

Master : Sécurité Systèmes et Services

Projet du Machine Learning :

# Credit Cards Fraud Detection



Credit Card Fraud Detection

Réalisé par :

AIT HMID Hafsa

MELLIANI Asmae

Encadré par :

Pr. Mohamed LAZAAR

Année universitaire : 2021/2022

## Table des matières

|  |           |
|--|-----------|
| <b>CHAPITRE 1 : Introduction générale .....</b>          | <b>4</b>  |
| 1.1. Organisation du projet.....                         | 4         |
| 1.2. Problématique .....                                 | 4         |
| 1.3. Hypothèses .....                                    | 4         |
| <b>CHAPITRE 2 : Prétraitement de données.....</b>        | <b>5</b>  |
| 2.1. Introduction .....                                  | 5         |
| 2.2. Processus de prétraitement .....                    | 5         |
| 2.3. La dataset utilisée .....                           | 7         |
| 2.4. Importation des bibliothèques.....                  | 7         |
| 2.5. Visualisation de données .....                      | 8         |
| <b>CHAPITRE 3 : Extraction des caractéristiques.....</b> | <b>13</b> |
| 3.1. Introduction .....                                  | 13        |
| 3.2. Processus d'extraction.....                         | 13        |
| 3.3. Algorithme PCA.....                                 | 13        |
| 3.4. Chi - square .....                                  | 17        |
| <b>CHAPITRE 4 : Implémentation de PMC ET SVM.....</b>    | <b>18</b> |
| 4.1. Perceptron multicouche (PMC).....                   | 18        |
| 4.2. Support Vecteur Machine (SVM).....                  | 19        |
| 4.3. La validation des modèles .....                     | 20        |
| <b>CHAPITRE 5: Conclusion générale .....</b>             | <b>21</b> |
| 5.1. Conclusion.....                                     | 21        |
| 5.2. Percepectives.....                                  | 21        |
| <b>Références .....</b>                                  | <b>23</b> |

## Table des Figures

|  |           |
|--|-----------|
| <b>FIGURE 1 : ILLUSTRATION D'UNE PARTIE DE LA DATA SET CREDITCARD.CSV .....</b>                | <b>7</b>  |
| <b>FIGURE 2 :LA DISTRIBUTION DES CAS LEGITIMES ET CAS FRAUDE .....</b>                         | <b>9</b>  |
| <b>FIGURE 3: LA DENSITE DES DISTRIBUTIONS DE LA CLASSE AMOUNT .....</b>                        | <b>9</b>  |
| <b>FIGURE 4: LA DENSITE DES DISTRIBUTIONS DE LA CLASSE TIME .....</b>                          | <b>10</b> |
| <b>FIGURE 5: LA MATRICE DE CORRELATION POUR TOUTES LES DIMENSIONS DE CREDITCARD.CSV .....</b>  | <b>11</b> |
| <b>FIGURE 6: LES HISTOGRAMMES D'AVANCEMENT DE TOUTES LES DIMENSIONS DE CREDITCARDS.CSV....</b> | <b>12</b> |
| <b>FIGURE 7: LA REDUCTION DE DIMENSIONNALITE PAR PCA EN DEUX COMPOSANTES .....</b>             | <b>14</b> |
| <b>FIGURE 8: LE DIAGRAMME DES TRANSACTIONS APRES UNDERSAMPLING.....</b>                        | <b>15</b> |
| <b>FIGURE 9: LA MATRICE DE CORRELATION APRES LE BALANCEMENT DE DATASET .....</b>               | <b>16</b> |
| <b>FIGURE 10: LA MATRICE DE CONFUSION POUR L'APPLICATION DE PMC .....</b>                      | <b>18</b> |
| <b>FIGURE 11: LA MATRICE DE CONFUSION AVEC L'APPLICATION DE SVM .....</b>                      | <b>19</b> |

### 1.1. Organisation du projet :

Ce présent travail sera structuré en 3 parties principales : nous présenterons comme première partie un prétraitement de données (Visualisation, Nettoyage, Normalisation...).

Dans la seconde partie, nous allons faire une extraction de caractéristiques en utilisant les méthodes qui fournissent des résultats satisfaisants.

Le troisième chapitre, sera la partie qui concerne l'implémentation des algorithmes de Machine Learning.

### 1.2. Problématique :

Alors que nous nous dirigeons vers le monde numérique, la cybersécurité devient un élément crucial de notre vie. Lorsque nous parlons de sécurité dans la vie numérique, le principal défi consiste à trouver l'activité anormale.

Lorsque nous effectuons une transaction lors de l'achat d'un produit en ligne, un bon nombre de personnes préfèrent les cartes de crédit. La limite de crédit des cartes de crédit nous aide parfois à faire des achats même si nous n'avons pas le montant à ce moment-là. Mais, d'autre part, ces fonctionnalités sont mal utilisées par les cyber-attaquants.

Dans ce projet on s'intéresse au problème de la détection des fraudes dans les cartes crédit.

### 1.3. Hypothèses :

Pour résoudre ce problème, nous avons besoin d'un système qui peut interrompre la transaction si elle semble louche.

Ici, vient le besoin d'un système capable de suivre le modèle de toutes les transactions et si un modèle est anormal, la transaction doit être abandonnée.

