Presentation finale fouille de donnée

Eric Papain Diakité Djeila
Master SIM Institut Francophone International
Année universitaire 2018/2020

Plan de Travail

Introduction au data mining

1. Projet de data mining et la méthodologie CRISP

2. Introduction à notre jeux de donnée Crédit Scoring

a. Contexte et type d'application
b. problématique
c. importation de notre jeux de donnée dans STATISTICA
d. Compréhension des données

3. Représentation Graphique

a. Histogramme des Variables
b. NUAGE DE POINTS ENTRE Durée du crédit et Montant de l'Emprunt
c. Relations entre notre variable d'intérêt *DOSSIER DE PRÊT* et la variable *SOLDE DU COMPTE COURANT AU MOMENT DU PRÊT*
d. *Relations entre notre variable d'intérêt DOSSIER DE PRÊT et la variable SEXE.*
e. Relations entre la variable *Remboursement des Prêts Antérieurs* et la variable *Nombre de Prêts Antérieurs dans cette Banque.*

4. Nettoyage des données

a. Données éparses
b. Valeurs manquantes
c. Doublons
d. Traitement des données atypiques à l'aide de graphiques

Introduction au data mining

L'informatique, définit comme étant la science du traitement automatique et rationnelle de l'information à l'aide d'un ordinateur, est une des sciences les plus évolutives de part son extension et son adaptation dans presque tous les domaines de la vie tels que l'économie, la santé, la circulation, les statistiques, les prévisions, les simulations et bien d'autres domaines encore.

Cette science n'a cessé de surprendre depuis son avènement aujourd'hui nous ne sommes plus juste au niveau du traitement automatique mais aujourd'hui nous parlons déjà du traitement intelligent par les ordinateurs et aussi à des prédictions grâce à des algorithmes existant dans plusieurs domaines car on assiste à des situation ou la machine décide d'apprendre et d'aider l'Homme dans le processus de prise de décision de communiquer avec des autres de se former de raisonner.

Dans le contexte actuel, nous assistons des fois a des situation où l'Homme veut anticiper sur un certain grand nombres de prise de décision en utilisant certaines informations existant, mais pourtant peuvent prédire en utilisant les information actuelle sur son environnement que se soit physique ou psychologique.cet état actuelle à poussé plein de chercheurs et d'ingénieurs a ouvrir une nouvelle voie de recherche qui est celle de la statistiques avancée plus connu sous le nom de DATA MINING ou encore ANALYSES DE DONNÉES.

Ainsi donc, il existe 03 grands types d'applications en Data mining

- Problèmes de Classification

par exemple prédire si un patient présente un fort risque de maladie cardiaque ou pas. La classification auras pour rôle trouver les variables présentant un lien fort avec la variable d'intérêts et de construire un modèle prédictif avec ces variables afin de classifier la variable d'intérêt.

-Problèmes de Régression

par exemple la mesure d'un procédé de fabrication industrielle.

-Problèmes de Segmentation(clustering)

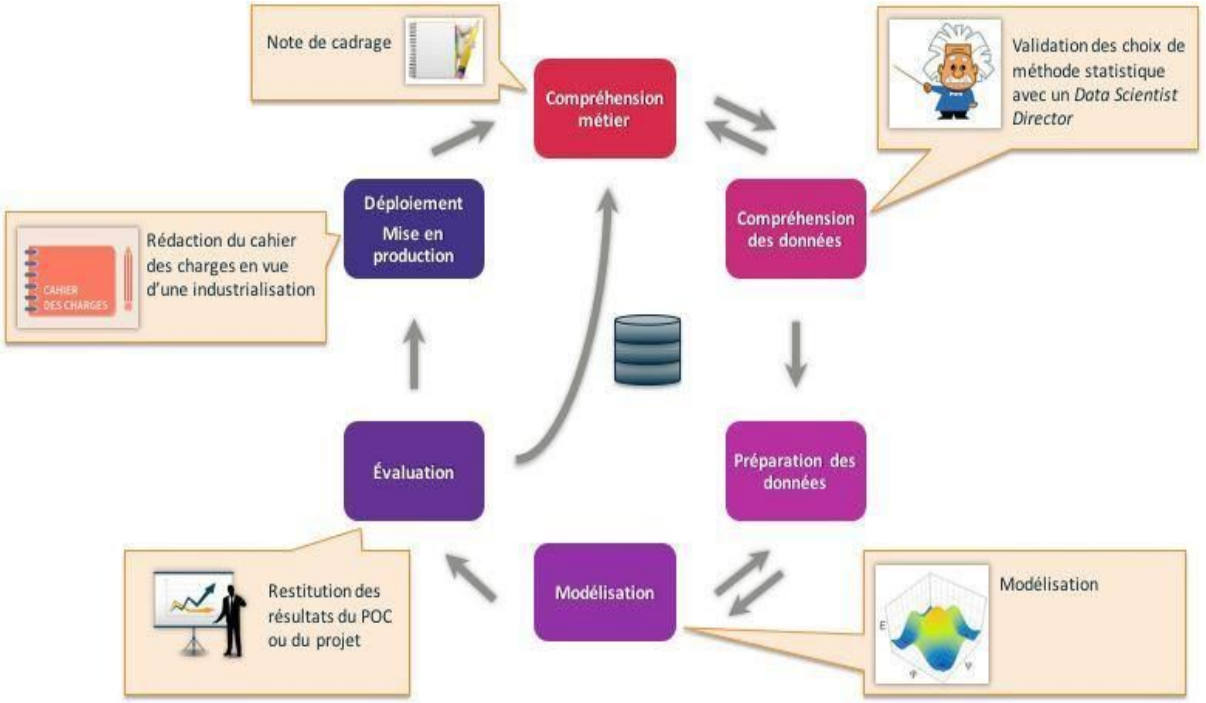
Dans ce cas il n'existe pas de variable d'intérêts. Les variables entrée servent à créer des groupes afin d'affecter chaque nouvel élément à un groupe donnée.

1. Projet de data mining et la méthodologie CRISP

CRISP(*Cross Industry Standard Process*) est une méthodologie d'analyse de donné, elle défini les étapes à suivre pour mener à bien un projet ces différentes étapes comprenant différentes étapes auxquels nous pouvons citer:

- Compréhension de la problématique
- Compréhension des données
- Préparation des données
- Modélisation
- Évaluation
- Déploiement

Faut noter que cette méthodologie d'analyse de donnée s'applique à tous les domaines d'activité indépendamment du logiciel de data mining utilisé.



Nous utiliserons dans notre contexte les outils d'analyse de donnée STATISTICA Data Miner

2. Introduction à notre jeux de donnée Crédit Scoring

a. Contexte et type d'application

Dans notre contexte, il s'agit d'une institution financière qui possède des données relatives à ses clients et aimerais les utiliser pour prédire le risque potentiel de crédits liée à ses futur client afin de savoir de combien ou bien si le crédit pourrait être accorder ou non.

Pour mener à bien notre études, nous suivrons un certains nombres d'étapes qui nous mèneras jusqu'à la production d'un modèle prédictif validé pour ce jeux de données précis.

b. problématique

Cette institution financière possède des données relatives à ses clients,Les clients sont classifiés comme ayant un « Bon » ou un « Mauvais » dossier de prêt.

- Dans notre contexte nous avons a faire a un problème de **Classification** dans le domaine de **l'apprentissage supervisé** qui utilise des algorithme d'intelligence artificielle pour la resolution de probleme statistique.

- **<u>REMARQUE ET INDICATION SPÉCIFIQUE À L'INSTITUTION</u>**
  - un emprunt se situe entre 300 et 30 000 usd
  - la durée de l'emprunt est de 3 a 52 mois
- **Notre Objectif** : prédire le risque de crédit des futurs emprunteurs de m'octroyer un crédit qu'aux emprunteurs présentant un faible risque et un fort potentiel

- **Solution :** pour répondre efficacement à notre problématique nous essayerons de modéliser la variable cible par l'algorithme des **Forêts Aléatoires.**

Nous divisons nos donnee en deux bloc 20% pour les test et 80% pour l'apprentissage et la formation du modèle.

c. importation de notre jeux de donnée dans STATISTICA

Dans cette partie, nous allons simplement convertir notre jeux de donnée du fichier csv vers un fichier excel que nous allons juste ouvrir dans notre outils d'analyse STATISTICA.

d. Compréhension des données

Notre jeux de donnée est disponible :
https://github.com/ericpapain/data-analysis-with-STATISTICA-DATA-MINER.git

Notre Jeux de donnée comporte 19 attributs et 1003 observations, elle est composé de variable continue et catégorielle:

ii. **Dossier de Prêt :** elle correspond à notre variable d'intérêt celle qu on souhaite prédire
iii. **Solde du Compte Courant :** solde dans le compte avant la demande de prêt
iv. **Durée du Prêt** : durée sur lequel le prêt doit être étendu
v. **Remboursement des Prêts Antérieurs :** représente les prêt antérieur rembourser
vi. **Objectif du Prêt :** raison de demander le prêt
vii. **Montant du Prêt** : montant du prêt demander par le client
viii. **Valeur de l'Épargne :** montant de l'épargne dans son compte avant le prêt
ix. **Ancienneté chez l'Employeur Actuel :** nombre d'année chez son employeur actuel

x. **Endettement en % du Revenu Disponible**

xi **Statut Marital : marié ou non**

xi. **Sexe**

xii. **Ancienneté dans le Ménage Actuel**

xiii. **Actifs les Plus Importants**

xiv. **Âge**

xv. **Autres Prêts en Cours**

xvi. **Type de Logement**

xvii. **Nombre de Prêts Antérieurs dans cette Banque**

xviii. **Occupation**

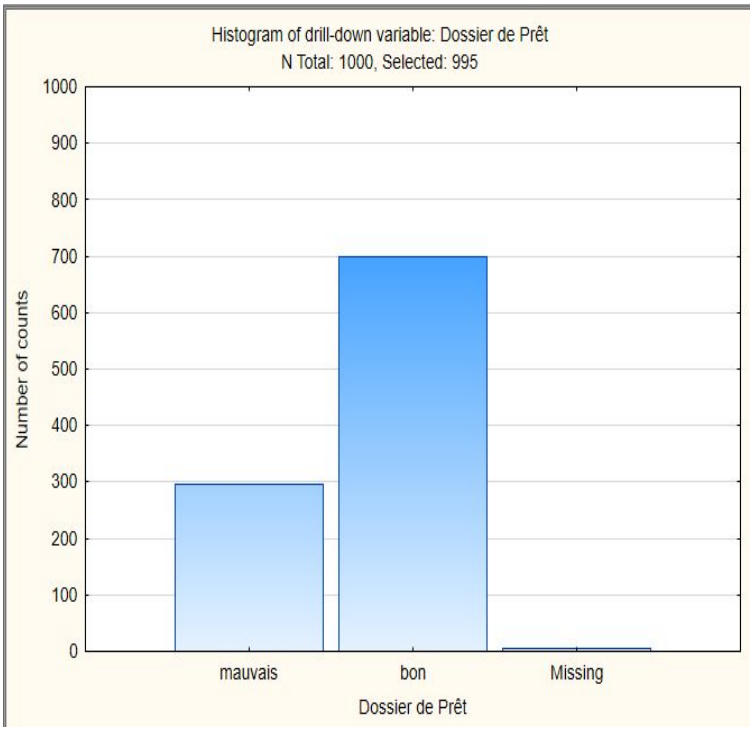
xix. **Ensemble :** représente la séparation de donnée (test, entraînement, validation)

3. Représentation Graphique de certain paramètre de notre jeux de donnée

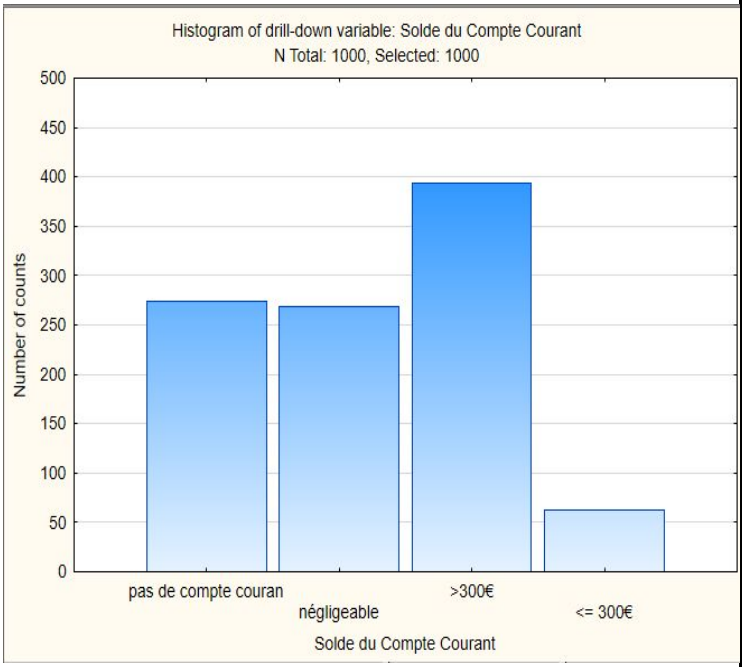
a. Histogramme des Variables

Ici nous utiliserons le module **Drill Down** de **STATISTICA** l'importance de ce module est qu'il laisse afficher la proportion des valeurs manquantes. on les résultats suivants :

