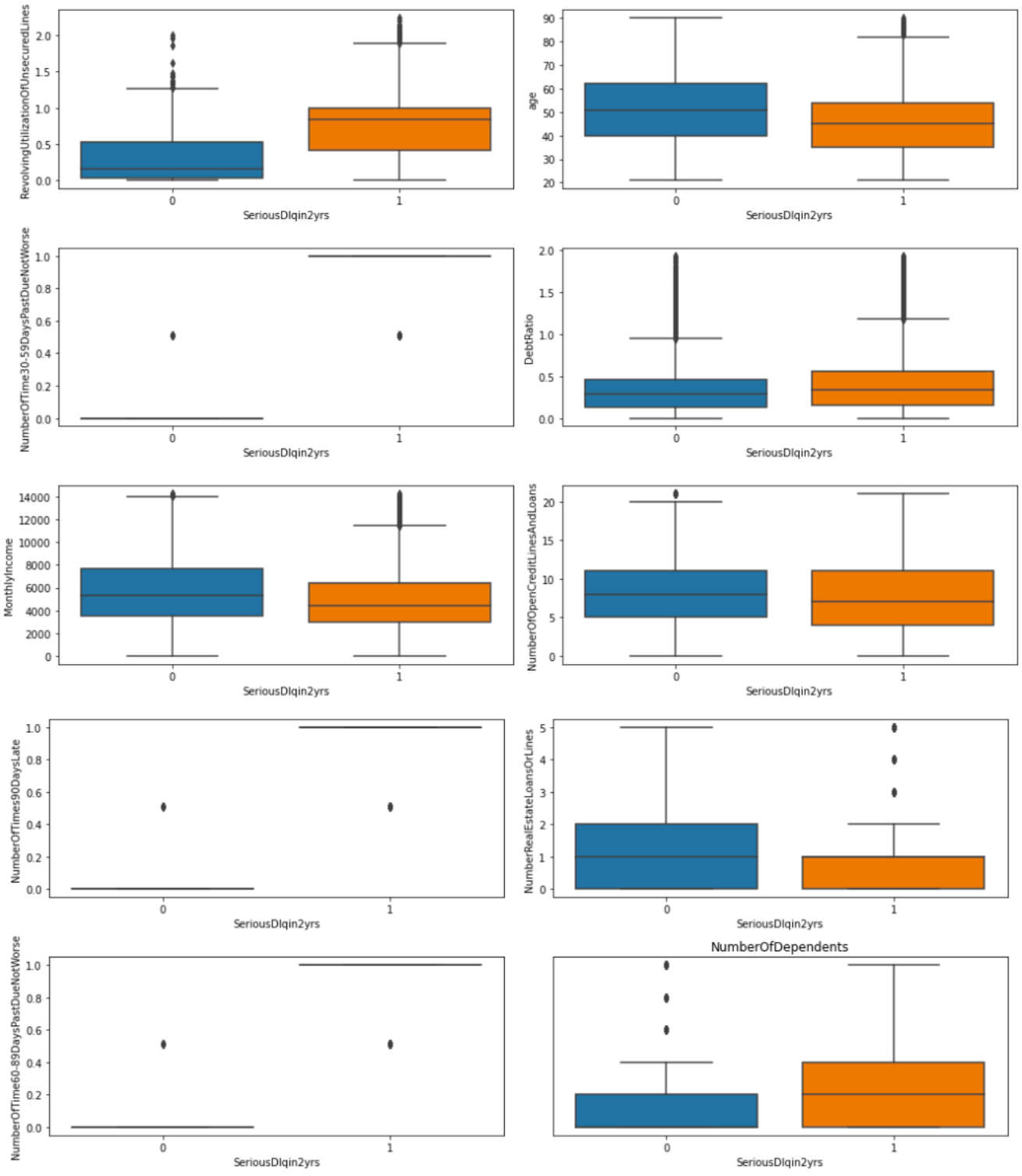
Figure 1: Pourcentage des valeurs manquantes

- Traitement des missing values :
Puisque le pourcentage des valeurs manquantes est trés faible alors on va remplacer chacune par la moyennes de la colonne qu'elle appartient. **- Heatmap après traitement des missing values :**



3.4 Identification des meilleurs prédicteurs parmi les variables:

On a representé les boites à moustahce des deux classes 0 et 1 pour chaque variable :



Les meilleurs variables dans le modèle de prédiction sont celles qui ont des médianes trop séparés entre chaque classe de **SeriousDlqin2yrs**.