Aujourd'hui, nous disposons de nombreux algorithmes d'apprentissage automatique qui peuvent nous aider à classer les transactions anormales. La seule exigence est les données passées et l'algorithme approprié qui peut mieux adapter nos données.

### 2.1. Introduction :

Le prétraitement des données est une technique d'exploration de données qui consiste à transformer les données brutes en un format compréhensible. Les données du monde réel sont souvent incomplètes, incohérentes, absentes de certains comportements ou tendances et sont susceptibles de contenir de nombreuses erreurs. C'est une méthode éprouvée pour résoudre ces problèmes puisque il prépare les données brutes pour un traitement ultérieur.

Le prétraitement des données est utilisé dans les applications basées sur des bases de données telles que la gestion de la relation client et les applications basées sur des règles (comme les réseaux neuronaux).

Dans les processus d'apprentissage automatique (ML), le prétraitement des données est essentiel pour coder l'ensemble de données sous une forme qui pourrait être interprétée et analysée par l'algorithme.

### 2.2 Processus de prétraitement :

Les données passent par une série d'étapes pendant le prétraitement :

#### **Nettoyage des données**

Les données sont nettoyées via des processus tels que le remplissage des valeurs manquantes ou la suppression de lignes avec des données manquantes, le lissage des données bruyantes ou la résolution des incohérences dans les données.

Le lissage des données bruyantes est particulièrement important pour les ensembles de données ML, car les machines ne peuvent pas utiliser des données qu'elles ne peuvent pas interpréter. Les données peuvent être nettoyées en les divisant en segments de taille égale qui sont ainsi lissés (binning), en les adaptant à une fonction de régression linéaire ou multiple (régression), ou en les regroupant en grappes de données similaires (clustering).

Des incohérences dans les données peuvent survenir en raison d'erreurs humaines (les informations ont été stockées dans un mauvais champ). Les valeurs dupliquées doivent être supprimées par

déduplication pour éviter de donner un avantage (biais) à cet objet de données.

## **Intégration Des Données**

Les données avec différentes représentations sont rassemblées et les conflits au sein des données sont résolus.

## **Transformation des données**

Les données sont normalisées et généralisées. La normalisation est un processus qui garantit qu'aucune donnée n'est redondante, tout est stocké dans un seul endroit et toutes les dépendances sont logiques.

## **Réduction de donnée**

Lorsque le volume de données est énorme, les bases de données peuvent devenir plus lentes, coûteuses à accéder et difficiles à stocker correctement. L'étape de réduction des données vise à présenter une représentation réduite des données dans un entrepôt de données.

Il existe différentes méthodes pour réduire les données. Par exemple, une fois qu'un sous-ensemble d'attributs pertinents est choisi pour sa signification, tout ce qui se trouve en dessous d'un niveau donné est ignoré. Des mécanismes de codage peuvent également être utilisés pour réduire la taille des données. Si toutes les données d'origine peuvent être récupérées après la compression, l'opération est étiquetée comme sans perte.

Si certaines données sont perdues, cela s'appelle une réduction avec perte. L'agrégation peut également être utilisée, par exemple, pour condenser d'innombrables transactions en une seule valeur hebdomadaire ou mensuelle, ce qui réduit considérablement le nombre d'objets de données.

## **Discrétisation des données**

Les données pourraient également être discrétisées pour remplacer les valeurs brutes par des niveaux d'intervalle. Cette étape implique la réduction d'un nombre de valeurs d'un attribut continu en divisant la plage d'intervalles d'attribut.

## Échantillonnage des données

Parfois, en raison de contraintes de temps, de stockage ou de mémoire, un ensemble de données est trop volumineux ou trop complexe pour être utilisé. Les techniques d'échantillonnage peuvent être utilisées pour sélectionner et travailler avec juste un sous-ensemble de l'ensemble de données, à condition qu'il ait approximativement les mêmes propriétés que celui d'origine.

### 2.3 La dataset utilisée :

On va travailler sur une dataset qui contient les transactions bancaires réelles effectuées par les titulaires de carte européens au cours de l'année 2013 (creditcard.csv).

L'avantage de cette dataset sur laquelle on a travaillé c'est qu'elle est déjà standardisée. Les caractéristiques V1, V2, ... V28 sont déjà obtenues avec PCA, les seules caractéristiques qui n'ont pas été transformées avec PCA sont '**Time**' et '**Amount**'.

|   | Time | V1        | V2        | V3       | V4        | V5        | V6        | V7        | V8        | V9        | ... |
|---|------|-----------|-----------|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----|
| 0 | 0    | -1.359807 | -0.072781 | 2.536347 | 1.378155  | -0.338321 | 0.462388  | 0.239599  | 0.098698  | 0.363787  | ... |
| 1 | 0    | 1.191857  | 0.266151  | 0.166480 | 0.448154  | 0.060018  | -0.082361 | -0.078803 | 0.085102  | -0.255425 | ... |
| 2 | 1    | -1.358354 | -1.340163 | 1.773209 | 0.379780  | -0.503198 | 1.800499  | 0.791461  | 0.247676  | -1.514654 | ... |
| 3 | 1    | -0.966272 | -0.185226 | 1.792993 | -0.863291 | -0.010309 | 1.247203  | 0.237609  | 0.377436  | -1.387024 | ... |
| 4 | 2    | -1.158233 | 0.877737  | 1.548718 | 0.403034  | -0.407193 | 0.095921  | 0.592941  | -0.270533 | 0.817739  | ... |

5 rows × 31 columns

Figure 1 : illustration d'une partie de la data set creditcard.csv

### 2.4 Importation des bibliothèques :

Dans cette étape, on va importer les bibliothèques python nécessaires pour l'implémentation de notre projet.