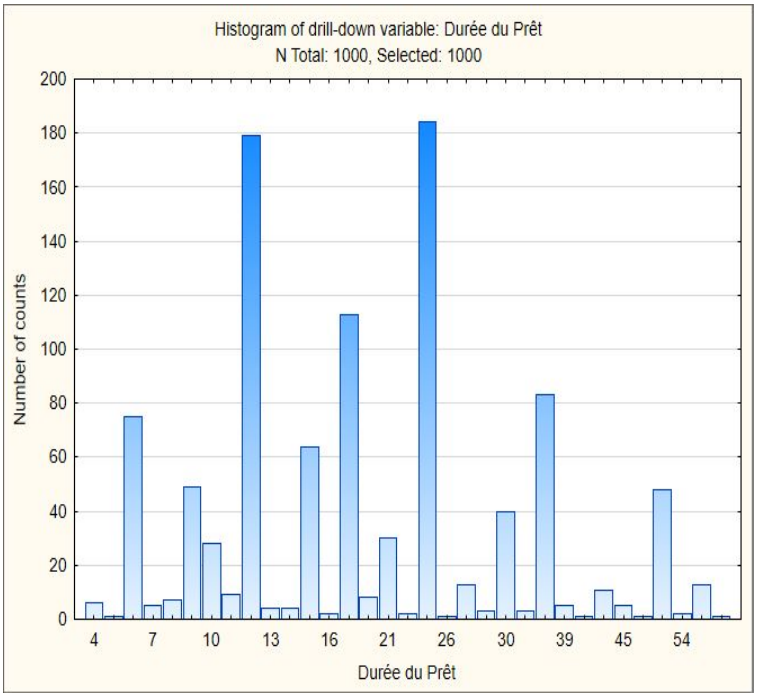
Dossier de Prêt



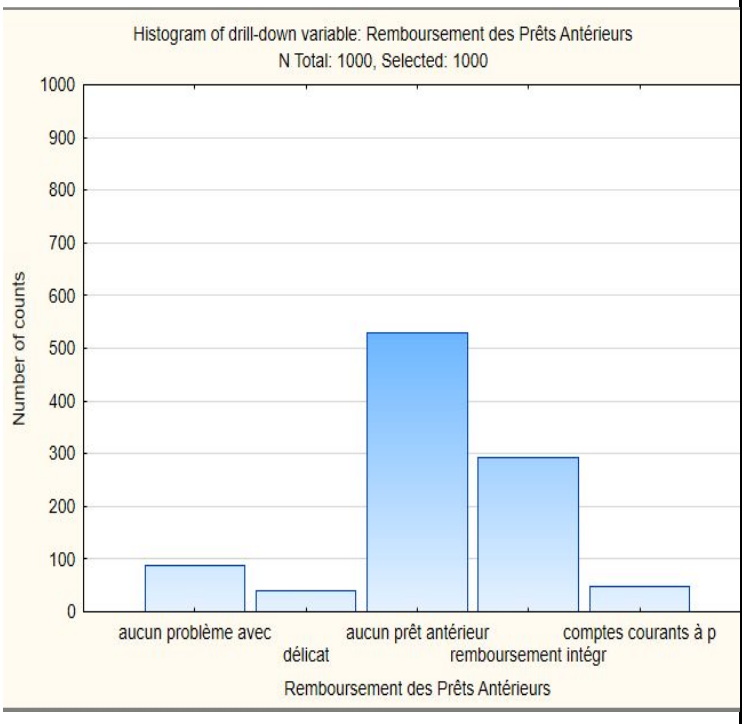
Solde du Compte Courant



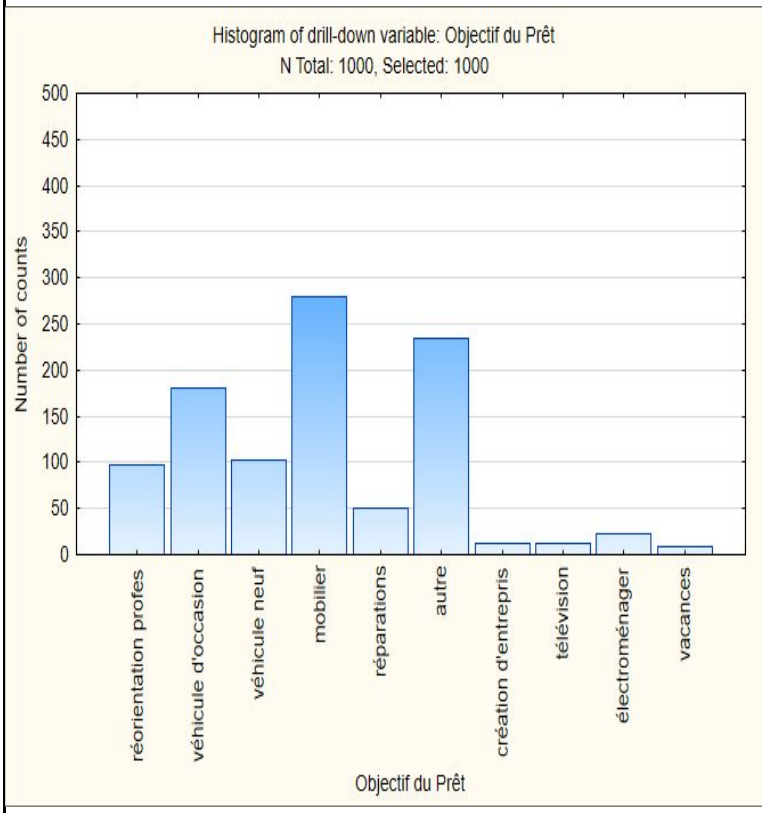
Durée du Prêt



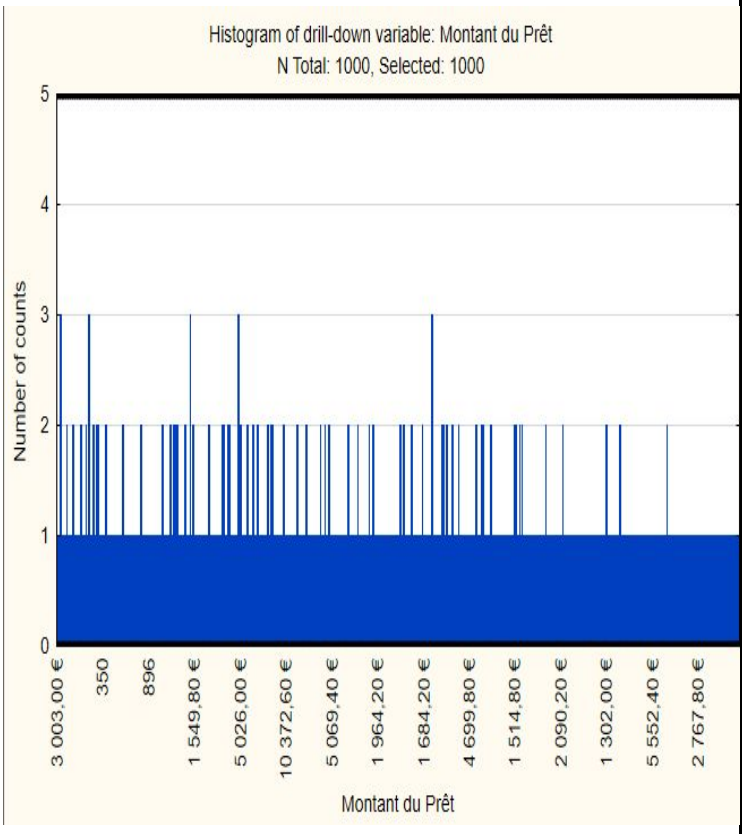
Remboursement des Prêts Antérieurs



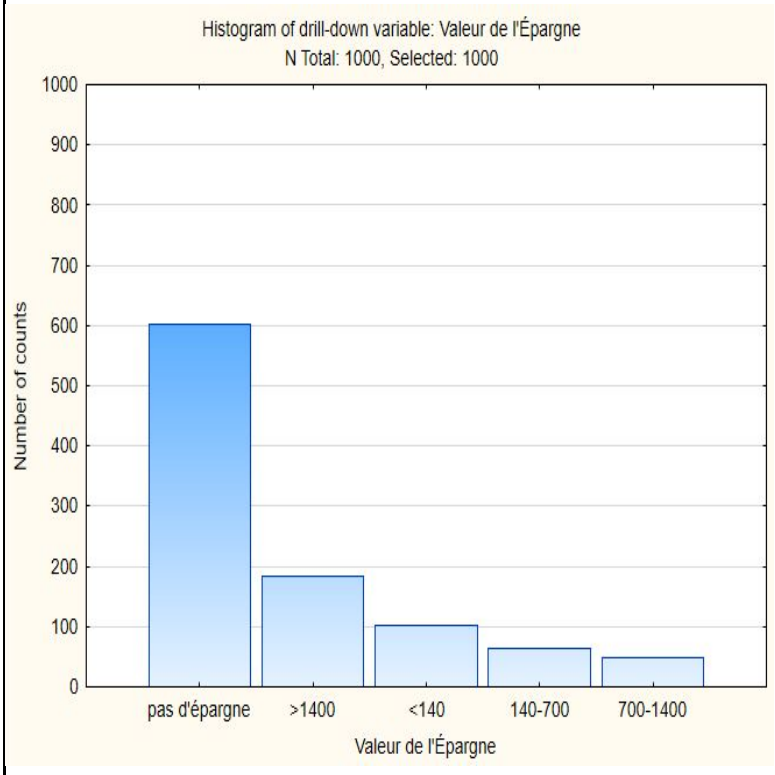
Objectif du Prêt



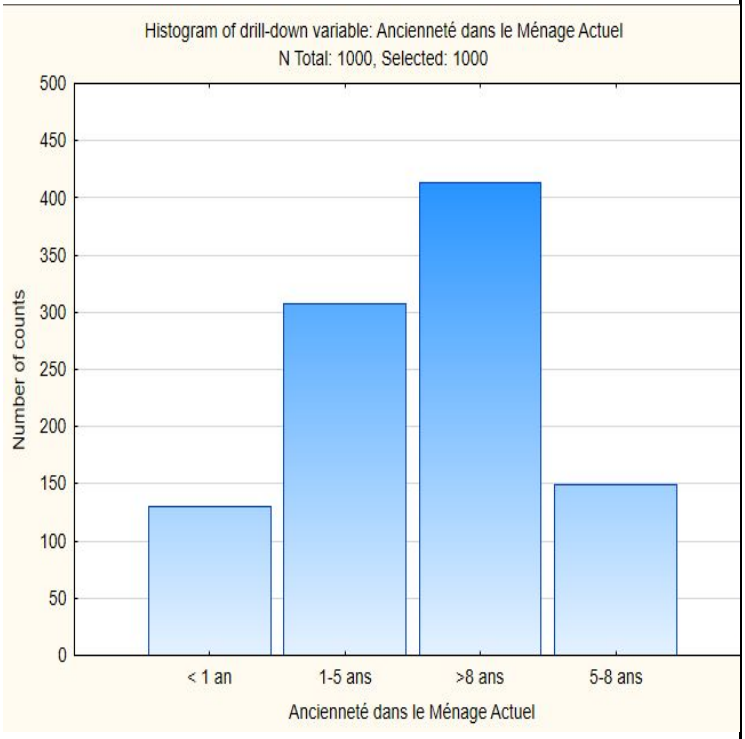
Montant du Prêt



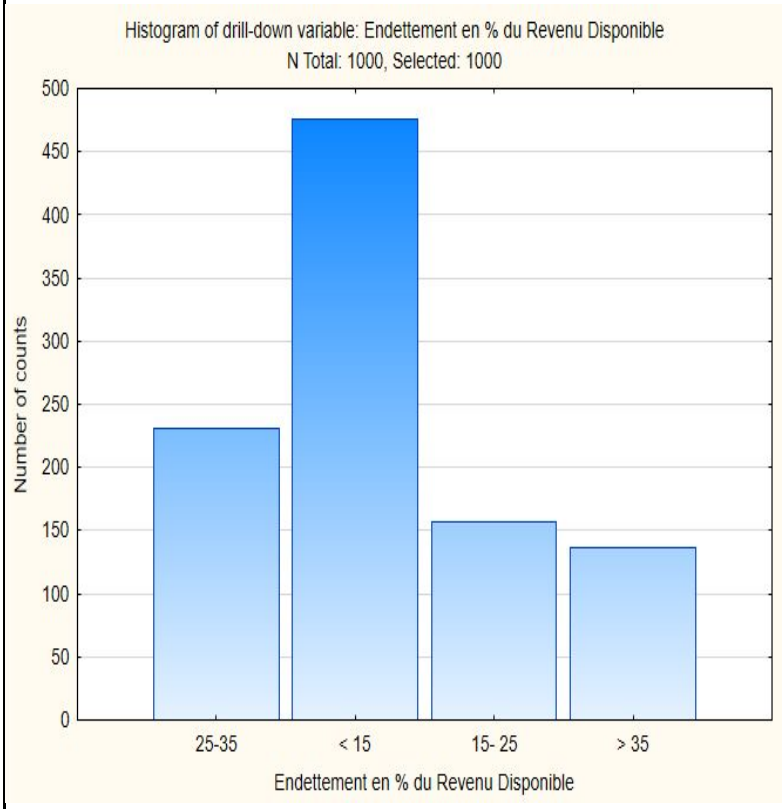
Valeur de l'Épargne



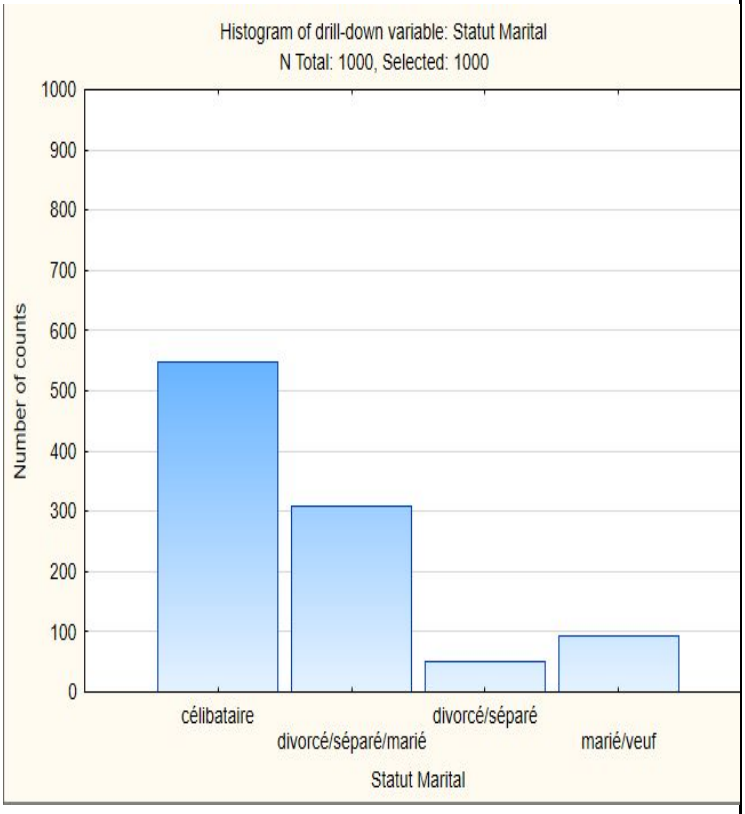
Ancienneté chez l'Employeur Actuel



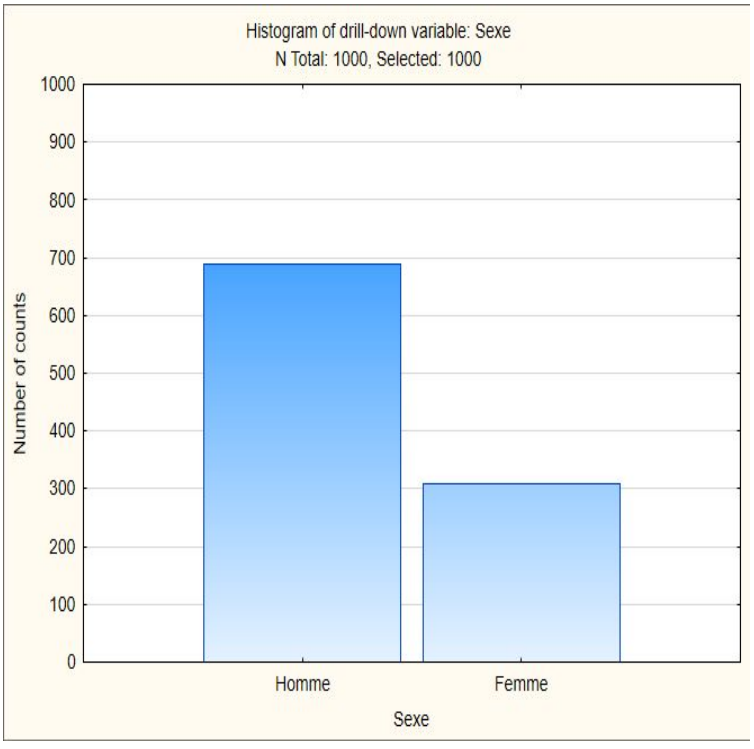
Endettement en % du Revenu Disponible



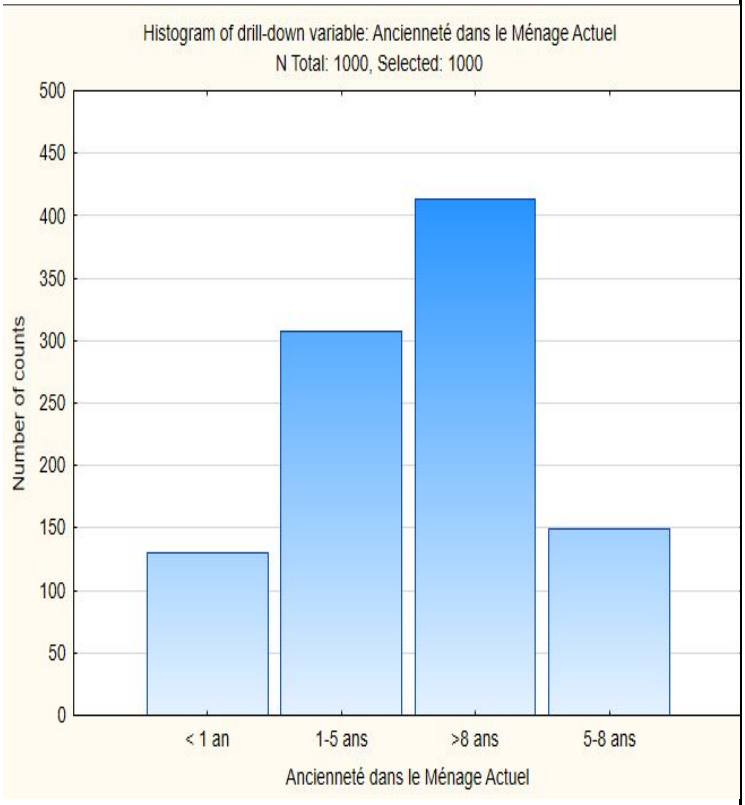
Statut Marital

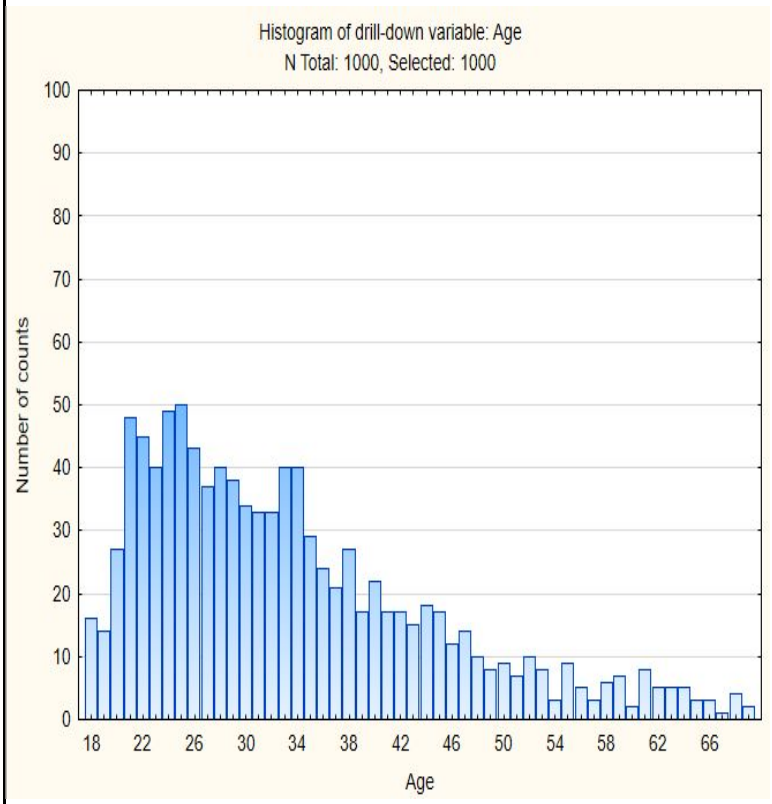


Sexe

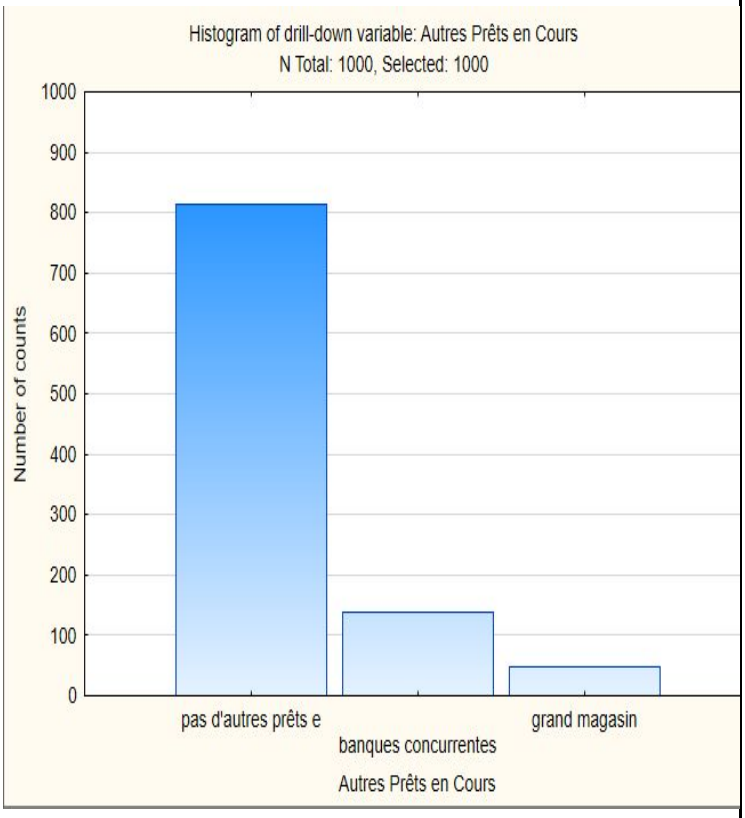


Ancienneté dans le Ménage Actuel

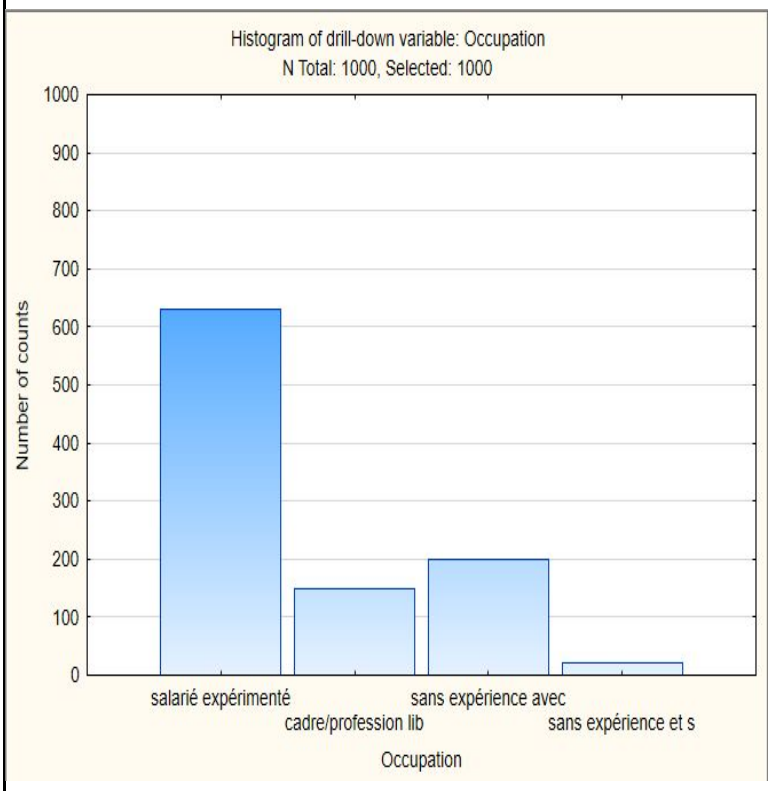




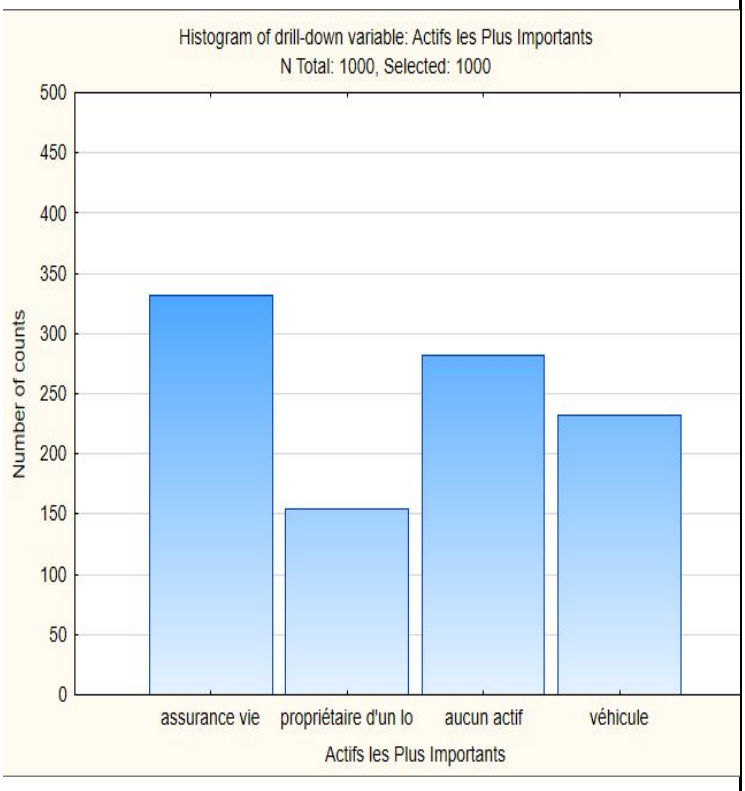
Autres Prêts en Cours



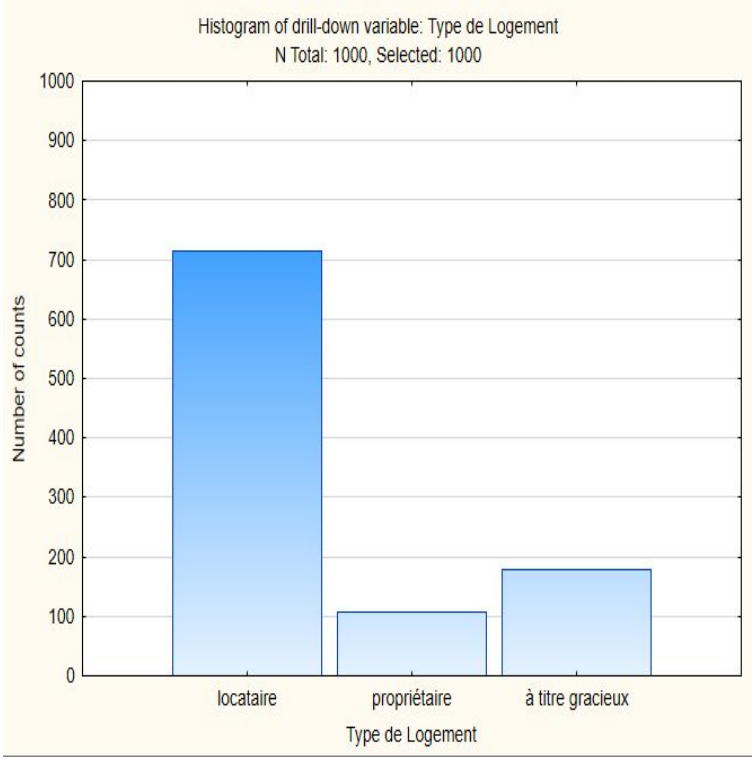
Occupation



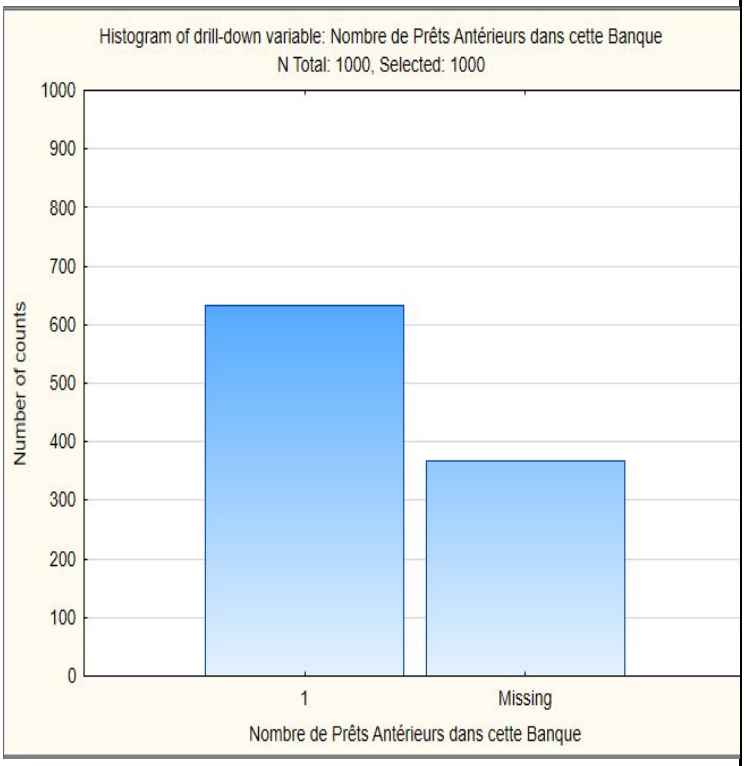
Actifs les Plus Importants



Type de Logement