4 Modèles de prévision

On a choisi une méthode de validation croisée basée sur la fonction `train_test_split` du package `sklearn.model_selection`

4.1 Application de LDA et QDA

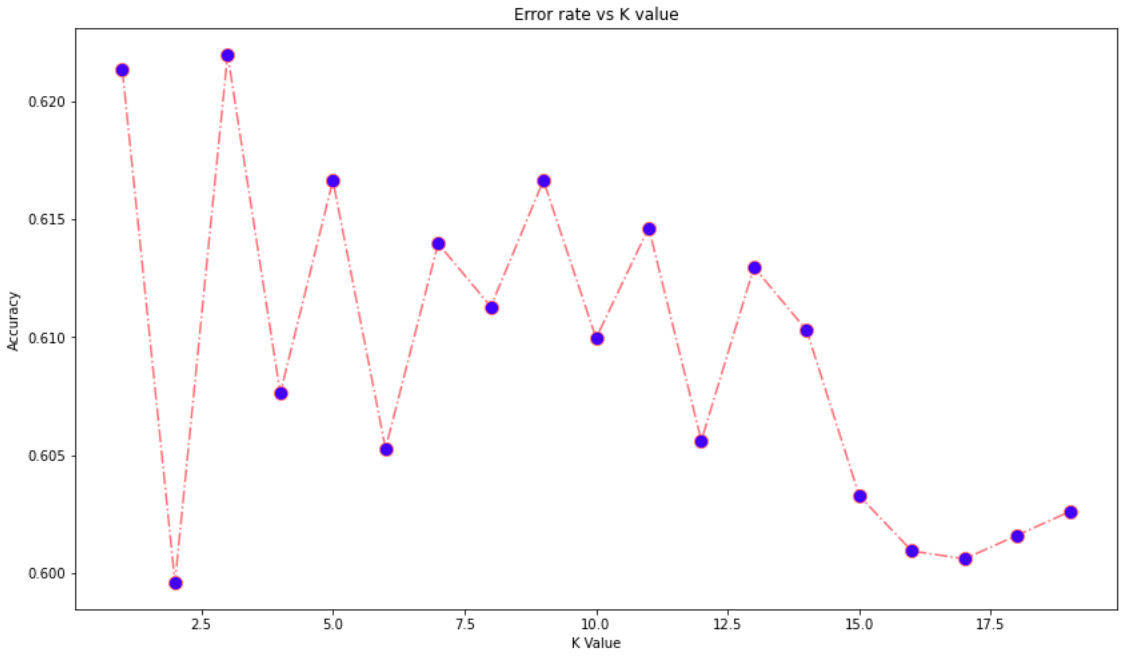
Voir le notebook

4.2 Régression logistique

Voir le notebook

4.3 Application d'autres modèles

• **Nearest neighbors classifier :**



5 Phase d'évaluation et règle de décision retenue

5.1 Comparaison des Godness of fit

La méthode prédifinie `cross_val_score` permet de calculer à quel niveau le modèle convient aux données existantes.

La précision moyenne de LDA: 0.999 (0.001)

La précision moyenne de QDA: 0.599 (0.123)

La précision moyenne du regression logistique: 0.999 (0.001)

La précision moyenne de Knn: 0.621 (0.012)

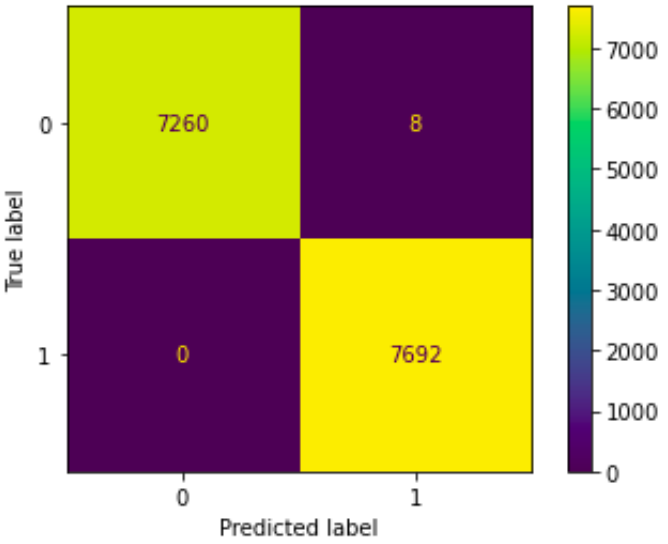
La précision moyenne de RFS: 0.999 (0.001)