#### La Bibliothèque Pandas:

Nous permet d'importer notre jeu de données et crée généralement une matrice d'entités contenant la variable de la dépendante et l'indépendante.

#### La Bibliothèque Sklearn:

Comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions

logistiques des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser.

### **La Bibliothèque Numpy:**

Fournit de multiples fonctions permettant notamment de créer directement un tableau depuis un fichier ou au contraire de sauvegarder un tableau dans un fichier, et manipuler des vecteurs, matrices et polynômes.

### **La Bibliothèque Scipy:**

Dépend de Numpy, elle fournit de nombreuses pratiques numériques conviviales et efficaces telles que des routines pour l'intégration et les problèmes d'optimisation en analyse numérique.

### **La Bibliothèque Matplotlib:**

Essai de rendre les tâches simples et les choses compliquées possibles. Nous pouvons générer des graphes, histogrammes, des spectres de puissance des graphiques à barres, des graphiques d'erreur, des nuages de dispersion, etc... .

### **La Bibliothèque Seaborn:**

C'est une librairie qui vient s'ajouter à Matplotlib, remplace certains réglages par défaut et fonctions, et lui ajoute de nouvelles fonctionnalités.

## **2.5 Visualisation de données :**

Cette étape permet la visualisation sous formats des figures et tableaux, la distribution des différentes données de notre dataset (`creditcard.csv`) en fonctions de différents facteurs.

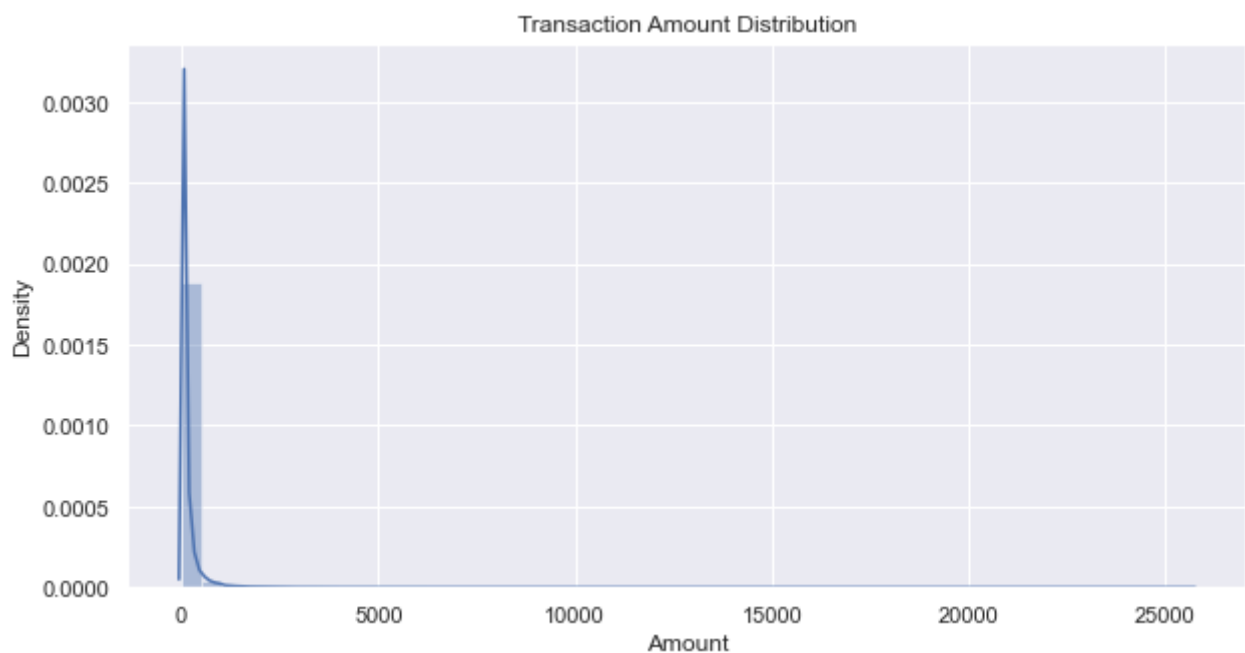
La courbe suivante par exemple illustre la distribution des cas légitimes et cas fraude pour notre dataset ([creditcard.csv](#)) : on remarque que le cas légitime (Normal) est plus dominant.