Nombre de Prêts Antérieurs dans cette Banque

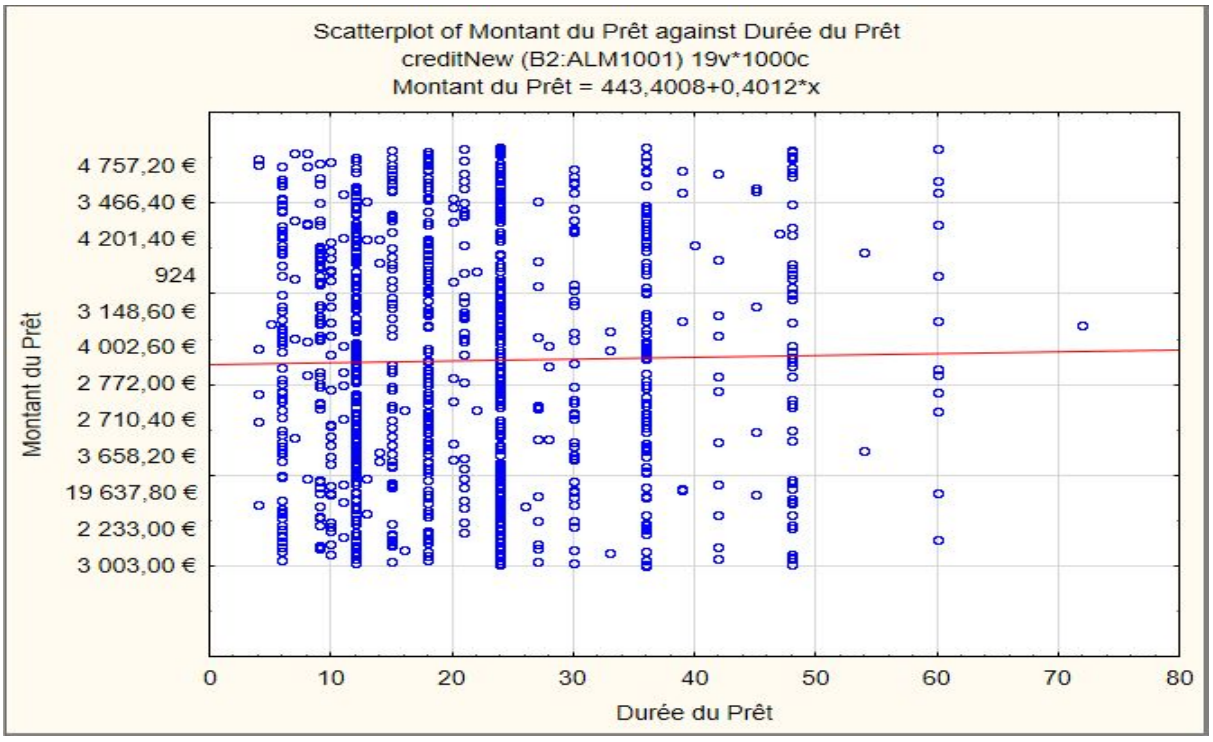


b. NUAGE DE POINTS ENTRE **Durée du crédit** et **Montant de l'Emprunt**

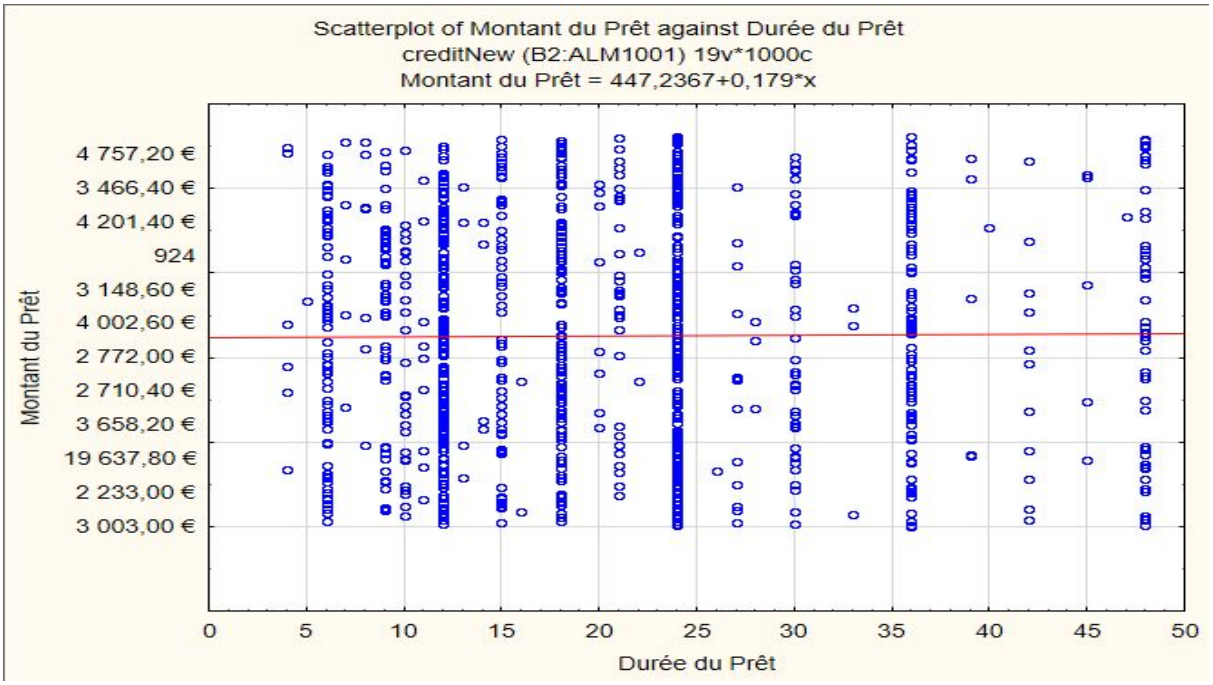
Pour ce faire nous allons sélectionner cette option dans notre outils statistica.

Rappel:

- un **emprunt** se situe entre **300 et 30 000 usd**
- la **durée de l'emprunt** est de **3 a 52 mois**



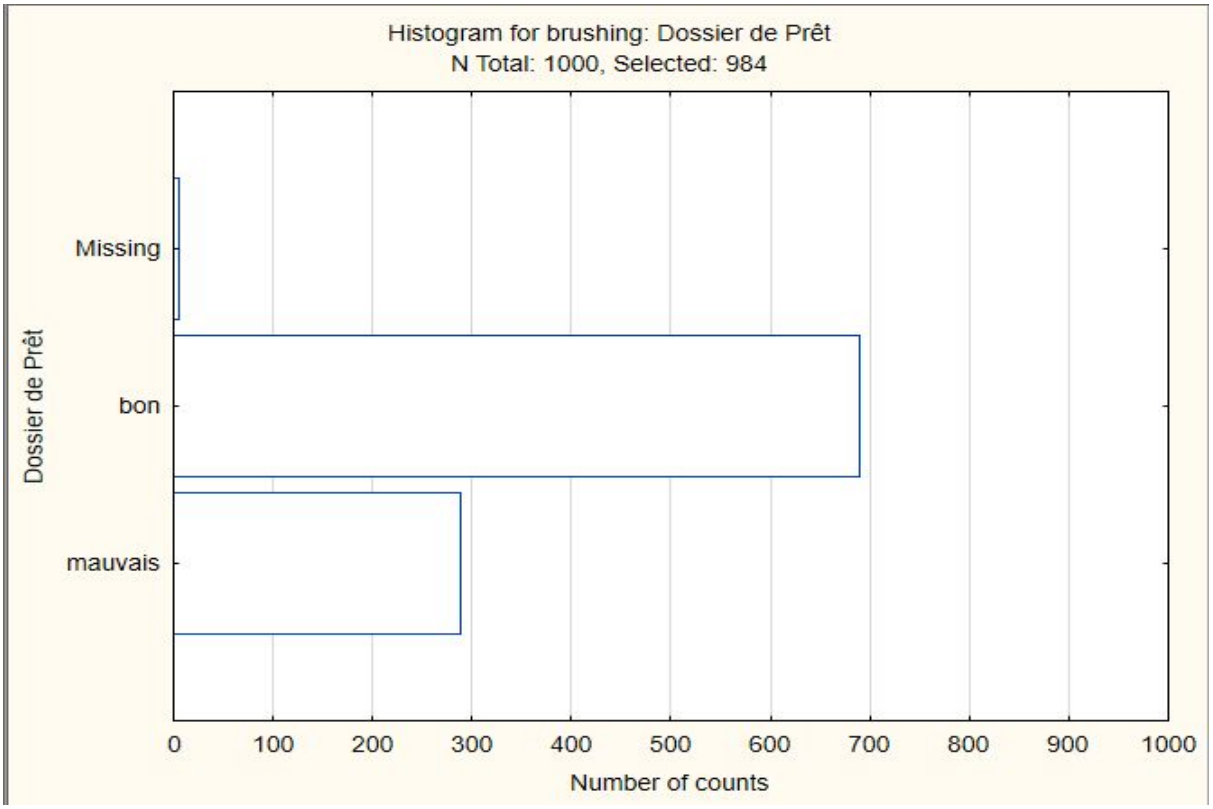
En regardant le nuage de point associé au duree de pret et au montant total nous constatons que, il ya certain point qui ne respecte pas les conditions fixés c'est à dire un **emprunt** se situe entre **300 et 30 000 usd et** la **durée de l'emprunt** est de **3 a 52 mois.** nous allons ce sujet les retirer de l'analyse en utilisant l'option de balayage que fournit STATISTICA pour les supprimer tous respectivement pour montant de l'emprunt et la durée de l'emprunt. on obtient après suppression le nuage suivant respectant les propriétés :



nous obtenons le nuage de point ci dessus après suppression des points aberrants.