La précision moyenne de DST: 0.999 (0.001)

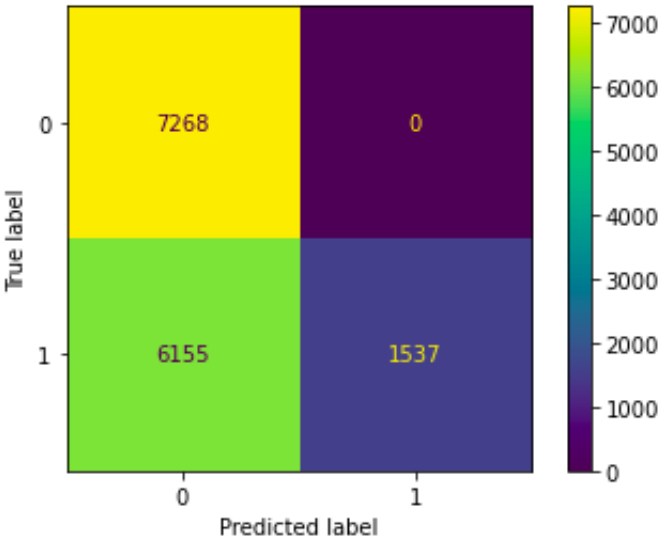
5.2 Pouvoir de prédiction

5.2.1 Matrices de confusion

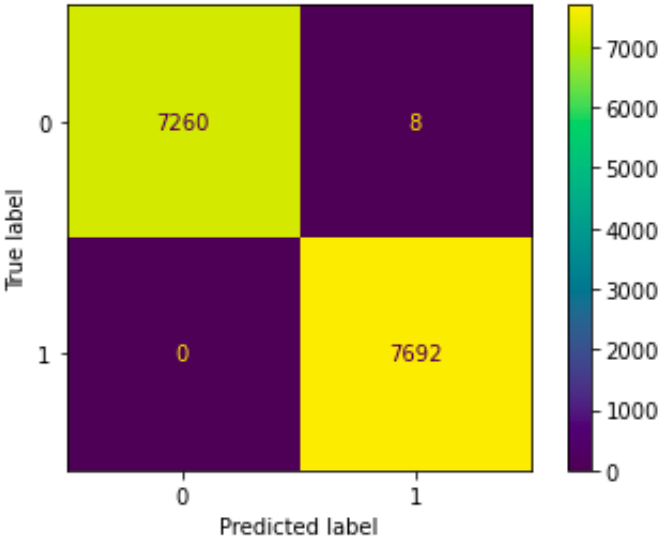
- **LDA Model :**



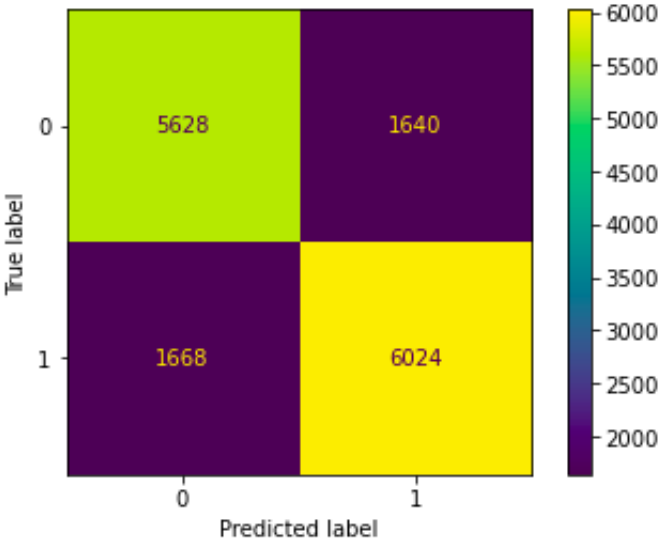
- **QDA Model :**



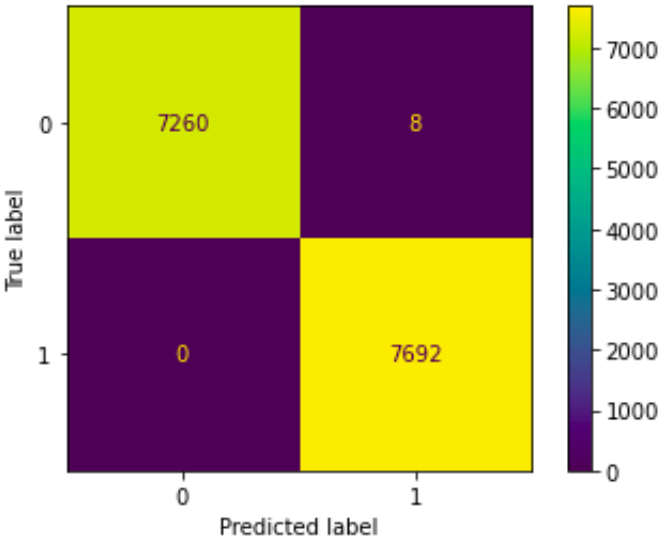