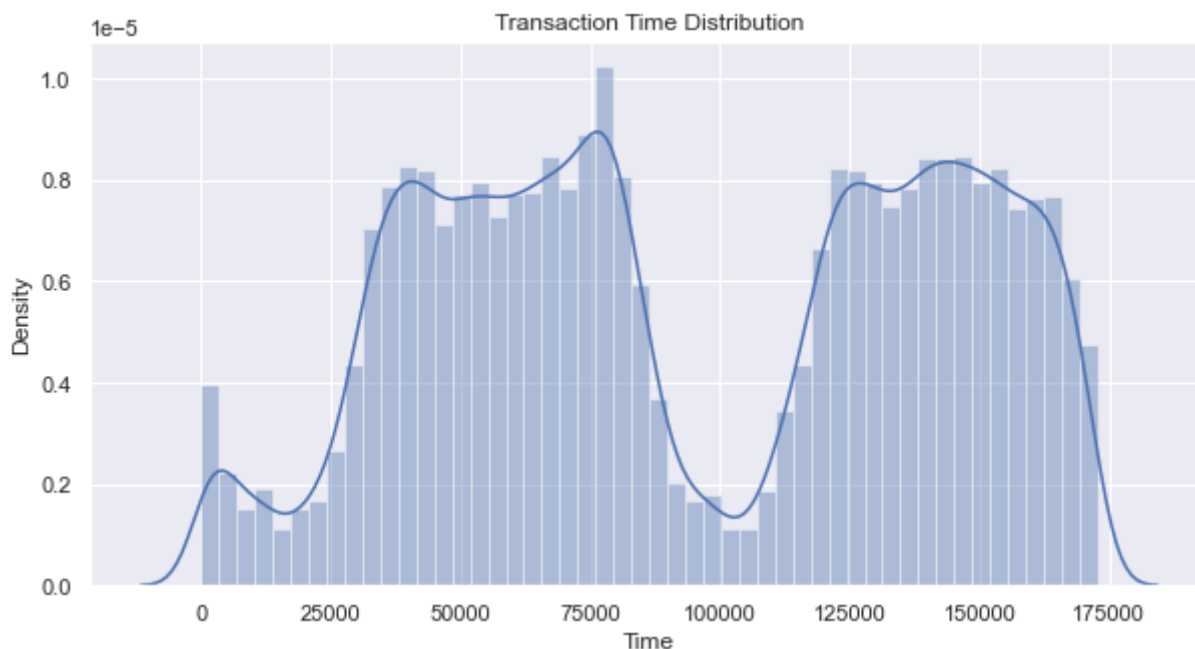
**Figure 2 :la distribution des cas légitimes et cas fraude**

La courbe qui suit nous montre que la densité des distributions de la classe Amount sur les 800-950 premières données est bien élevée, mais depuis ce chiffre elle a chuté rapidement jusqu'elle est devenue nulle sur tout le reste de l'ensemble.



**Figure 3: la densité des distributions de la classe Amount**

Pour le diagramme suivant, il illustre bien que la densité des distributions de la classe Time (en secondes) est variable sur l'ensemble de notre dataset de traitement, donc elle n'est pas stable. Donc il est difficile de prédire le temps de détection des fraudes.



**Figure 4: la densité des distributions de la classe Time**

Normalement si on travaille sur des dimensions réduites, on pourra bien lire les données à partir de notre modèle.

De même, on peut générer cette fois-ci une matrice de corrélation pour toutes les dimensions de notre dataset : [creditcard.csv](#)



On présente sous forme d'histogramme l'avancement de toutes les dimensions de notre dataset:



Figure 6: les histogrammes d'avancement de toutes les dimensions de creditcards.csv

La visualisation ci-dessus nous permet de voir les transactions de chacune des variables sous format des histogrammes.

### **Remarque :**

Le travail sur des données de grandes dimensions ne donnent pas toujours des résultats satisfaisants, raison pour laquelle nous aurons besoin d'une solution optimale pour la réduction des données. Ceci sera réalisé dans la prochaine étape.

### 3.1 Introduction :

L'extraction des caractéristiques consiste à réduire le jeu de données afin d'établir des modèles optimaux et de fournir des résultats satisfaisants.

Ce chapitre sera essentiellement consacré à la présentation de processus d'extraction de caractéristiques.

Cette partie comprend les deux algorithmes :

✓ *L'algorithme PCA*

✓ *L'algorithme Chi-square*

### 3.2 Processus d'extraction :

L'extraction de caractéristiques est souvent la phase la plus longue et la plus pénible de la création du projet de ML. Dans cette phase, on commence par utiliser nos données brutes et nos connaissances du domaine pour créer des caractéristiques qui feront fonctionner les algorithmes du **Machine Learning**. Dans cette partie, nous verrons ce qui fait une bonne caractéristique et comment la représenter dans notre modèle de **Machine Learning**.

Nous allons essayer de présenter deux méthodes vues dans le cours du Machine Learning : PCA et chi-square

### 3.3 Algorithme PCA

**L'analyse en composantes principales (ACP)** permet d'analyser et de visualiser un jeu de données contenant des individus décrits par plusieurs variables quantitatives.

C'est une méthode statistique qui permet d'explorer des données dites multi variées (données avec plusieurs variables). Chaque variable pourrait être considérée comme une dimension différente. Si vous avez plus de 3 variables dans votre jeu de données, il pourrait être très difficile de visualiser les données dans une "hyper-espace" multidimensionnelle.