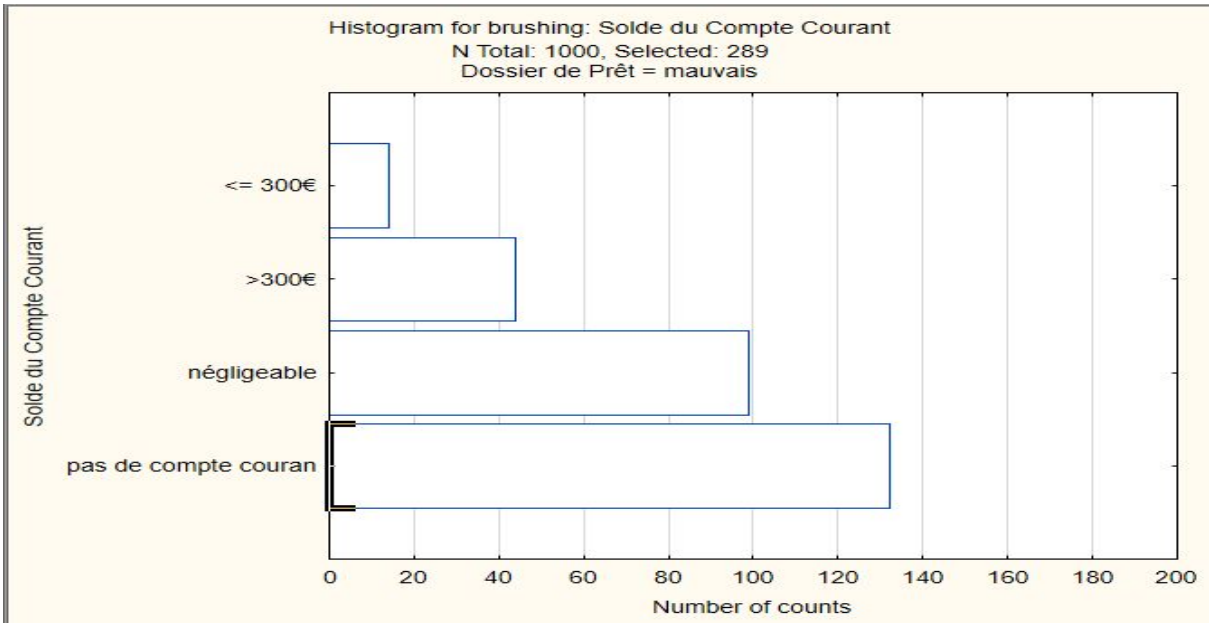
Nous venons de supprimer de nos données toutes les informations aberrantes.

c. Relations entre notre variable d'intérêt ***DOSSIER DE PRÊT*** et la variable ***SOLDE DU COMPTE COURANT AU MOMENT DU PRÊT***

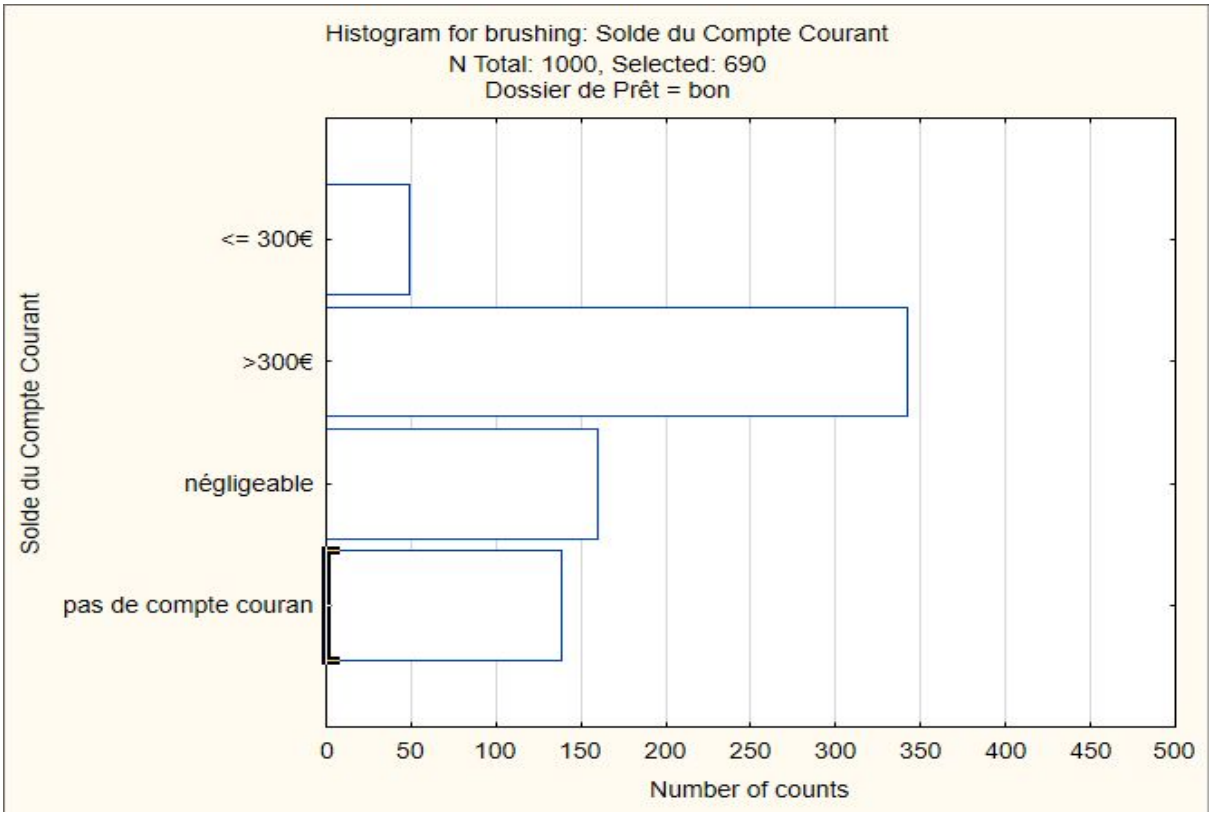


Voici la répartition de notre variable d'intérêt **DOSSIER DE PRÊT.**

Remarque : Dans STATISTICA, en utilisant l'option de balayage, nous pouvons voir la modalité pour les **mauvais dossier de prêt** concernant le ***SOLDE DU COMPTE COURANT AU MOMENT DU PRÊT*** . on peut le voir sur la representations ci dessous.

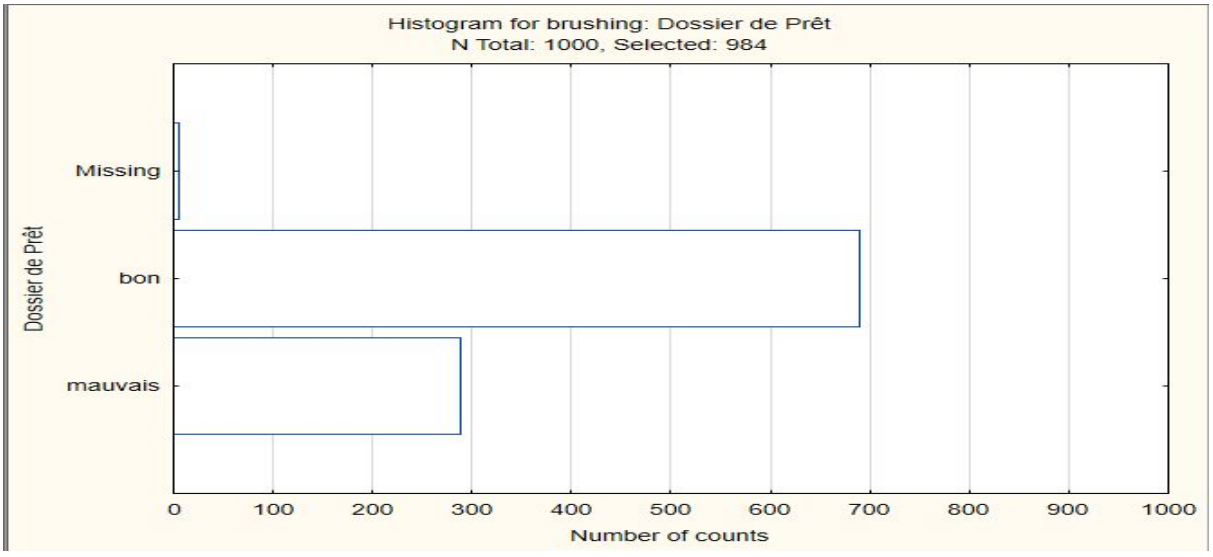


Remarque 1: D'après cet histogramme, nous constatons que la majorité de personne ayant ***un mauvais dossier de prêt*** sont ceux n'ayant ***pas de compte courant*** et ou sont des personnes donc ***le solde du compte courant est négligeable.***

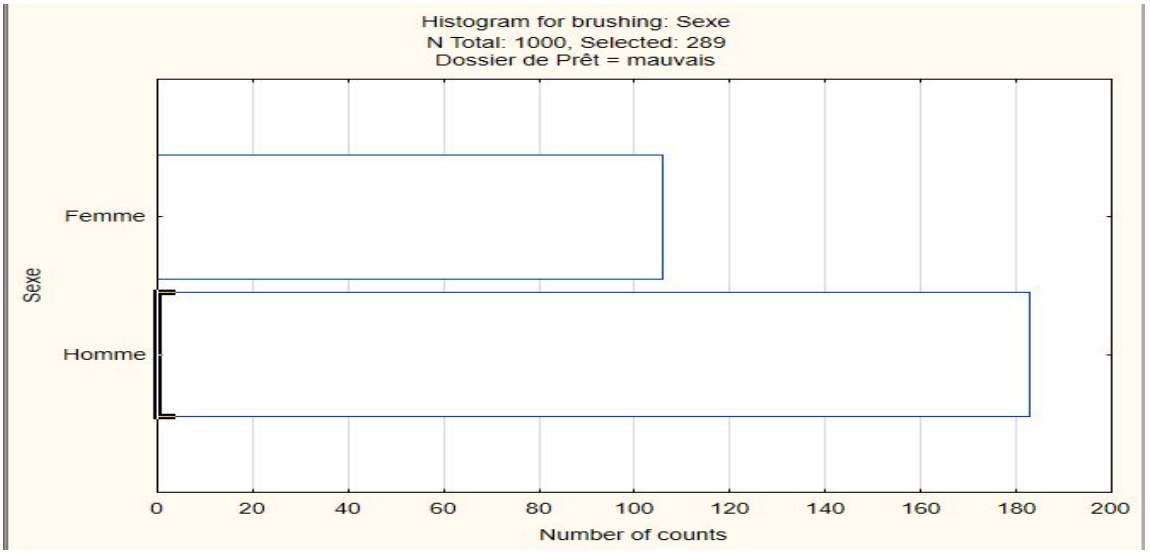


Remarque 2: De même, d'après l'histogramme des dossier de prêt représenter plus haut, nous constatons également que la majorité de personne ayant un bon dossier de prêt sont ceux ayant plus de 300 euro dans leurs compte courant c'est à dire un solde non négligeable sur leurs comptes courant et ou sont des personnes donc le solde du compte courant est négligeable.

d. Relations entre notre variable d'intérêt ***DOSSIER DE PRÊT*** et la variable ***SEXE.***

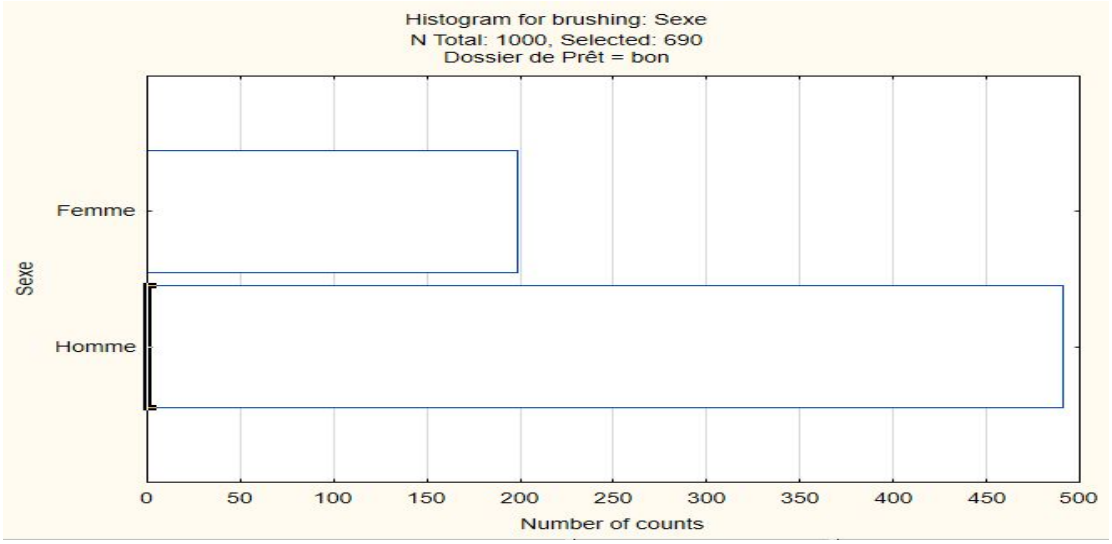


Remarque 1: en utilisant toujours l'histogramme en broche de notre dossier de prêt associé au ***SEXE***, on obtient:



Remarque 1: D'après cet histogramme ci-dessus, nous constatons que la majorité de personne ayant ***un mauvais dossier de prêt*** sont des **Hommes.**

De même en regardant la modalité des personne ayant le bon dossier on a l'histogramme suivante:

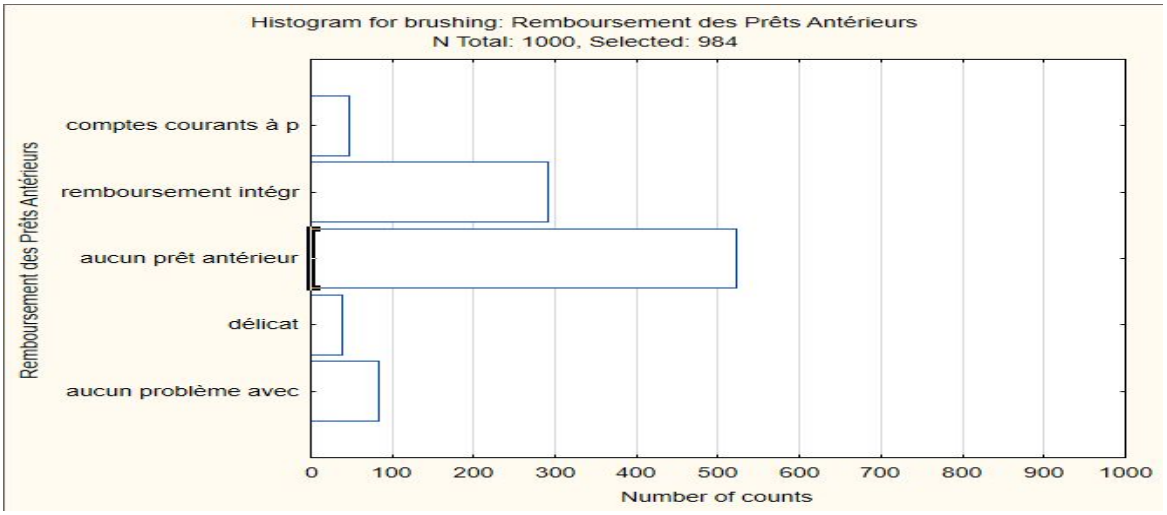


Remarque 2: D'après cet histogramme, nous constatons que la majorité de personne ayant ***un mauvais dossier de prêt et ceux ayant un bon dossier de prêt*** on presque les même modalité c'est à dire son sensiblement identique pour les bons et les mauvais dossier de prêt.

