# Les Réseaux de Neurones avec *STATISTICA*

Au cours des deux dernières décennies, l'intérêt pour les *réseaux de neurones* s'est accentué. Cela a commencé par les succès rencontrés par cette puissante technique dans beaucoup de domaines, aussi variés que la finance, médecine, ingénierie, géologie et même physique.

Le succès radical des réseaux de neurones sur presque toutes les autres techniques statistiques peut être attribué à la puissance, *polyvalence* et facilité d'utilisation. Les réseaux de neurones sont des techniques de modélisation et de prévision très sophistiqués capable de modéliser des fonctions de modélisation et relations extrêmement complexes.

La capacité d'apprentissage par des exemples est une des nombreuses fonctionnalités des réseaux de neurones qui permet à l'utilisateur de modéliser des données et d'établir des règles précises de gouvernance sur la relation sous jacente entre divers données d'attribut. L'utilisateur des réseaux de neurones réunit des données représentatives, et ensuite exécute des *algorithmes d'apprentissage* qui peuvent apprendre automatiquement la structure des données. Bien que l'utilisateur ait besoin d'avoir quelques connaissances heuristiques sur la façon de *sélectionner* et *préparer* les données, le réseau de neurones approprié et interpréter les résultats, le niveau de connaissance de l'utilisateur exigé pour appliquer avec succès les réseaux de neurones est bien inférieur à celui exigé dans la plupart des outils et techniques statistiques traditionnelles, particulièrement quand les algorithmes des réseaux de neurones sont cachés derrière des programmes informatiques comme bien conçus qui emmène l'utilisateur du début à la fin en quelques clics.

Voyons un exemple d'utilisation des réseaux de neurones sur une thématique où l'on utilise principalement la régression logistique : le *Crédit Scoring*.

## Classification des Scores de Crédit avec les Réseaux de Neurones

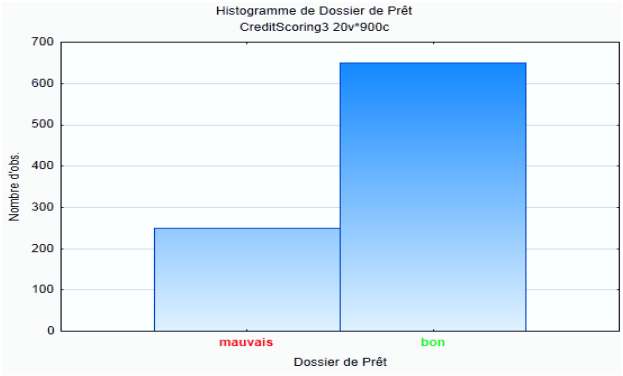
Un fichier de données contient les observations d'individus sollicitant un crédit. L'information comprend les détails depuis la demande de crédit avec le solde, durée, paiements, etc. La variable cible est *Dossier de Prêt,* qui est catégorielle par nature. Les clients sont classés soit en Bon ou Mauvais risque de crédit. Le but est de prévoir précisément les risques de crédit des clients (Bon ou Mauvais).

| | 1 Dossier de Prêt | 2 Solde du Compte Courant | 3 Durée du Prêt | 4 Remboursement des Prêts Antérieurs | 5 Objectif du Prêt | M |
|---|---|---|---|---|---|---|
| 1 | mauvais | pas de compte courant | 36 | aucun problème avec les prêts antérieurs | réorientation professionnelle | |
| 2 | bon | négligeable | 48 | délicat | réorientation professionnelle | |
| 3 | mauvais | >300€ | 36 | aucun prêt antérieur | véhicule d'occasion | |
| 4 | bon | pas de compte courant | 24 | remboursement intégral | véhicule neuf | |
| 5 | bon | >300€ | 24 | aucun prêt antérieur | réorientation professionnelle | |
| 6 | bon | négligeable | 12 | aucun prêt antérieur | réorientation professionnelle | |
| 7 | mauvais | pas de compte courant | 30 | aucun prêt antérieur | véhicule d'occasion | |
| 8 | bon | négligeable | 15 | remboursement intégral | mobilier | |
| 9 | bon | >300€ | 15 | remboursement intégral | mobilier | |
| 10 | mauvais | négligeable | 27 | remboursement intégral | mobilier | |
| 11 | mauvais | négligeable | 24 | aucun prêt antérieur | véhicule d'occasion | |
| 12 | mauvais | pas de compte courant | 18 | aucun prêt antérieur | réparations | |
| 13 | mauvais | négligeable | 36 | aucun problème avec les prêts antérieurs | réorientation professionnelle | |
| 14 | bon | >300€ | 6 | aucun problème avec les prêts antérieurs | réorientation professionnelle | |
| 15 | bon | pas de compte courant | 12 | aucun prêt antérieur | autre | |
| 16 | bon | >300€ | 42 | aucun prêt antérieur | mobilier | |
| 17 | bon | pas de compte courant | 48 | aucun prêt antérieur | véhicule neuf | |
| 18 | mauvais | négligeable | 24 | comptes courants à problèmes | réparations | |
| 19 | bon | négligeable | 24 | remboursement intégral | autre | |
| 20 | bon | négligeable | 48 | aucun prêt antérieur | création d'entreprise | |
| 21 | bon | <= 300€ | 12 | aucun prêt antérieur | mobilier | |
| 22 | bon | >300€ | 24 | remboursement intégral | véhicule d'occasion | |
| 23 | bon | pas de compte courant | 18 | remboursement intégral | autre | |
| 24 | bon | >300€ | 12 | aucun prêt antérieur | autre | |
| 25 | mauvais | >300€ | 10 | aucun prêt antérieur | autre | |
| 26 | bon | négligeable | 48 | aucun problème avec les prêts antérieurs | création d'entreprise | |
| 27 | mauvais | négligeable | 24 | aucun problème avec les prêts antérieurs | véhicule d'occasion | |