L'analyse en composantes principales est utilisée pour extraire et de visualiser les informations

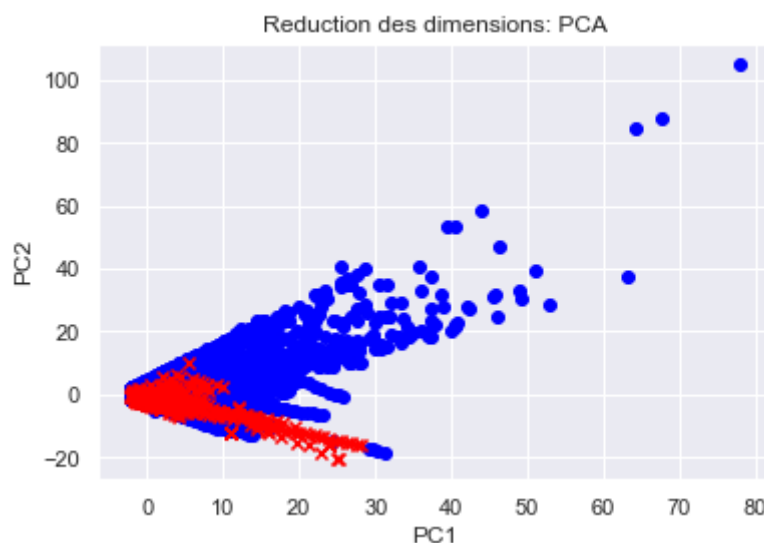
importantes contenues dans une table de données multi-variées. L'ACP synthétise cette information en seulement quelques nouvelles variables appelées **composantes principales**. Ces nouvelles variables correspondent à une combinaison linéaire des variables originels. Le nombre de composantes principales est inférieur ou égal au nombre de variables d'origine.

L'information contenue dans un jeu de données correspond à la variance ou l'*inertie totale* qu'il contient. L'objectif de l'ACP est d'identifier les directions (i.e., *axes principaux* ou composantes principales) le long desquelles la variation des données est maximale.

En d'autres termes, l'ACP réduit les dimensions d'une donnée multi variée à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d'information.

L'avantage de cette dataset sur laquelle on a travaillé c'est qu'elle est déjà standardisée avec PCA. Les caractéristiques V1, V2, ... V28 sont déjà obtenues avec PCA, les seules caractéristiques qui n'ont pas été transformées avec PCA sont 'Time' et 'Amount'.

Raison pour laquelle, nous n'aurons même pas besoin de refaire la standardisation de tous dataframes et des données, mais plutôt on s'intéresse à la transformation de **Time et Amount**.



**Figure 7: La réduction de dimensionnalité par PCA en deux composantes**

Nous avons utilisé le PCA pour réduire la dimensionnalité de notre ensemble de données à 2 dimensions afin de pouvoir le tracer. Cela nous dira s'il y a une certaine séparation entre la classe positive et la classe négative. Le graphique nous montre les deux composants principaux et nous pouvons voir que les transactions frauduleuses sont regroupées dans la partie inférieure gauche du graphique.

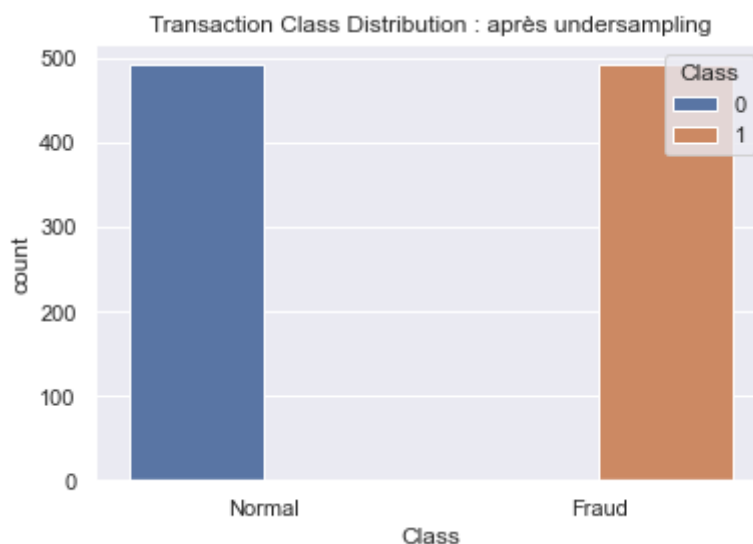
**On en déduit que notre dataset est toujours imbriquée.**

## Alors comment peut-on faire pour avoir une dataset balancée entre les cas Fraud et les Cas Legit ?

L'idée est simple, pour travailler sur une dataset balancée, nous allons prendre un échantillon aléatoire de notre ensemble de données et nous concaténons 2 dataframes : si axis = 0 ce dataframe sera ajouté un par un. Et ceci est connu sous le nom du **susample**.

EN effet le **subsample (undersampling)** est une technique permettant d'équilibrer les ensembles de données inégaux en conservant toutes les données dans la classe minoritaire et en diminuant la taille de la classe majoritaire. C'est l'une des nombreuses techniques que les Datascientist peuvent utiliser pour extraire des informations plus précises à partir d'ensembles de données initialement déséquilibrés.

**Nous aurons alors des données distribuées uniformes avec 492 transactions frauduleuses et 492 transactions normales. C'est comme ceci que nous avons obtenu l'amélioration de notre modèle dans ce qui suit.**



**Figure 8: Le diagramme des transactions après Undersampling**

Maintenant après avoir appliqué la technique du **Undersampling**, nous avons obtenu notre dataframe balancée correctement : 492 Fraud et 492 Legit.



## Matrice de corrélation :

Les matrices de corrélation sont l'essence même de la compréhension de nos données. Nous voulons savoir s'il existe des caractéristiques qui influencent fortement si une transaction spécifique est une fraude.