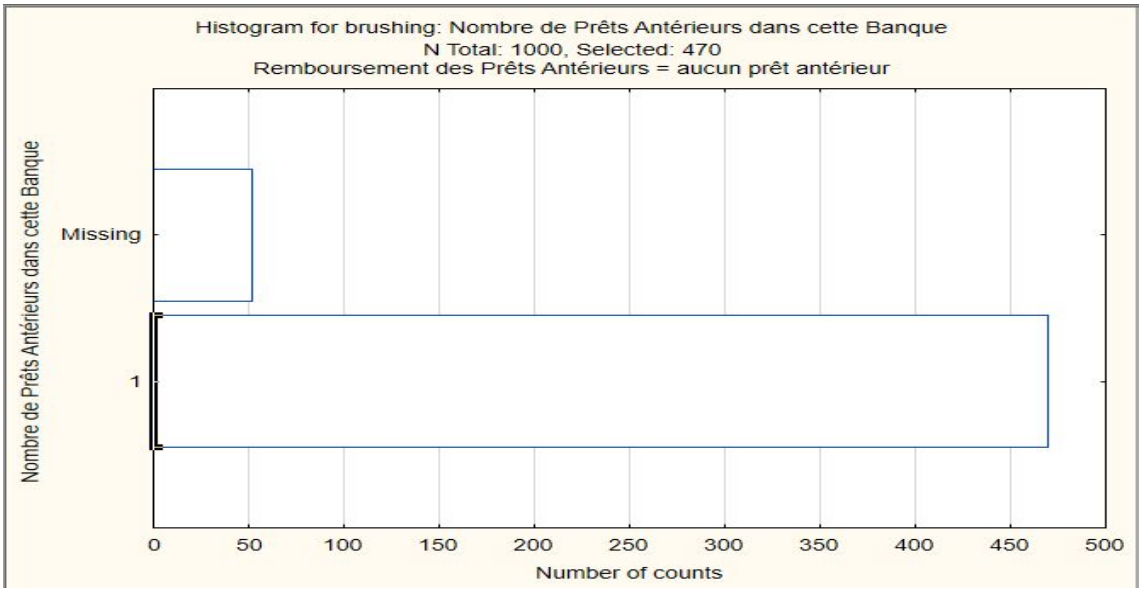
conclusion: la variable SEXE ne semble donc pas très pertinente pour discriminer notre variable d'intérêt ***DOSSIER DE PRÊT***.

e. Relations entre la variable ***Remboursement des Prêts Antérieurs*** et la variable ***Nombre de Prêts Antérieurs dans cette Banque.***

en utilisant l'histogramme de relation entre notre nombre de prêt antérieurs dans cette banque; on obtient:



D'après ce qui précède nous avons constaté que la majeure partie des client dans l'analyse effectuée plus haut n'avais jamais effectuer de prêt auparavant.



en regardant la répartition du nombre de prêt dans cette population, on constate que plusieurs d'entre eux aurait effectué plusieurs prêt dans cette institution ce qui ***est très incohérent.***

4. Nettoyage des données

a. Données éparses

dans cette section nous traiterons les valeurs manquantes dans notre jeux de données.

a ce sujet nous allons utiliser notre outils et le résultat obtenue est le suivant après suppression des donnée éparse on obtient des généralement dans statistica la suppression des éléments ayant plus de 10% de valeurs manquante après application du filtrage, notre jeux de donnée est resté inchangé en terme de nombre de variable mais par ailleurs deux observations on été supprimer.

b. Valeurs manquantes

pour le traitement de nos valeurs manquantes, dans nos donnée plus haut nous avons plutôt des valeurs manquantes au niveau de dossier de prêt et de puisque si nous ne les traitons pas les algorithme de data mining vont simplement les ignorer, alors dans notres cas de figure nous avons utilisé la méthode de remplacement par la moyenne qui est une fonction fournie par statistica data mining.

après ce traitement nous obtenons une nouvelle fiche d'analyse de données.

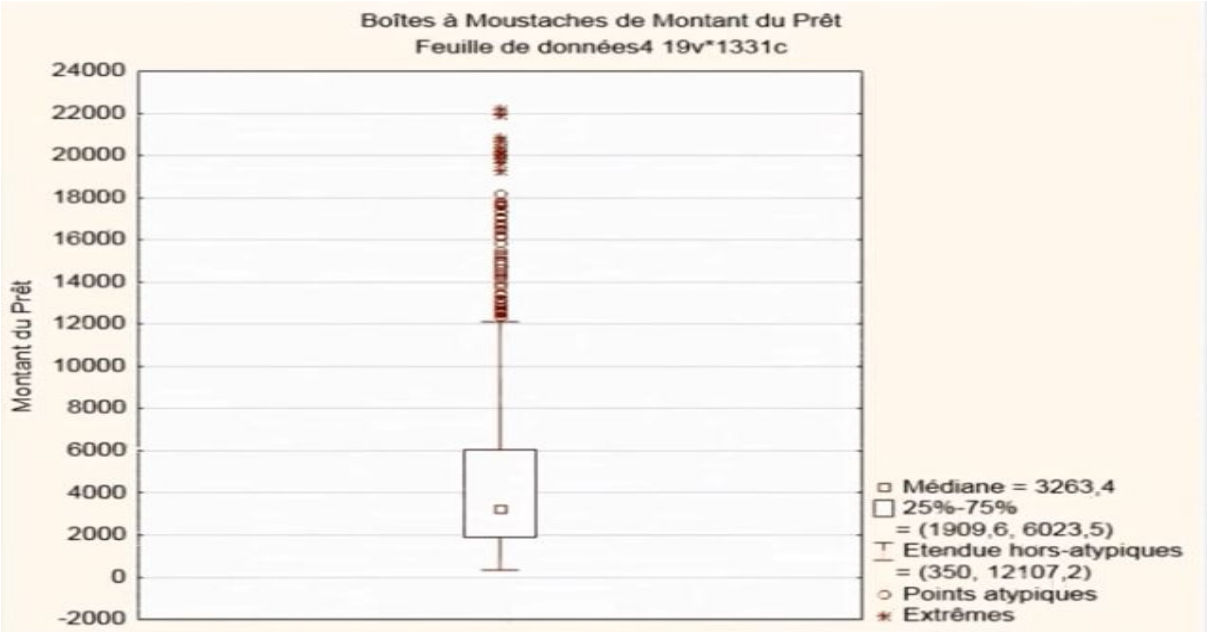
c. Doublons

Elle correspondent au valeurs doubler dans notre cas toujours dans l'onglet de traitement de donnée, nous pouvons observer 3 suppression de notre jeux de donnée.

d. Traitement des données atypiques à l'aide de graphiques

puisque nous manipulons les valeurs continues, nous utiliserons dans notre cas les boite a moustaches.

En examinant les nouvelle donnée, nous constatons qu'il subisse un certain grand nombre de point atypique voir extrême mais ces données sont légitime car répondent au critère donner pour le montant de l'emprunt. on peut le voir dans le schéma de la boite a moustache ci dessous.

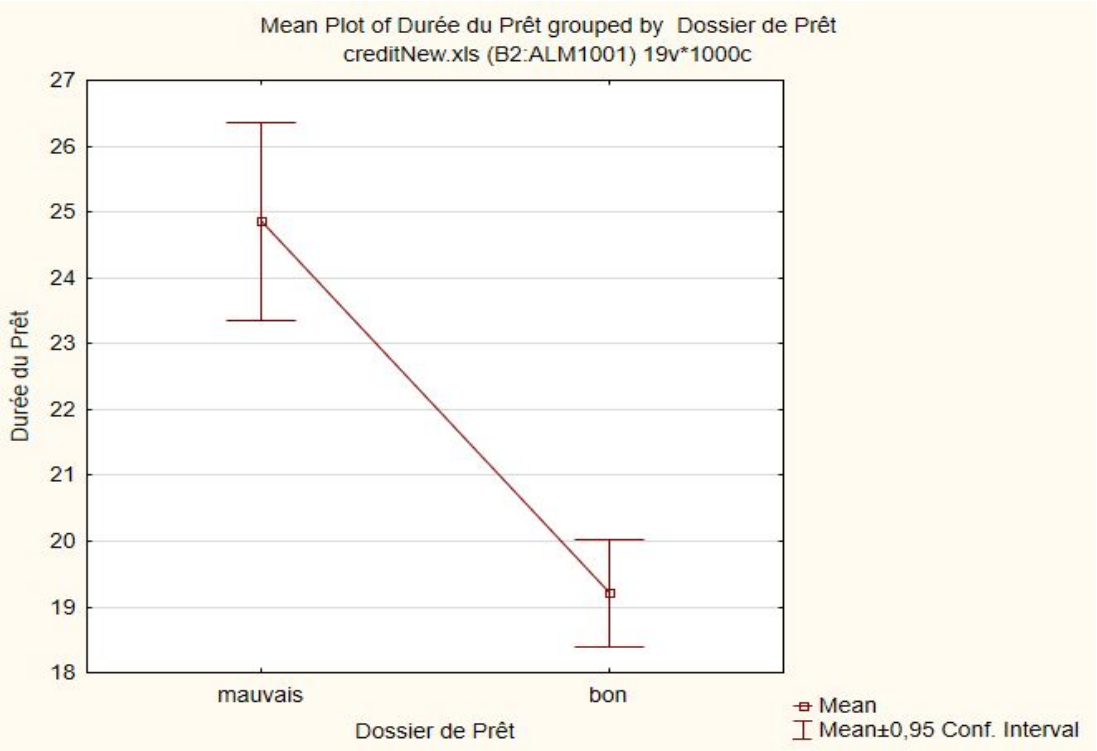


5. Exploration Graphique

Nous avons précédemment nettoyer et fiabilisé nos donnée avec différentes procédure comme le traitement des valeurs manquante et des valeurs aberrantes a présent nous disposons d'une base propre et fiable que nous explorerons graphiquement afin de visualiser, de découvrir les relations qui existe entre les différentes variables et notamment avec notre variable cible la variable dossier de prêt qui traduit en fait le risque de défaut.

➔ LIEN ENTRE LA VARIABLE CONTINUE **DURÉE DE PRÊT** ET LA VARIABLE CATÉGORIELLE **DOSSIER DE PRÊT**

Pour se faire nous construirons à ce sujet un diagramme de moyenne pour faire ressortir le lien entre ces deux variables.



on constate que la moyenne de mauvais dossier est plus élevée que celle des bon dossier ce qui peut nous faire comprendre intuitivement que la qualité du dossier est influencée aussi par la durée de prêt. Donc nous pouvons par la tirer la conclusion les durée plus élevée sont les durée plus à risque.

6. Échantillonnage des données

Dans cette partie nous nous focalisons à la répartition de nos donnée en données **d'apprentissage, de test et de validation**.

a. Échantillons d'apprentissage, de test et de validation

l'échantillon d'apprentissage permet de créer des modèle de data mining, **les données de test** permettent et interviennent à la phase de construction pour se rassurer du bon entrainement du modèle et n'intervient pas dans la phase de **validation du modèle.**

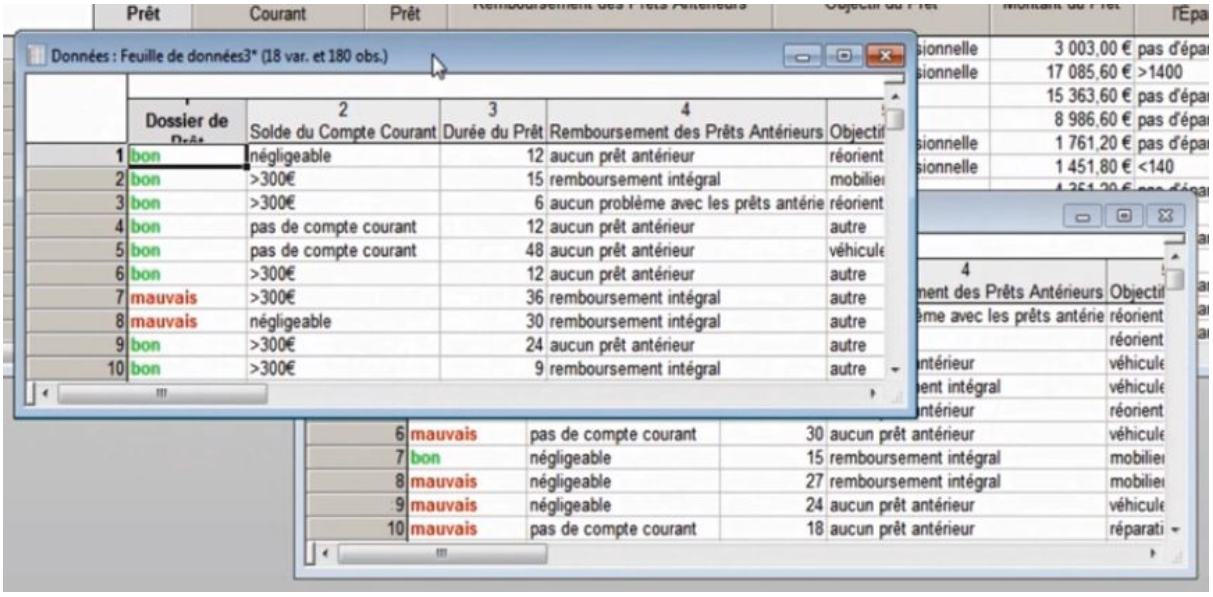
Dans notre cas nous allons utiliser un modèle **d'échantillonnage aléatoire** fourni par STATISTICA pour repartir nos données en 2 échantillons,

- d'une part des données qui serviront à la construction du modèle, avec des donnée **d'apprentissage et de test**
- et d'autre part des données pour **la validation.**

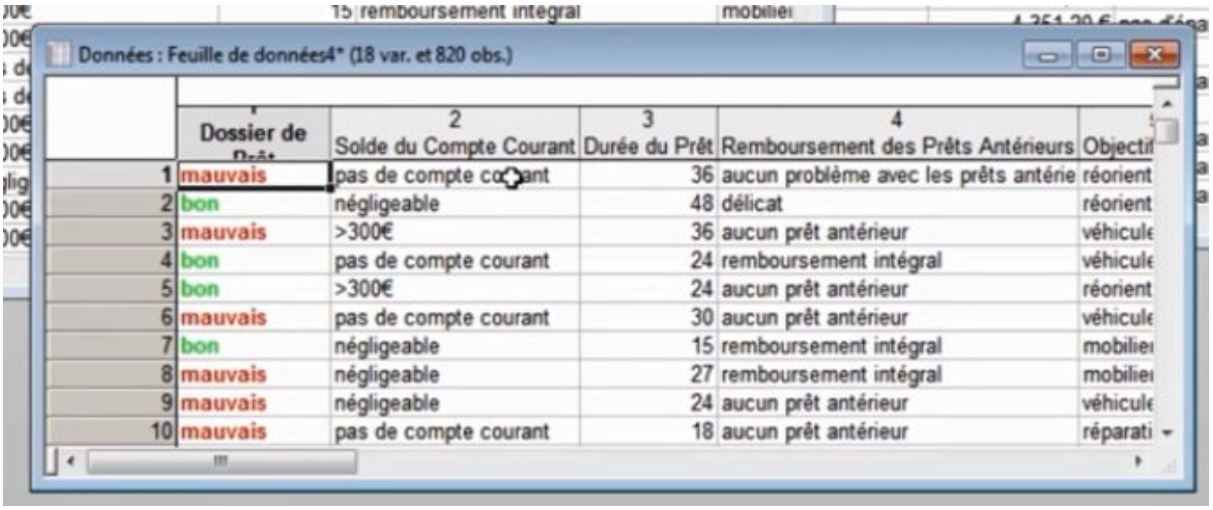
b. taille de l'échantillon

cela se fait dans statistica en allant dans l'onglet ***donnée***

- ***nous spécifions ici que nous souhaitons que l'échantillon de validation représente 15% des donnée***
- ***l'échantillon de construction et de test constituent donc 85% de données .***



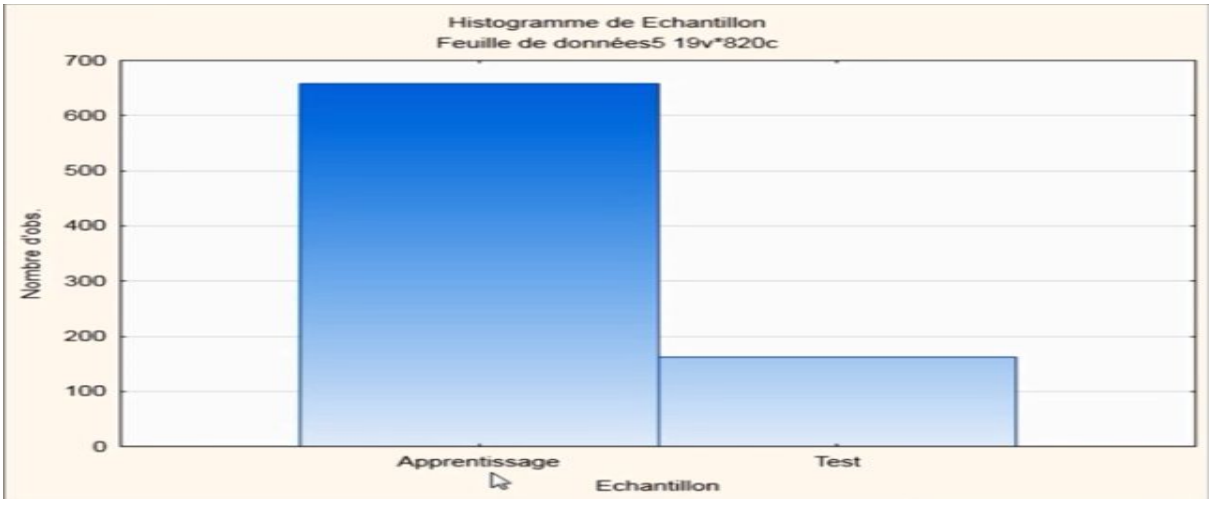
nous constatons que les donnée de validation contiennent environ 180 observations comme nous le voyons sur sur les capture ci-dessus se qui représentent environs 15% de nos données. ces données serviront lors de la phase de validation et serviront à dire dans quel condition ses modèle doivent être généralisé.