## Prélude à l'Analyse : Équilibrage de la Variable Cible

Dans le cadre de la classification, il est optimale d'avoir une répartition homogène des modalités de la variable cible. En effet, si une modalité, par exemple 'Bon' de notre variable cible *Dossier de Prêt*, est sur représentée, la modalité 'Mauvais' risque d'être mal modélisée. Un autre problème subsiste dans ce type de jeu de données : l'évaluation des modèles sur ces données. A titre d'exemple, prenons un jeu de données constitué à 90% de 'Bon' dossier et de 10% de 'Mauvais' dossier. Un modèle ayant tendance à toujours classifier en 'Bon' dossier aura une performance de 90% sur l'ensemble de nos données !

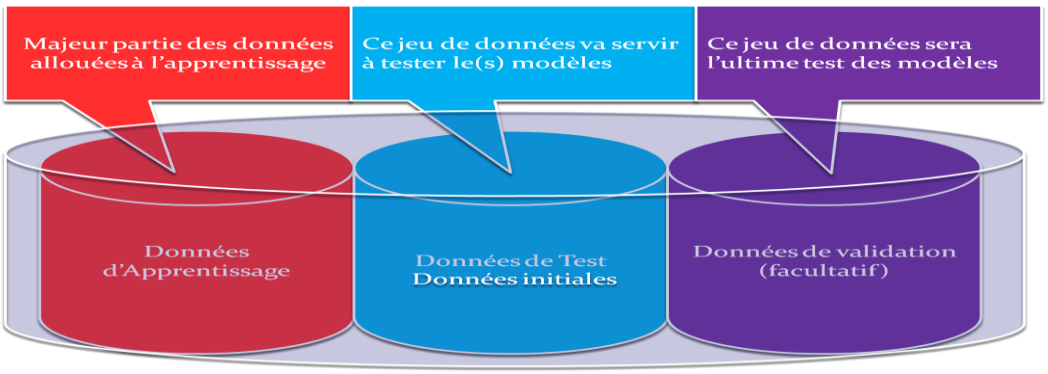
Regardons la répartition des modalités de la variable cible *Dossier de Prêt.*




L'histogramme nous montre qu'environ 70% des clients dans nos données sont classifiés comme '*Bon*'. Nous sommes dans le cas d'un jeu de données déséquilibré.

Pour remédier à ce problème, nous allons effectuer un échantillonnage aléatoire stratifié sur la variable *Dossier de Prêt*. Ce procédé va sélectionner aléatoirement un nombre d'observations spécifié par l'utilisateur pour chaque modalité d'une variable donnée (dans notre cas, *Dossier de Prêt*).