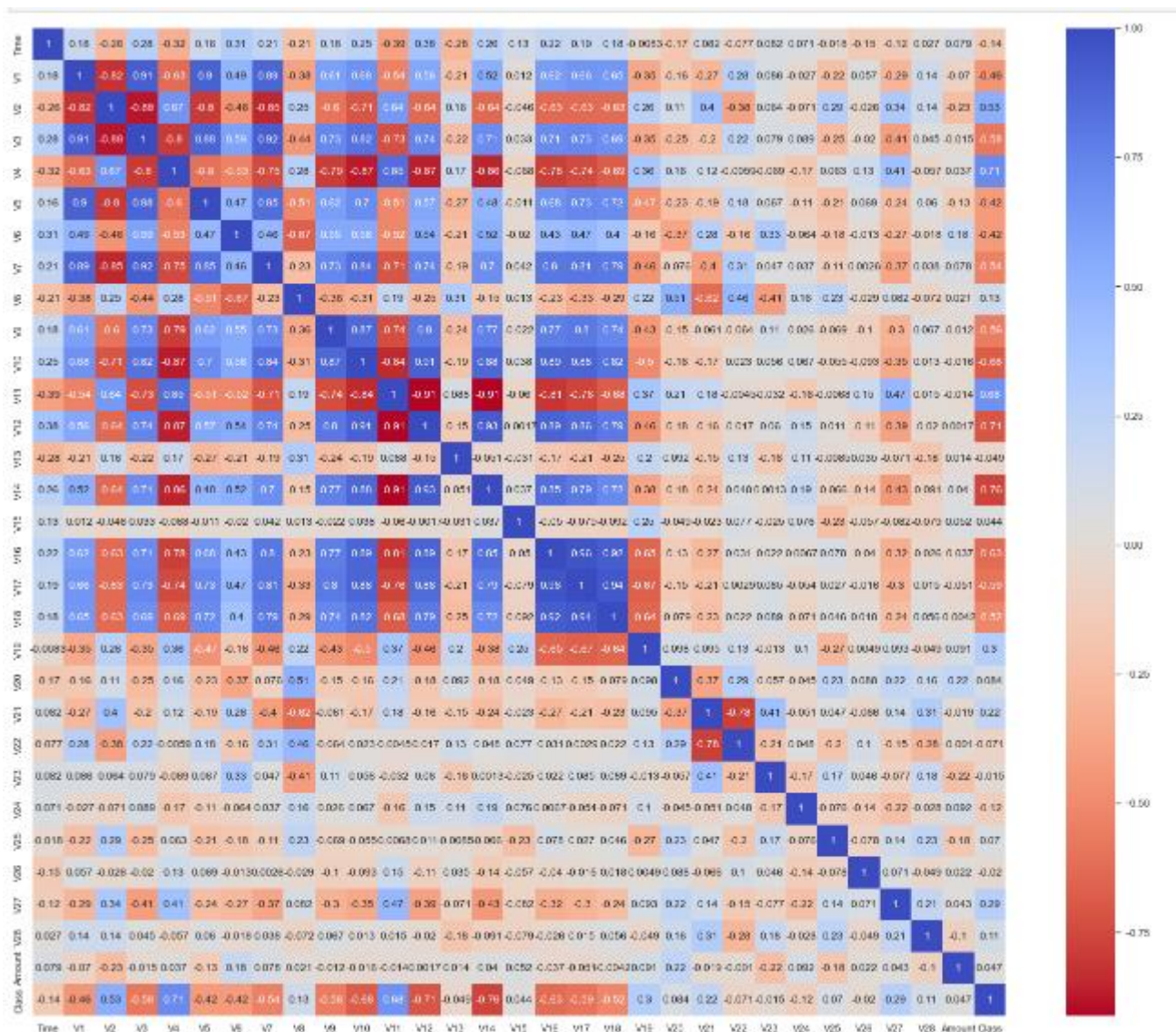


Figure 9: la matrice de corrélation après le balancement de dataset

### Interprétation de ces résultats :

**Corrélations négatives :** V17, V14, V12, V16 et V10 sont corrélés négativement

**Corrélations positives :** V2, V4, V11 et V19, V21, V27 sont positivement corrélés.



### 3.4 Chi - square :

Le Chi-square est un test d'hypothèse statistique valide lorsque la statistique de test est distribuée par le chi carré sous l'hypothèse nulle. C'est un test la dépendance entre les variables catégorielles.

Nous nous demandons toujours où le test Chi-Square est utile dans l'apprentissage automatique et comment ce test fait une différence. La sélection des fonctionnalités est un problème important dans l'apprentissage automatique, où nous aurons plusieurs fonctionnalités en ligne et nous devons sélectionner les meilleures fonctionnalités pour construire le modèle. Le test du chi-Square nous aide à résoudre le problème de sélection des fonctionnalités en testant la relation entre les fonctionnalités.

#### Étapes pour effectuer le test du chi-Square :

1. Définir l'hypothèse.
2. Construisez un tableau de contingence.
3. Trouvez les valeurs attendues.
4. Calculez la statistique Chi-Square.
5. Accepter ou rejeter l'hypothèse nulle.

Cette étape consiste à tourner les algorithmes du perceptron multicouche PMC et support vecteur machine SVM afin de pouvoir choisir le meilleur modèle pour la représentation des données.