	Dossier de Prêt	2 Solde du Compte Courant	3 Durée du Prêt	4 Remboursement des Prêts Antérieurs	Objectif
1	mauvais	pas de compte courant	36	aucun problème avec les prêts antérie	réorient
2	bon	négligeable	48	délicat	réorient
3	mauvais	>300€	36	aucun prêt antérieur	véhicule
4	bon	pas de compte courant	24	remboursement intégral	véhicule
5	bon	>300€	24	aucun prêt antérieur	réorient
6	mauvais	pas de compte courant	30	aucun prêt antérieur	véhicule
7	bon	négligeable	15	remboursement intégral	mobilier
8	mauvais	négligeable	27	remboursement intégral	mobilier
9	mauvais	négligeable	24	aucun prêt antérieur	véhicule
10	mauvais	pas de compte courant	18	aucun prêt antérieur	réparati

par ailleurs les 85% de donné restante aussi se trouve aussi dans une autre feuille de donnée comportant comme nous le voyons dans l'entête près de 820 observations, ces donnée seront divisés en échantillon d'apprentissage et en échantillon de test et utiliser pour la modélisation.comme nous le voyons plus haut.

nous allons ainsi utiliser par la suite une formule nous permettant de séparer nos 85% en deux échantillon d'apprentissage et ne échantillon de test ce qui permet de répartir 20% de données pour les tests et 80% des 85% pour l'apprentissage, il faut noter que ces 20% de donnée de test permettent d'éviter un sur apprentissage c'est à dire que le modèle apprennent uniquement sur des mêmes valeurs comme nous pouvons le voir dans l'histogramme d'échantillonnage ci-dessous:

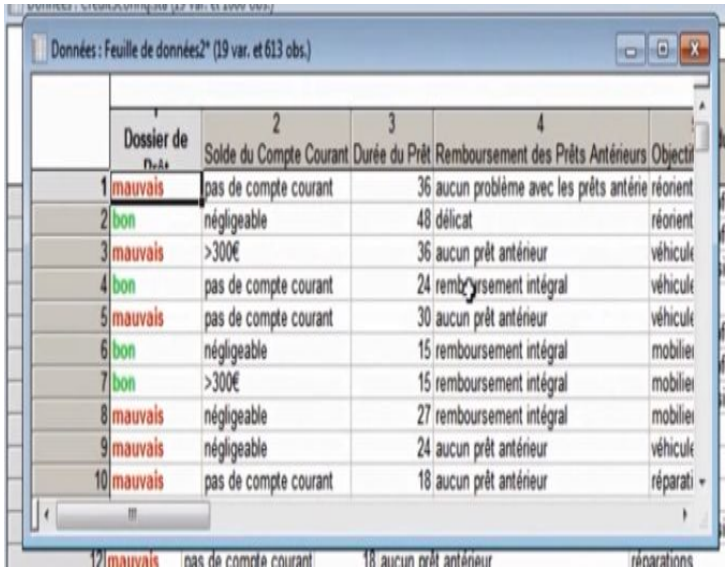


c. Échantillonnage aléatoire stratifié

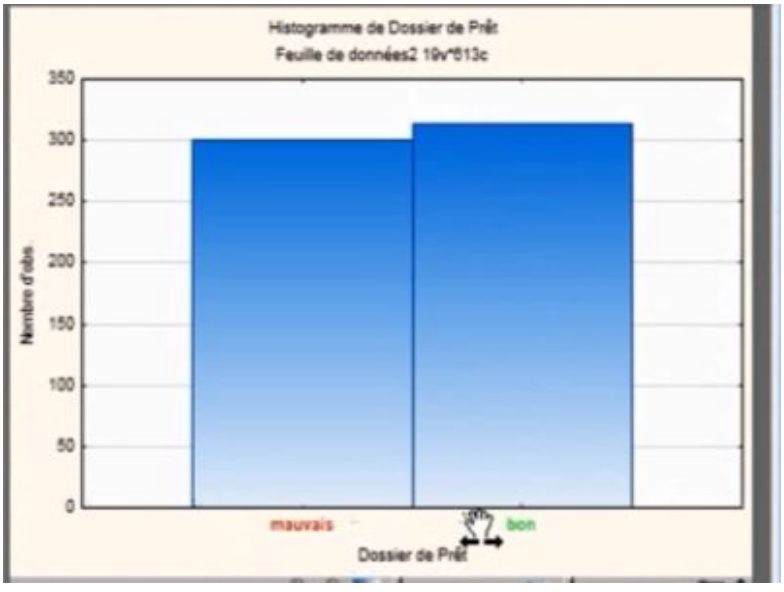
en se penchant sur les proportion de bon dossier de prêt nous constatons qu'il occupe près de 70% des observations et pourtant les mauvais dossiers de prêt occupe plutôt 30% des observations ce qui pourrait fausser notre apprentissage a cet effet il nous faut créer un rééquilibrage des donnée au niveau de notre variables d'intérêt qui est le **DOSSIER DE PRÊT.** a cet effet nous allons appliquer à notre variable d'intérêt dans STATISTICA **une méthode d' échantillonnage stratifié (la variable de stratification est la variable dossier de prêt)** afin de créer une sorte de rééquilibrage entre les différentes observations.

on obtient les résultat suivant en créer un filtre à 300 observations maximum car nous avons le nombre de mauvais dossier de prêt qui apparaît à 300 observations.

après ce traitement on obtient la nouvelle feuille de donnée suivante et le nouveau histogramme de **bon** et **mauvais dossier de prêt** ci dessous:



	Dossier de Prêt	2 Solde du Compte Courant	3 Durée du Prêt	4 Remboursement des Prêts Antérieurs	Objecti
1	mauvais	pas de compte courant	36	aucun problème avec les prêts antérie	réorient
2	bon	négligeable	48	délicat	réorient
3	mauvais	>300€	36	aucun prêt antérieur	véhicule
4	bon	pas de compte courant	24	remboursement intégral	véhicule
5	mauvais	pas de compte courant	30	aucun prêt antérieur	véhicule
6	bon	négligeable	15	remboursement intégral	mobilier
7	bon	>300€	15	remboursement intégral	mobilier
8	mauvais	négligeable	27	remboursement intégral	mobilier
9	mauvais	négligeable	24	aucun prêt antérieur	véhicule
10	mauvais	pas de compte courant	18	aucun prêt antérieur	réparati
12	mauvais	pas de compte courant	18	aucun prêt antérieur	réparations



d. Filtre de sélection des meilleurs prédicteurs

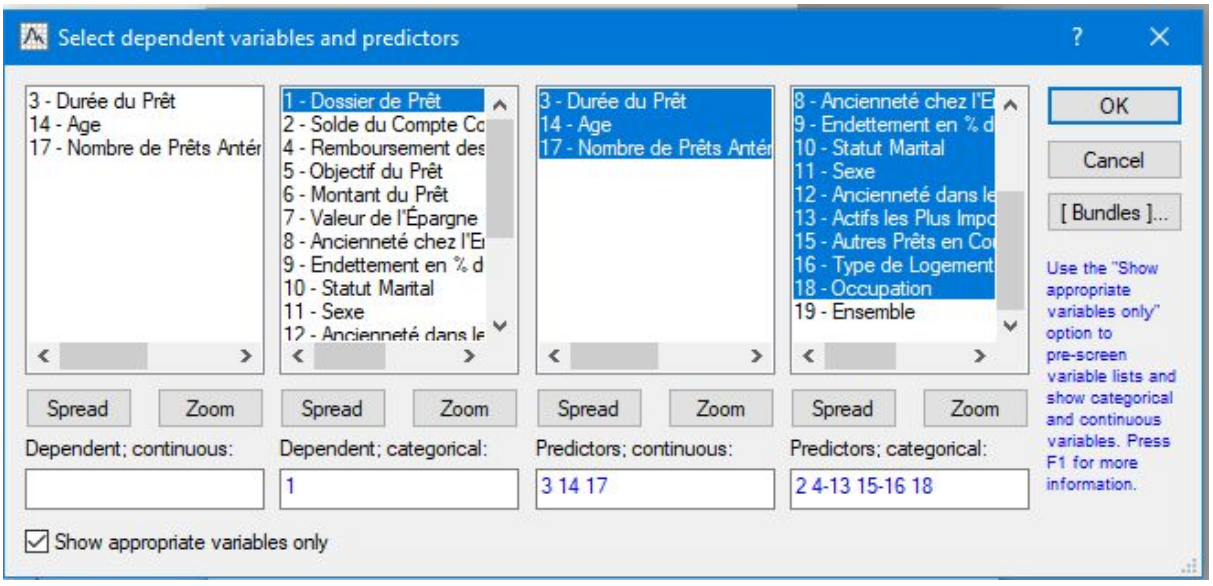
Dans cette section, nous nous occuperons de la sélection des variables d'intérêt qui sont les meilleures prédicteurs de notre variable d'intérêt.

cette méthode de filtrage permet également de supprimer des variables très difficile à suivre.

Dans notre outils nous avons la possibilité de définir par défaut le nombre de prédicteur soit la possibilité de définir un seuil de probabilité en utilisant un test significatif du CHI-2.

STATISTICA propose aussi un outil simple qui nous permettra de gerer les meilleurs prédicteurs en les rangeant sous forme de groupe suivant la nature de chaque variable (dans notre cas variable continue et la variable catégorielle) : les analyse permettent d'obtenir les capture ci-dessous :

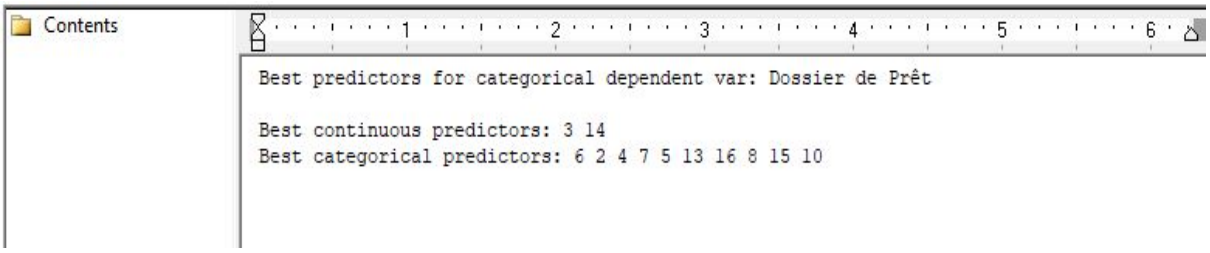
i. **sélection des meilleurs prédicteur continues et catégorielle**



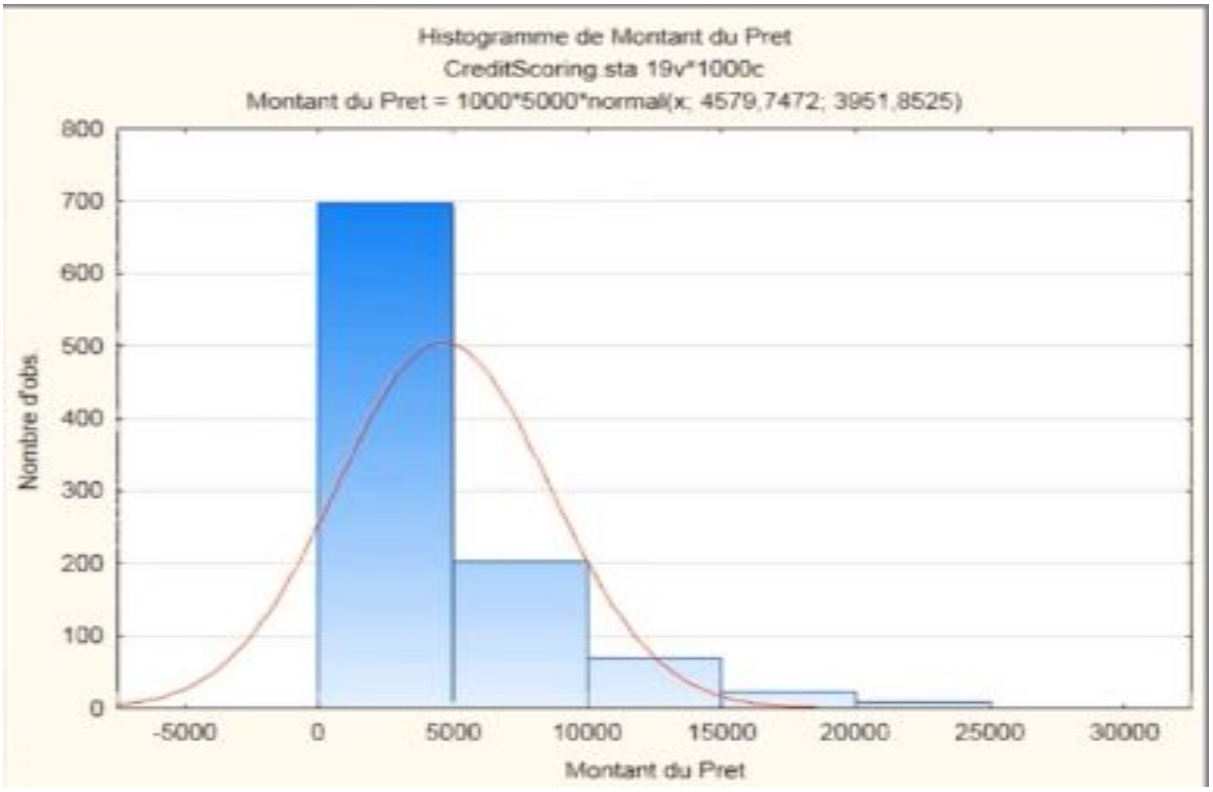
ii. **affichage des meilleures prédicteurs sous forme de tableaux**

	Best predictors for categorical dependent var: Dossier de Prêt (creditNew.xls (B2:ALM1001))	
	Chi-square	p-value
Montant du Prêt	931,7460	0,404516
Solde du Compte Courant	123,7209	0,000000
Remboursement des Prêts Antérieurs	61,6914	0,000000
Durée du Prêt	54,4481	0,000000
Valeur de l'Épargne	36,0989	0,000000
Objectif du Prêt	33,3564	0,000116
Actifs les Plus Importants	23,7196	0,000029
Age	19,4818	0,012485
Type de Logement	18,6740	0,000088
Ancienneté chez l'Employeur Actuel	18,3683	0,001045
Autres Prêts en Cours	12,8392	0,001629
Statut Marital	9,6052	0,022238

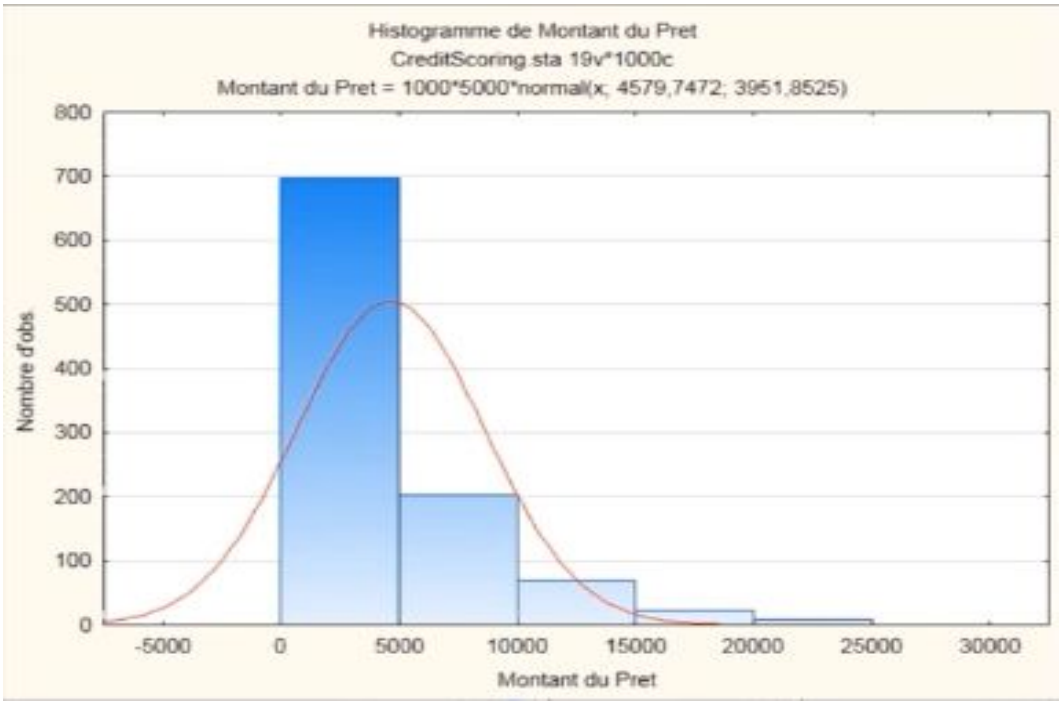
iii. affichage des meilleures prédicteurs sous la forme d'identifiant ligne colonne.



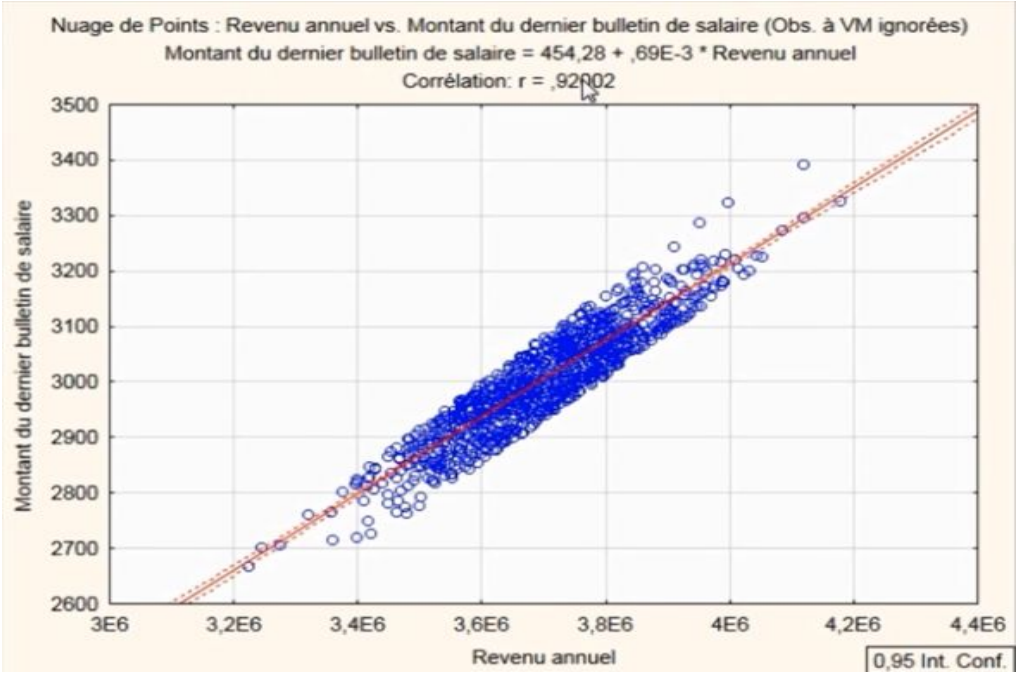
iv. histogramme du groupe des variables prédictive continue



iv. histogramme du groupe des variables prédictive catégorielle



v. nuage de point : revenue annuel vs montant du dernier bulletin de salaire



Nous remarquons dans la représentation du nuage de point une tres fortes corrélations et redondance entre ses deux variables **(r =0.92002)**.