## Partitionnement des Données



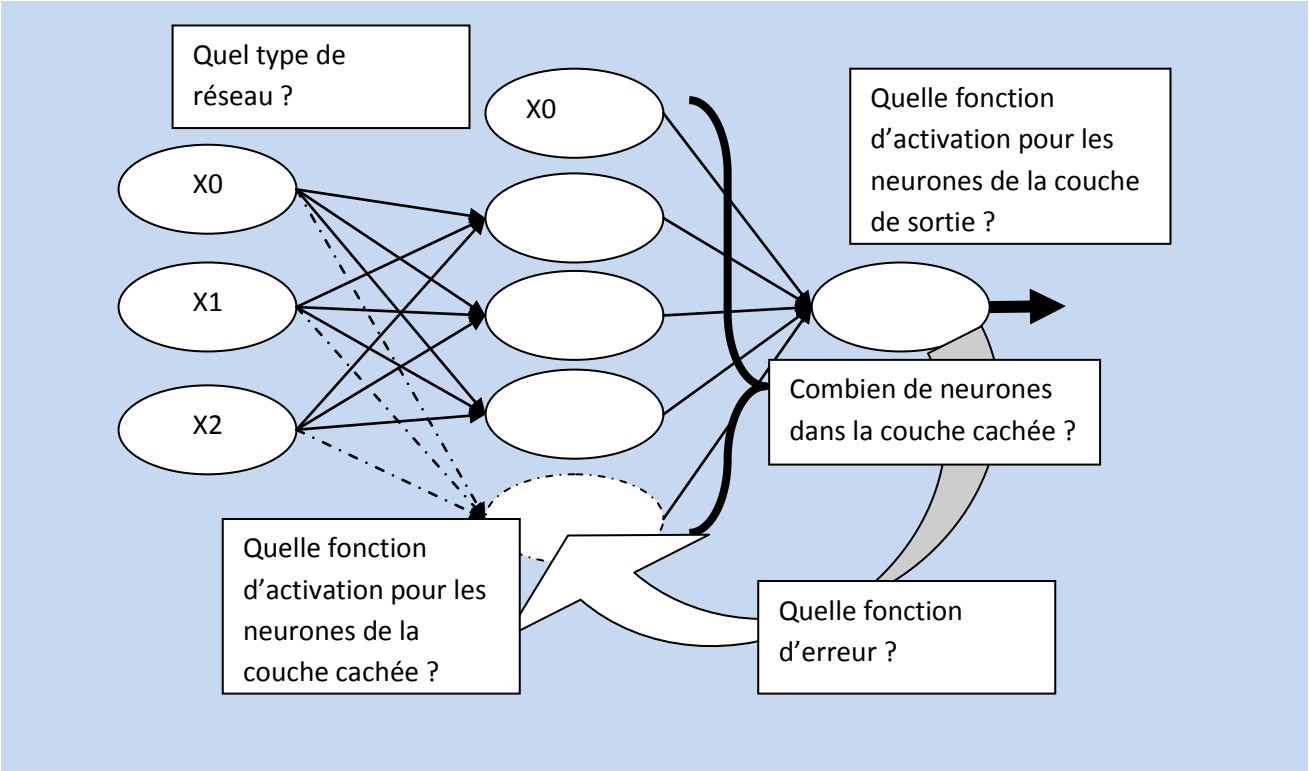
## Construction des Modèles

La variable cible étant donc *Dossier de Prêt*, une variable de type catégorielle. Les variables dites explicatives seront :

- De type continu : la *Durée du crédit*, le *Montant du prêt,* et l'*Age*

- De type catégorielles : le *Solde du compte courant,* l'état des *Remboursement des prêts antérieurs*, l'*Objectif du prêt*, la *Valeur d'épargne*, l'*Ancienneté chez l'employeur actuel* , l'*Endettement en % du revenu disponible*, le Sta*tut marital*, le *Sexe*, l'*Ancienneté dans le ménage actuel*, les *Actifs les plus importants*, les *Autres prêts en cours*, le *Type de logement*, le *Nombre de prêts antérieurs dans cette banque*, et l'*Occupation*

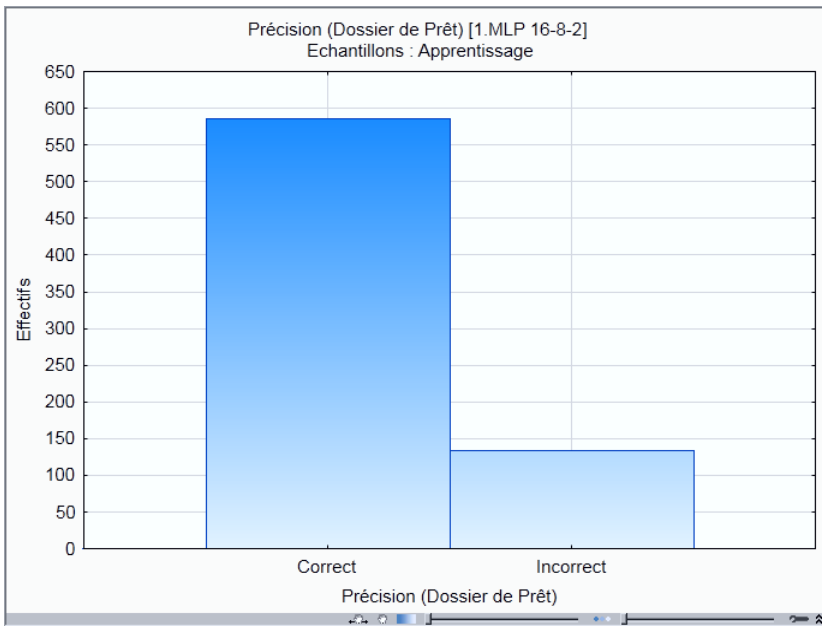
La principale difficulté de la construction d'un réseau de neurones est de déterminer la configuration qui optimise la modélisation. En effet, on ignore quelle configuration *fonction d'activation des neurones de la couche cachée - fonction d'activation des neurones de la couche de sortie - nombre de neurones dans la couche cachée – fonction d'erreur – type de réseau* est optimale à notre problématique.