### 4.1. Perceptron multicouche (PMC)

Le Perceptron multicouche (PMC) est un algorithme d'apprentissage supervisé qui apprend une fonction par formation sur un ensemble de données, où est le nombre de dimensions pour l'entrée et est le nombre de dimensions pour la sortie. Étant donné un ensemble de fonctionnalités et une cible, il peut apprendre un approximateur de fonction non linéaire pour la classification ou la régression. Elle est différente de la régression logistique, en ce sens qu'entre la couche d'entrée et la couche de sortie, il peut y avoir une ou plusieurs couches non linéaires, appelées couches cachées.

Après l'application de PMC sur notre dataset (**creditcard.csv**) on a obtenu les résultats suivants :

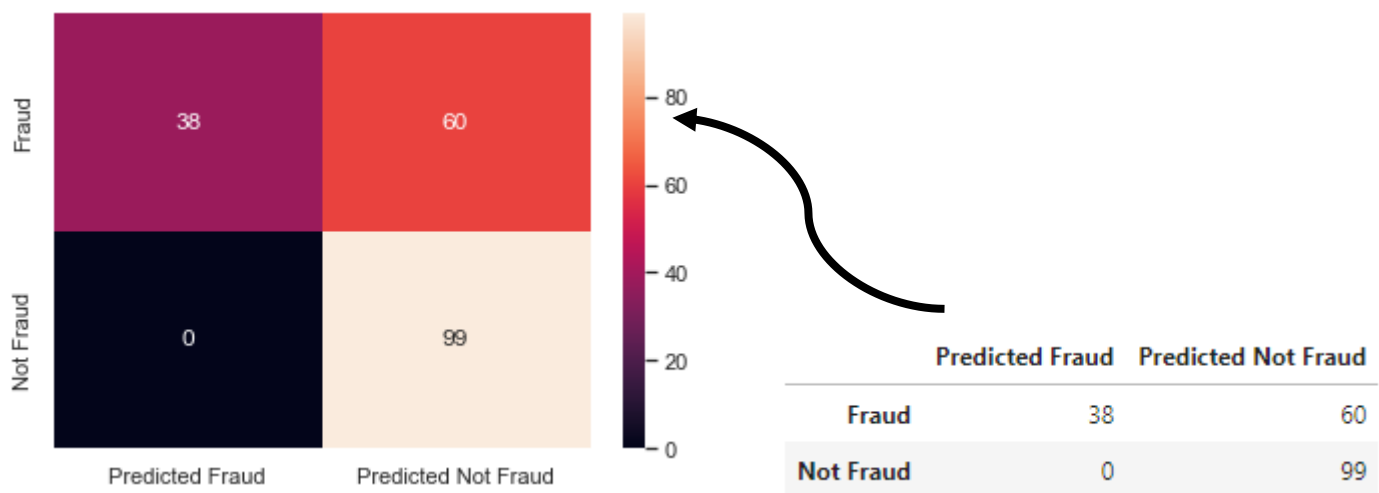


Figure 10: la matrice de confusion pour l'application de PMC

## 4.2 Support Vecteur Machine (SVM) :

Les machines à vecteurs de support (SVM) sont un ensemble de méthodes d'apprentissage supervisé utilisées pour la classification, la régression et la détection des valeurs aberrantes.

### Les avantages des SVM sont :

- 1- Efficace dans les espaces de grande dimension.
- 2- Toujours efficace dans les cas où le nombre de dimensions est supérieur au nombre d'échantillons.
- 3- Utilise un sous-ensemble de points d'apprentissage dans la fonction de décision (appelés vecteurs de support), il est donc également efficace en mémoire.
- 4- Polyvalent : différentes fonctions du noyau peuvent être spécifiées pour la fonction de décision. Des noyaux communs sont fournis, mais il est également possible de spécifier des noyaux personnalisés.

### Les inconvénients des SVM incluent :

Si le nombre de caractéristiques est bien supérieur au nombre d'échantillons, évitez le sur-ajustement dans le choix des fonctions du noyau et le terme de régularisation est crucial. Les SVM ne fournissent pas directement des estimations de probabilité, celles-ci sont calculées à l'aide d'une validation croisée quintuple coûteuse.

L'application de SVM sur la dataset (**creditcards.csv**) donne les résultats suivants :