7. Introduction aux méthodes de partitionnement récursif

ici il s'agira de trouver un et de modéliser un modèle permettant de caractériser un dossier de crédit comme BON ou MAUVAIS.
La méthode du partitionnement récursif fait référence au processus de constructions d'un arbres de décision. il s'agit en fait d'une certe question qui nous mènerait jusqu'à une prédiction finale.

- dans cette modélisation, les variables prédictives permettent de créer les branches des arbres afin de diviser et de repartir des observations dans des groupe de plus en plus homogènes.

- ses groupes homogènes constituerons les feuilles de l'arbres que nous appellerons des noeuds.
- lorsqu'on effectue suffisamment de division jusqu'au noeud terminaux, nous arrivons au noeuds terminaux ou nous n'aurons plus de division possible.
- dans notre outils statistica, les noeud terminaux sont représentés en rouge.

8. Construction des modèles de Classification

Principes Fondamentaux

Les arbres de classification permettent de prévoir l'affectation d'observations ou d'objets à des classes d'une variable dépendante catégorielle à partir de leurs mesures sur une ou plusieurs variables prédictives. Les arbres de décision constituent l'une des principales techniques utilisées en Data Mining.

Le but des arbres de classification consiste à prévoir ou expliquer les réponses d'une variable dépendante catégorielle.

a. Classification par l'arbre de C&RT (classification and regression trees)

Les Arbres de classification C&RT (Classification et régression) est une méthode de classification et de prévision basée sur un système d'arborescence. Cette méthode utilise la technique de partition récursive afin de diviser les données d'apprentissage en segments présentant des champs de sortie similaires. L'Arbre C&RT examine en premier lieu les champs d'entrée, afin de définir la meilleure segmentation : celle-ci est mesurée en fonction de la réduction de l'index d'impureté résultant de la segmentation. Le découpage définit deux sous-groupes qui sont à leur tour découpés en deux nouveaux sous-groupes : le découpage se poursuit jusqu'à ce que l'un des critères d'arrêt soit atteint. Toutes les divisions sont binaires (deux sous-groupes uniquement).

Conditions requises. L'apprentissage d'un modèle d'arbre C&RT requiert l'utilisation d'au moins un champ *Entrée* et d'un champ *Cible*. Les champs cible et d'entrée peuvent être continus (intervalle numérique) ou catégoriels.

Puissance. Les modèles d'arbre C&RT s'avèrent relativement robustes en présence de problèmes (par exemple, des données manquantes ou un nombre trop important de champs). Leur temps d'apprentissage est généralement court. De plus, les modèles d'arbre C&RT sont généralement plus faciles à comprendre que d'autres types de modèle dans la mesure où les règles extraites de ces modèles sont relativement simples à interpréter.

Conditions d'arrêt

Elle correspondent au critères utiliser pour trouver l'arbre de taille idéale, elle permet de déterminer quel noeud continuerons à être divisé et quel noeud seront des noeud terminaux les méthodes existantes sont :

- Elagage selon l'erreur de classement (ici si nous définissons le nombre de noeud maximum, alors se nombre de mot défini permettra d'arrêter la division de l'arbre)

- Elagage selon l'écart
- l’arrêt direct de type FACT

Validation Croisée

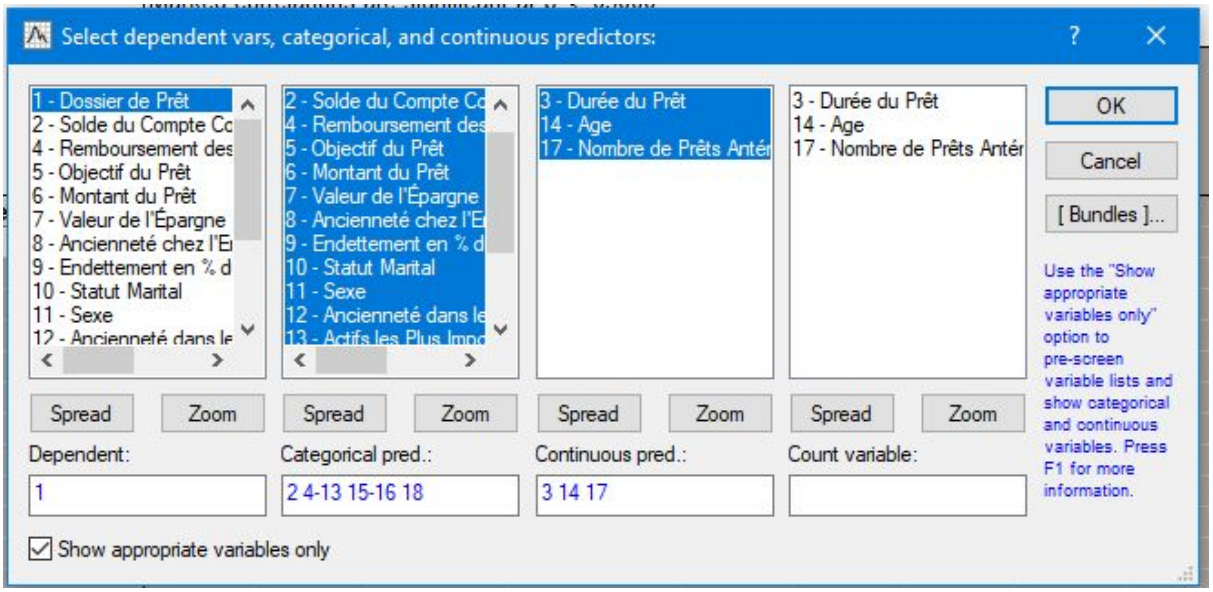
C’est une technique utilisée pour éviter le surajustement des données c’est à dire quand un modèle s’ai bien modéliser un jeux de donnée spécifique mais qui ne parvient pas a restituer la relation sous jacente entre les variables.

Substituts

L’utilisation de substitut est importante quand dans les donnée il existe des donnée manquantes dans ce cas, l’utilisation des valeurs manquantes permette de remplacer les valeurs manquantes afin de continuer l’apprentissage.

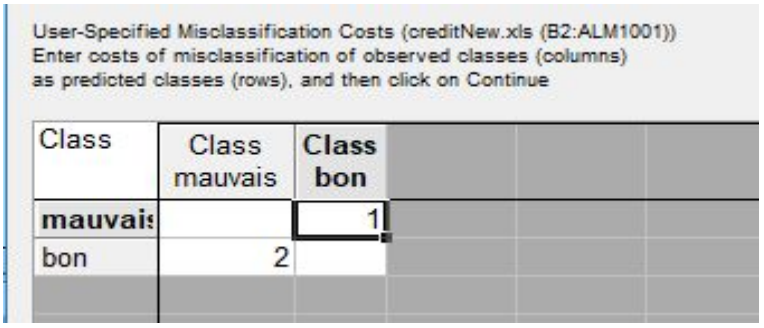
➔ CONSTRUCTION DE L’ARBRE

- **sélection des variable catégorielle et continue et la variable dépendante**



nous voyons dans la capture ci dessus la sélection de nos différentes variables catégorielle et continues pour lancer la construction du modèle et la variable dépendante.

- **définition des coûts d'erreur de classement**

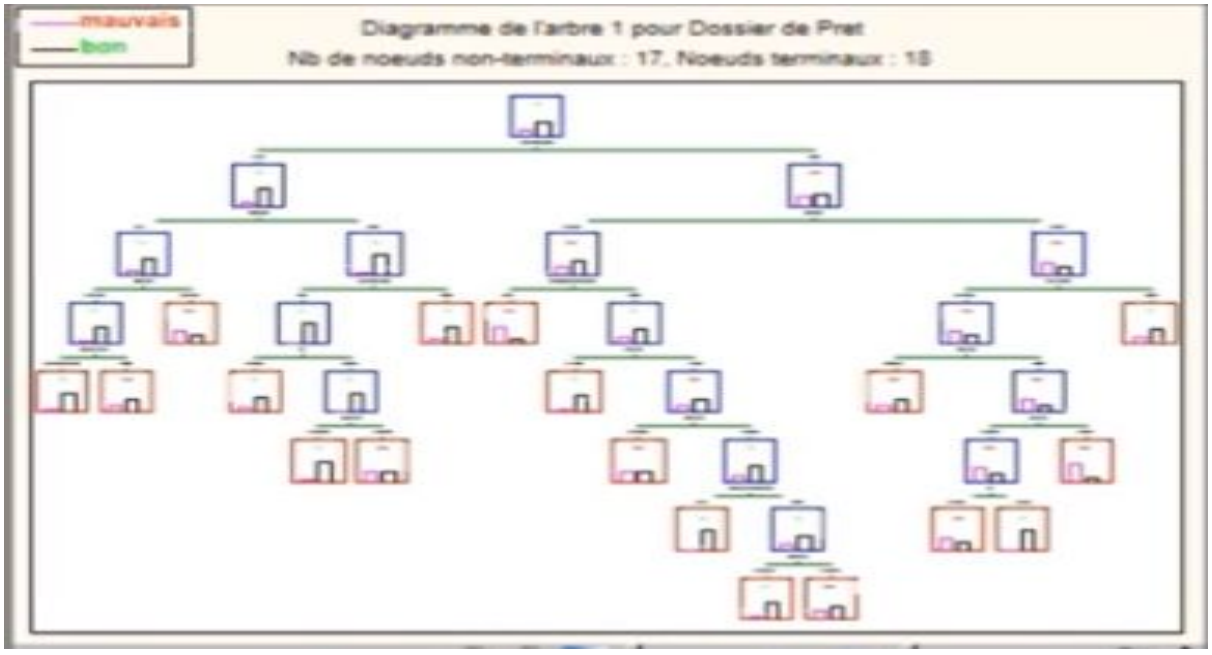


Dans la capture ci dessus, nous avons défini pour les cou d'erreur de classement 2 pour que la prédiction peut prédire un bon dossier mauvais et 1 la prédiction d'un mauvais dossier bon .

Ceci permet d'éviter les cas de figure où nous aurons à valider un mauvais dossier comme bon dossier car il serait mieux pour la structure de classer un bon dossier comme mauvais que de classifier un mauvais dossier comme bon. c'est ce qu'explique la capture ci-dessus.

après exécution de l'apprentissage nous obtenons les différents schéma suivant:

- **Diagramme de l'arbre pour le dossier de prêt**



Ce diagramme étant un peu complexe nous allons l'explorer de différentes façon en affichant son contenue sous différentes formes afin de comprendre la représentation des données.

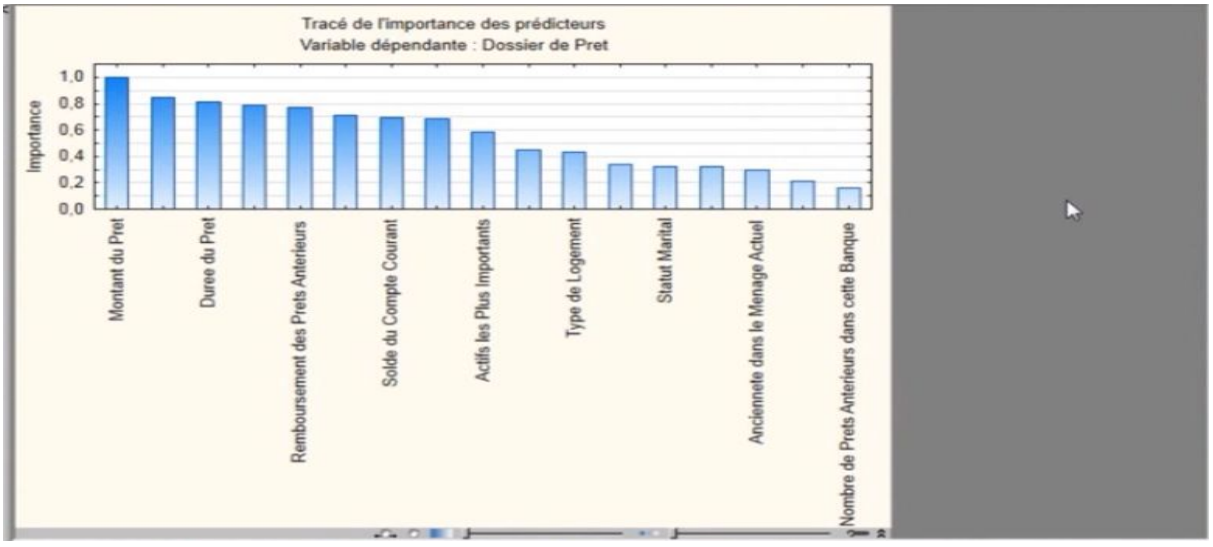
- **Structure de l'arbre**

Structure de l'arbre 1 (CreditScoring.sta)
Variable dépendante : Dossier de Pret
Options : Réponse catégorielle, Arbre numéro 1

N° du Noeud	Branche gauche	Branche droite	Taille du noeud	N de la classe mauvais	N de la classe bon	Catég. sélect.	Var. de division
1	2	3	1000	300	700	bon	Solde du Compte Coura
2	4	5	457	60	397	bon	Objectif du Pr
4	6	7	240	45	195	bon	Montant du Pr
6	8	9	229	38	191	bon	Autres Prets en Cou
8			189	24	165	bon	
9			40	14	26	mauvais	
7			11	7	4	mauvais	
5	14	15	217	15	202	bon	Solde du Compte Coura
14	16	17	190	9	181	bon	Ag
16			7	2	5	bon	
17	18	19	183	7	176	bon	Montant du Pr
18			181	6	175	bon	
19			2	1	1	mauvais	
15			27	6	21	bon	
3	26	27	543	240	303	mauvais	Duree du Pr
26	28	29	306	106	200	mauvais	Remboursement des Prets Anterieu
28			28	21	7	mauvais	
29	30	31	278	85	193	bon	Duree du Pr
30			80	14	66	bon	
31	32	33	198	71	127	mauvais	Montant du Pr
32			73	37	36	mauvais	
33	34	35	125	34	91	bon	Anciennete chez l Employeur Actu

On peut voire pour un début dans la représentation des données la répartition des données ainsi que les catégorie de division, nous remarquons également la répartition entre les bons et les mauvais dossier de prêt.

- **Tracé de l'importances des prédicteurs**



- **Valeurs prévues vs valeurs observé**

Valeurs prévues 1 (CreditScoring.sta)
Variable dépendante : Dossier de Pret
Options : Réponse catégorielle, Arbre numéro 1, Ech. analyse

	Observé observée	Prévu prévue	Probabilité de mauvais	Probabilité de bon	Noeud terminal
1	mauvais	mauvais	0,623077	0,376923	44
2	bon	bon	0,292683	0,707317	39
3	mauvais	mauvais	0,636364	0,363636	7
4	bon	mauvais	0,357143	0,642857	40
5	bon	bon	0,126984	0,873016	8
6	bon	mauvais	0,506849	0,493151	32
7	mauvais	mauvais	0,623077	0,376923	44
8	bon	bon	0,166667	0,833333	36
9	bon	bon	0,033149	0,966851	18
10	mauvais	mauvais	0,623077	0,376923	44
11	mauvais	mauvais	0,623077	0,376923	44
12	mauvais	mauvais	0,506849	0,493151	32
13	mauvais	mauvais	0,623077	0,376923	44
14	bon	bon	0,126984	0,873016	8
15	bon	bon	0,166667	0,833333	36
16	bon	bon	0,033149	0,966851	18
17	bon	mauvais	0,357143	0,642857	40
18	mauvais	mauvais	0,623077	0,376923	44
19	bon	mauvais	0,623077	0,376923	44
20	bon	bon	0,292683	0,707317	39
21	bon	bon	0,222222	0,777778	15
22	bon	bon	0,126984	0,873016	8

Nous pouvons voir la répartition de valeurs prévue et la proportion de valeurs observer après la création du modèle.