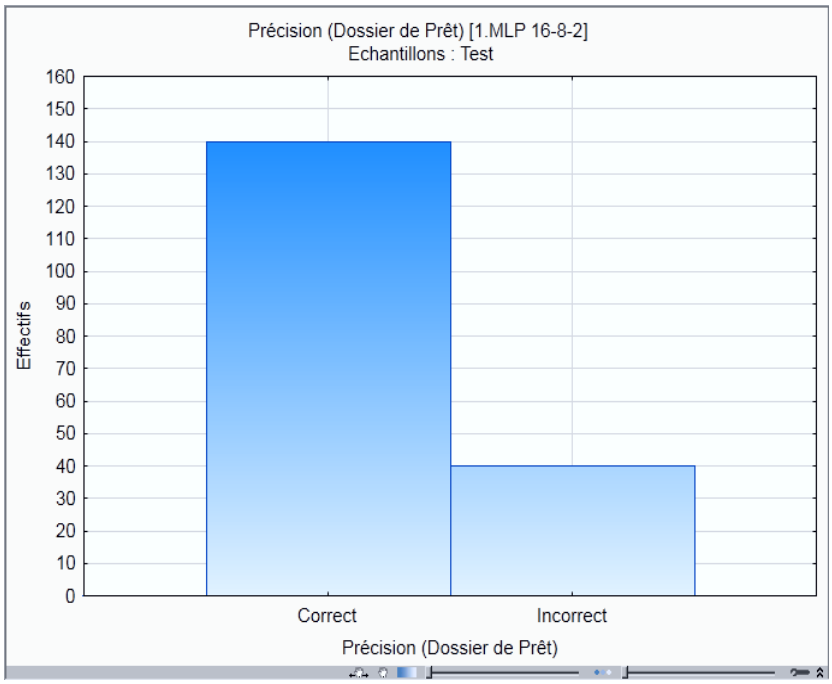

## Les Résultats !

Il a été construit une centaine de réseaux dont nous avons conservé le meilleur : celui présentant un bon compromis entre les performances en apprentissage et en test.

L'histogramme de précision ci-après nous montre que le modèle a classifié correctement la majeure partie des dossiers (plus de 80%) sur les données d'apprentissage




De même, l'histogramme de précision ci-dessous nous montre que le modèle a classifié correctement 78% des dossiers sur les données de test.




Les réseaux de neurones sont souvent comparés à une boite noire à juste titre car, en effet, il est très difficile de comprendre les calculs effectués durant le processus d'apprentissage. Néanmoins, il est possible en résultat à ces calculs d'estimer la sensibilité des variables d'entrées et ainsi de savoir quelles sont les variables les plus explicatives du phénomène à modéliser.

Dans notre cas, les variables *Solde du compte courant*, *Valeurs de l'épargne et Remboursement des prêts antérieur* ont été estimées comme les plus explicatives.

### Validation du Modèle

Afin de valider ce modèle, il a été estime l'erreur de classification sur un jeu de données de validation constitué de 50 'Bon' dossiers et 50 'Mauvais' dossiers.

La matrice de confusion, un outil de mesure de qualité d'un modèle de classification présentant le nombre de dossier classer correctement et incorrectement par un modèle, nous montre que notre réseau de neurones présente des résultant intéressant (taux d'erreur de 29%) :

| **Dossier de Prêt** | **Prévision du modèle de *Réseau de neurones* en tant que 'Mauvais' dossier** | **Prévision du modèle de *Réseau de neurones* en tant que 'Bon' dossier** |
|---|---|---|
| Observée 'Mauvais' | 39 | 11 |
| Observée 'Bon' | 18 | 32 |

### Réseaux de Neurones VS Régression Logistique

Il a été construit un modèle de régression logistique sur les mêmes données et les mêmes variables en entrées ayant servi à construire notre réseau de neurones.

La matrice de confusion obtenue en appliquant ce modèle de régression logistique nous montre que le modèle de régression logistique modélise de manière satisfaisante notre problématique mais avec un taux d'erreur légèrement supérieur à notre réseau de neurones (31% d'erreur pour la régression logistique contre 29% pour notre réseau de neurones) :

| **Dossier de Prêt** | **Prévision du modèle de *Régression logistique* en tant que 'Mauvais' dossier** | **Prévision du modèle de *Régression logistique* en tant que 'Bon' dossier** |
|---|---|---|
| Observée 'Mauvais' | 34 | 16 |
| Observée 'Bon' | 15 | 35 |