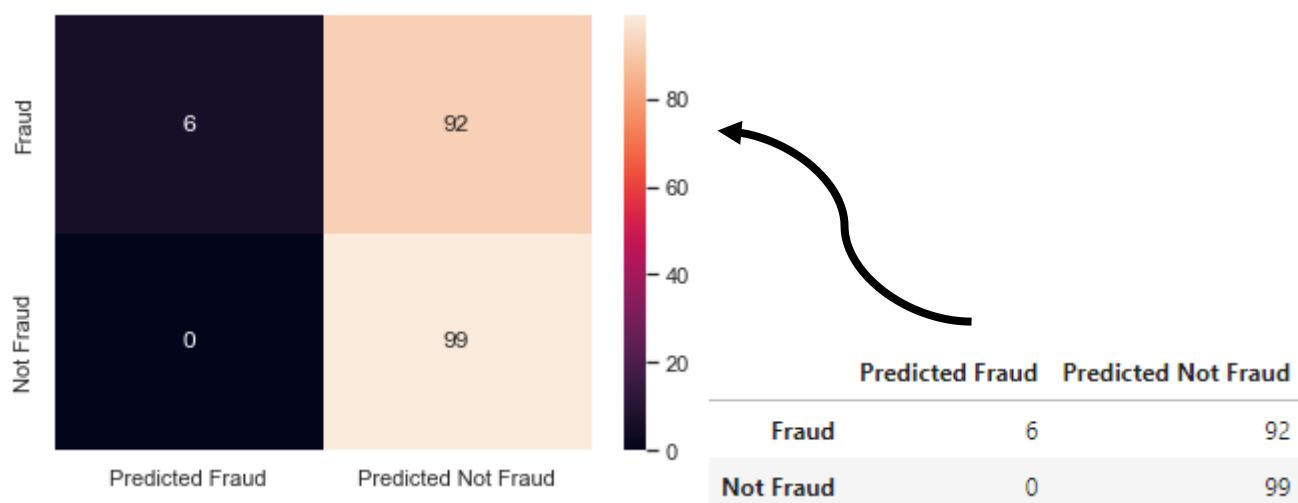


Figure 11: la matrice de confusion avec l'application de SVM

### 4.3 La validation des modèles :

Dans cette dernière étape, nous devons choisir lequel des modèles traités est meilleur pour la détection des Fraudes dans notre Dataset : **creditcard.csv**

Alors comment peut-on sélectionner le meilleur modèle selon quels critères ?

D'après avoir calculé les valeurs estimées du PMC ET SVM avec et sans cross validation, et en s'appuyant également sur la matrice de confusion, on peut sélectionner le modèle qui a la valeur de validation la plus proche à 1 (dans le cas de Cross Validation):

**\*Perceptron multicouche : 0.627344 (cross validation)**

**\*Support vecteur machine : 0.548778 (cross validation)**

**\*Le meilleur modèle est : Le perceptron Multicouche**

On conclut que pour le cas de notre dataset **creditcar.csv** avec des données bien spécifiques, nous avons trouvé que La méthode du PMC est plus optimale que SVM pour la détection des fraudes des cartes de crédits.

### 5.1 Conclusion

En guise de conclusion, La détection de fraude par carte de crédit présente plusieurs caractéristiques qui en font une tâche difficile.

Le défi consiste à reconnaître les transactions frauduleuses par carte de crédit afin que les clients des sociétés de cartes de crédit ne soient pas facturés pour des articles qu'ils n'ont pas achetés.

Les principaux défis liés à la détection de la fraude par carte de crédit sont :

- D'énormes données sont traitées chaque jour et la construction du modèle doit être suffisamment rapide pour répondre à l'arnaque à temps.

- Données déséquilibrées, c'est-à-dire que la plupart des transactions (99,8 %) ne sont pas frauduleuses, ce qui rend très difficile la détection des frauduleuses

- Disponibilité des données car les données sont pour la plupart privées.

- Les données mal classées peuvent être un autre problème majeur, car toutes les transactions frauduleuses ne sont pas détectées et signalées.

Après l'étude de cette problématique sur notre dataset [creditcard.csv](#) on a constaté que la méthode du PMC est plus optimale que SVM pour la détection des fraudes des cartes de crédits.

### 5.2 Perspectives :

Bien que nous n'ayons pas pu atteindre l'objectif d'une précision de 100 % en matière de détection de fraudes, nous avons fini par créer un système qui peut, avec suffisamment de temps et de données, se rapprocher de cet objectif. Comme avec n'importe quel projet, il y a toujours une amélioration. La nature même de ce projet permet à plusieurs algorithmes d'être intégrés ensemble sous forme de modules et leurs résultats peuvent être combinés pour augmenter la précision du résultat final. Ce modèle peut encore être amélioré avec l'ajout de plusieurs algorithmes.

Cependant, la sortie de ces algorithmes doit être dans le même format que les autres. Une fois que la condition est satisfaite, les modules sont faciles à ajouter et cela offre une grande modularité et polyvalence au projet.

Plus de marge d'amélioration peut être trouvée dans l'ensemble de données. Comme on a démontré auparavant, la précision des algorithmes augmente lorsque la taille du jeu de données est augmentée. Par conséquent, plus de données rendent le modèle plus précis dans la détection des fraudes et permettent la réduction du nombre de faux positifs. Cependant, cela nécessite un soutien officiel des banques elles-mêmes.

## Références

### Livres et articles

- **Data science : fondamentaux et études de cas : Machine Learning, EDITION EYROLLES 2015.**
- **“Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Veal” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017**
- **“Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonapat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014**
- **Apprentissage statistique: Réseaux de neurones. Cartes topologiques. Machines à vecteurs supports EDITION EYROLLES 2008 .**
- **Big Data et Machine Learning - 2e éd. - Les concepts et les outils de la data science, EDITION DUNOD 2016.**
- **Apprentissage machine - Clé de l'intelligence artificielle - Une introduction pour non-spécialistes, EDITION ELLIPSE 2019.**

### Webographie

- [www.DataScienceCentral.com](http://www.DataScienceCentral.com)
- [www.SmartDataCollective.com](http://www.SmartDataCollective.com)
- [www.Kaggle.com](http://www.Kaggle.com)
- [www.InsideBIGDATA.com](http://www.InsideBIGDATA.com)
- [wwwSimplyStatistics.org](http://wwwSimplyStatistics.org)