- **Regression logistique Model :**



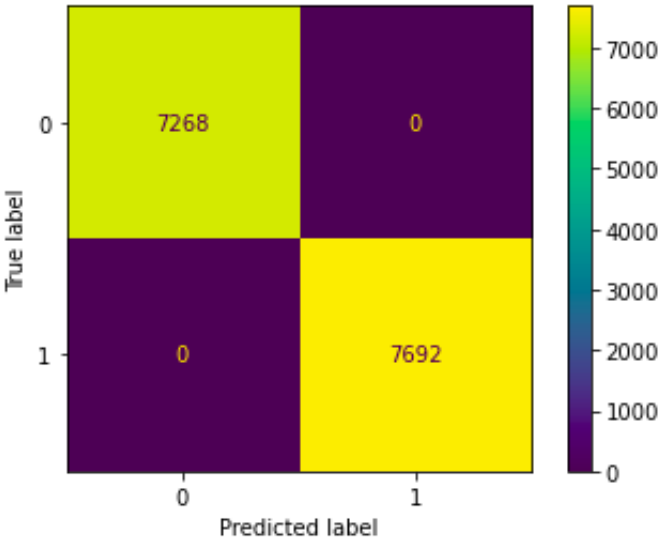
- **Knn Model :**



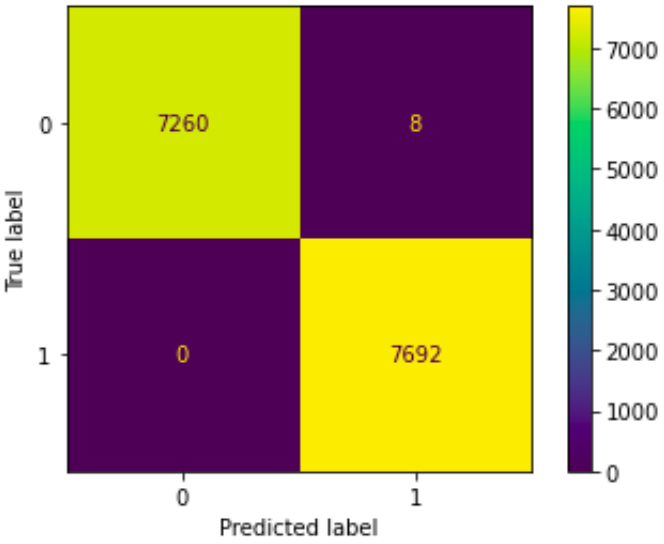
- **Random forest Model :**



- **random tree Model :**



- **SVM Model :**


7260
8
0
7692
0
1
True label
0
1
Predicted label
7000
6000
5000
4000
3000
2000
1000
0