- **code de déploiement**

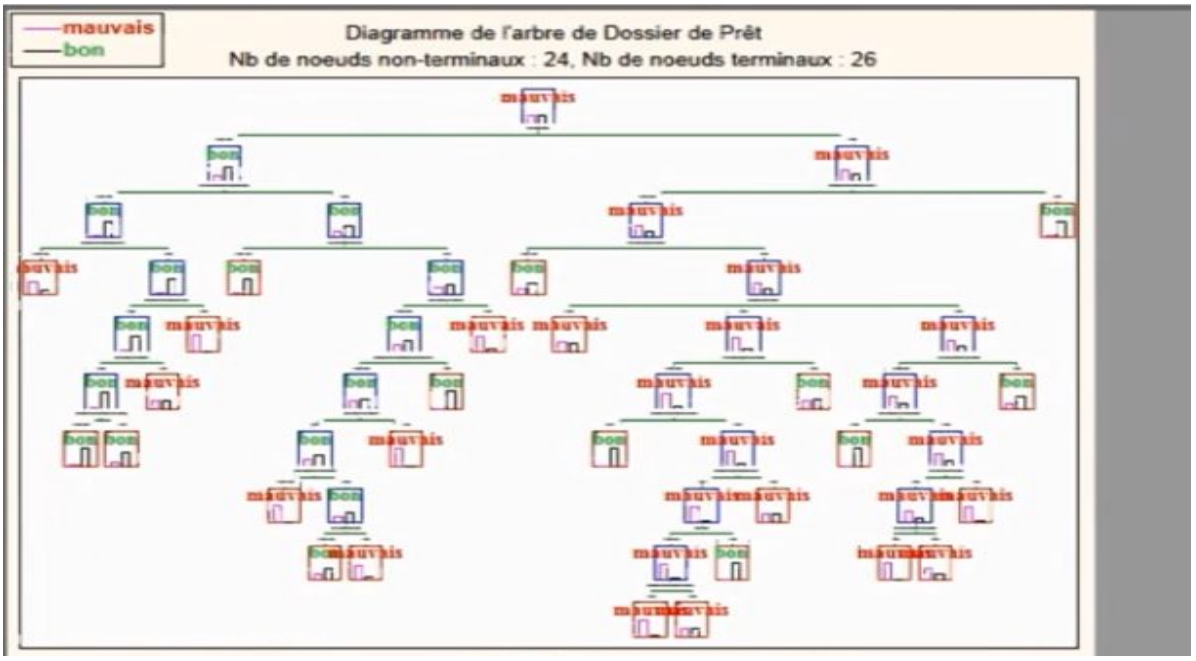
```
                        </Node>
                    </Node>
                </Node>
                <Node score="bon">
<targetPrediction name="mauvais" value="2.92682926829268e-001"/>
<targetPrediction name="bon" value="7.07317073170732e-001"/>
                    <CompoundPredicate booleanOperator="and">
                        <SimplePredicate field="Valeur de l Epargne"
operator="notEqual" value="pas d epargne"/>
                        <SimplePredicate field="Valeur de l Epargne"
operator="notEqual" value="&lt;140"/>
                        <SimplePredicate field="Valeur de l Epargne"
operator="notEqual" value="140-700"/>
                    </CompoundPredicate>
                </Node>
            </Node>
        </Node>
</Node>
        </TreeModel>
</PMML>
```

Nous obtenons également le code PMML qui nous permettra de déployer notre modèle dans n'importe quel environnement data mining de votre choix.

b. Classification par l'arbre de CHAID

Contrairement à la méthode C&RT cette méthode permet de construire des arbres non binaires. elle est généralement utiliser pour des gros volume de donnée.

Ces ajustement et traitement sont assez distincte nous afficherons ici juste les résultats obtenue en appliquant la méthode de classification CHAID.



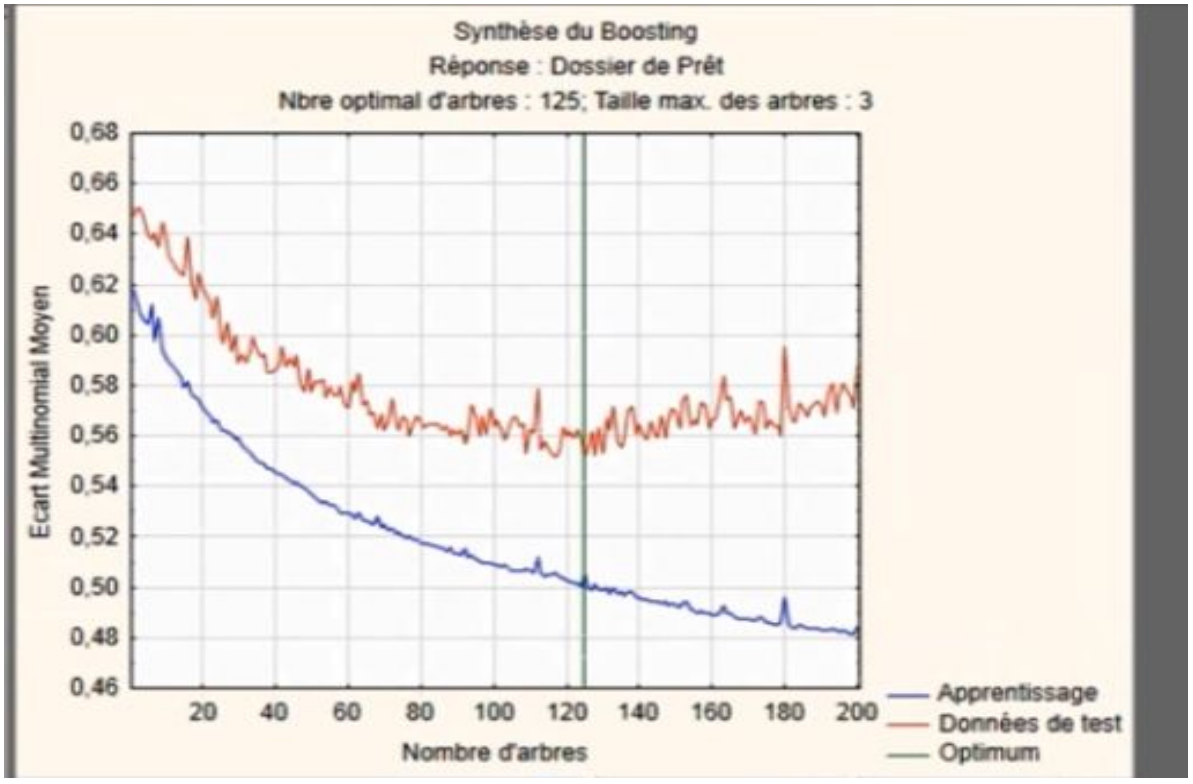
Nous pouvons remarquer qu'il n'existe pas de structure binaire dans la structure de l'arbre produite par l'algorithme CHAID

c. Classification par Boosting d'arbres

le Boosting d'arbre créer une série d'arbre de décision très simple et chacun de ces arbres pris séparément possède un faibles pouvoir prédictif mais une fois utiliser ensemble, tous ses arbres peuvent devenir d'excellent prédicteur. les prédictions du boosting d'arbre correspond donc à la prédiction de tous ses arbres simples pris dans leurs ensembles.

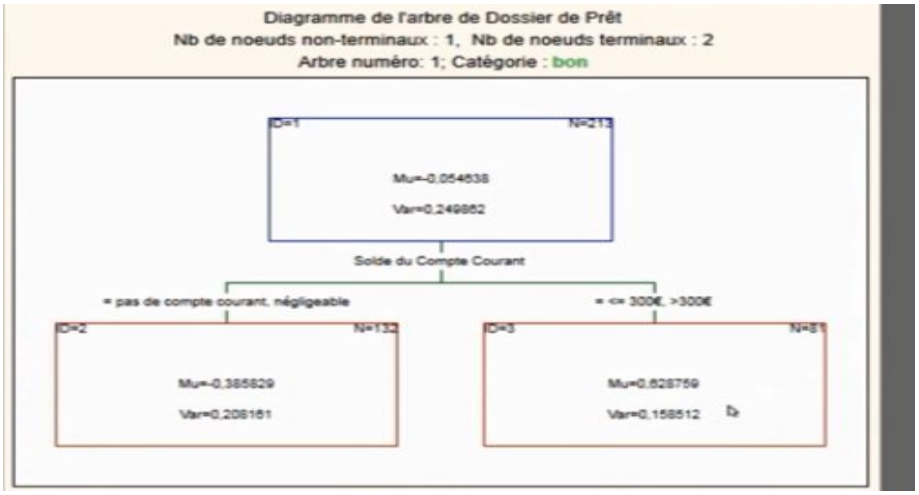
remarques le taux d'apprentissage est fixé a 0.1 et nous aurons à fixer dans notre cas la construction de près de 200 arbres de décision.

le graphe construit a partir du boosting d'arbres permet d'obtenir le graphe suivant qui représentent la synthèse de l'algorithme de BOOSTING ARBRE sur nos données :



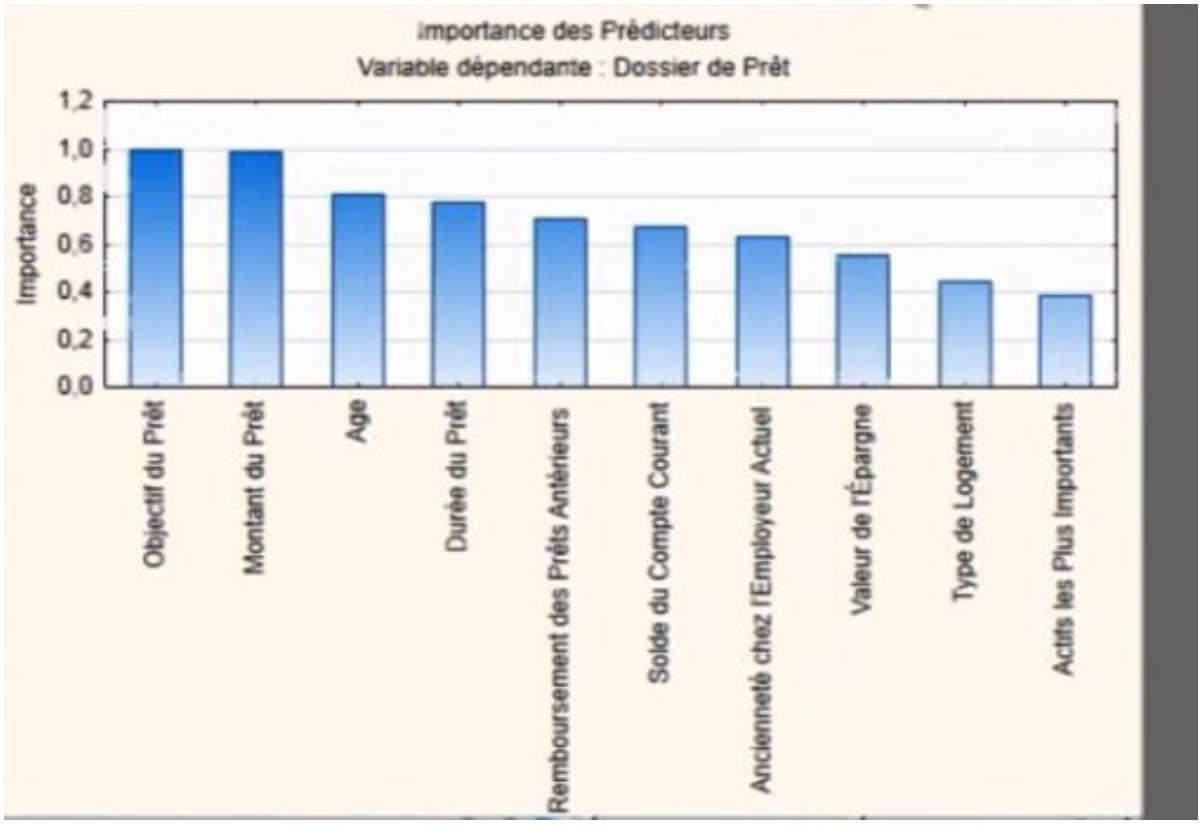
Nous dans le graphe de synthèse une très bonne réussite sur les donnée de test du modèle construit.

- **Diagramme de l'arbre de dossier de prêt (binaire)**

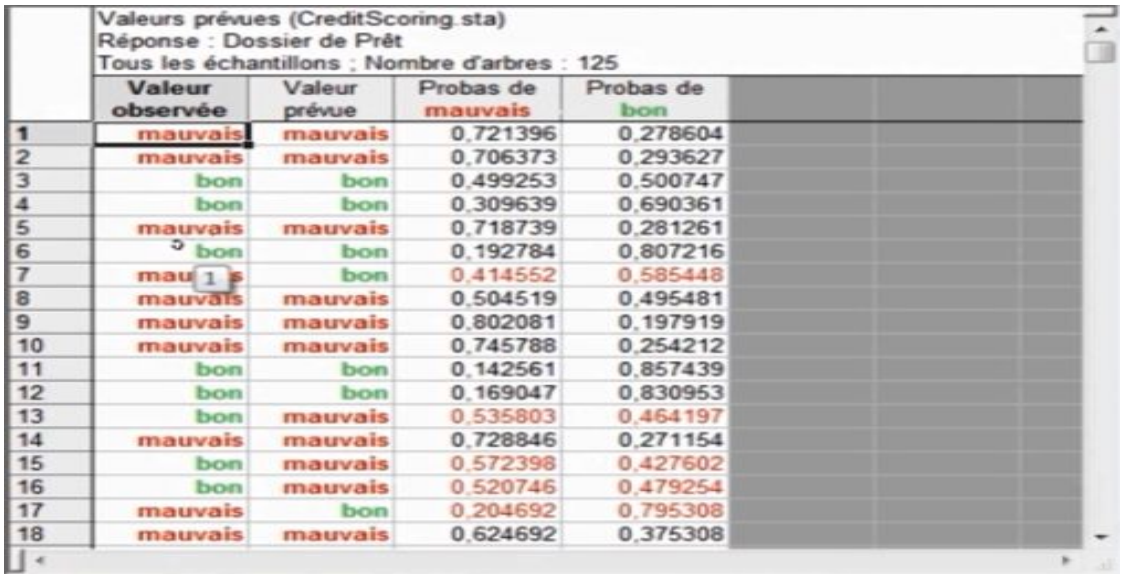


Cet arbre binaire représente l'affichage d'un seul noeud du boosting alors ce noeud ou cet affichage ne fourni pas assez d'informations sur la méthode du coup nous procéderons à l'affichage sous forme de tableau pour une bonne lisibilité et applicabilité.

- **importance des prédicteurs dans cette méthode**



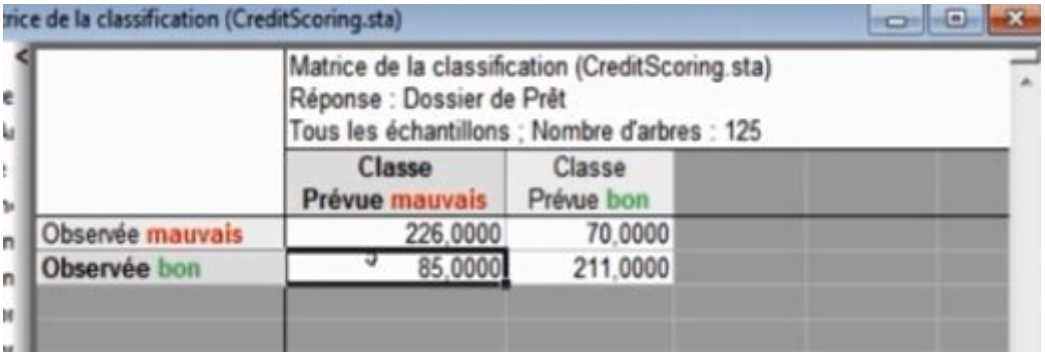
- **valeurs prévues vs valeurs observé**



Valeurs prévues (CreditScoring.sta)
Réponse : Dossier de Prêt
Tous les échantillons ; Nombre d'arbres : 125

	Valeur observée	Valeur prévue	Probas de mauvais	Probas de bon
1	mauvais	mauvais	0,721396	0,278604
2	mauvais	mauvais	0,706373	0,293627
3	bon	bon	0,499253	0,500747
4	bon	bon	0,309639	0,690361
5	mauvais	mauvais	0,718739	0,281261
6	bon	bon	0,192784	0,807216
7	mauvais	bon	0,414552	0,585448
8	mauvais	mauvais	0,504519	0,495481
9	mauvais	mauvais	0,802081	0,197919
10	mauvais	mauvais	0,745788	0,254212
11	bon	bon	0,142561	0,857439
12	bon	bon	0,169047	0,830953
13	bon	mauvais	0,535803	0,464197
14	mauvais	mauvais	0,728846	0,271154
15	bon	mauvais	0,572398	0,427602
16	bon	mauvais	0,520746	0,479254
17	mauvais	bon	0,204692	0,795308
18	mauvais	mauvais	0,624692	0,375308

- **Matrice de répartition**



Matrice de la classification (CreditScoring.sta)
Réponse : Dossier de Prêt
Tous les échantillons ; Nombre d'arbres : 125

	Classe Prévue mauvais	Classe Prévue bon
Observée mauvais	226,0000	70,0000
Observée bon	85,0000	211,0000

En observant la matrice de répartition de bon contre de faux dossier, on constate que notre algorithme a prédit 85 dossiers bon comme étant mauvais et de même a prédit 70 dossier mauvais comme étant bon dossier.

d. Méthode de Classification par les forêts aléatoires

i. Présentation de la méthode des forêts aléatoires utiliser

STATISTICA met en oeuvre le modèle de classification dit "**des Forêts Aléatoires**", développé par Breiman. Mais cet algorithme est également applicable à des problématiques de régression. Une Forêt Aléatoire est constituée d'un ensemble d'arbres simples de prévision, chacun étant capable de produire une réponse lorsqu'on lui présente un sous-ensemble de variables

explicatives ou prédicteurs. Pour les problématiques de classification, la réponse prend la forme d'une classe qui associe un ensemble (classe) de valeurs indépendantes (prédicteur) à une des catégories présente dans la variable dépendante.

Une forêt aléatoire est constituée d'un nombre arbitraire d'arbres simples, qui permettent de voter pour la classe la plus populaire (classification), ou dont les réponses sont combinées (moyennées) pour obtenir une estimation de la variable dépendante (régression). En utilisant des ensembles d'arbres, nous parvenons à améliorer significativement la prévision (c'est-à-dire avec une meilleure capacité à prévoir de nouvelles données).

La réponse de chaque arbre dépend du sous-ensemble de prédicteurs choisis de façon indépendante (avec remise) et avec la même distribution pour tous les arbres de la forêt qui est un sous-ensemble des valeurs des prédicteurs du jeu de données original. Dans le module *STATISTICA Forêts Aléatoires*, la taille optimale du sous-ensemble de variables prédictives est donnée par la formule $\mathbf{log_2}\ \boldsymbol{M}\mathbf{+1}$, où $\boldsymbol{M}$ représente le nombre d'entrées.

Pour les problèmes de classification, étant donné un ensemble d'arbres simples et un ensemble aléatoire de variables prédictives, la méthode des Forêts Aléatoires va définir une fonction d'erreur qui va déterminer dans quelle mesure le nombre moyen de votes pour la classe correcte dépasse le vote moyen des autres classes de la variable dépendante. Cette mesure constitue donc une bonne manière d'effectuer des prévisions, mais nous permet également d'associer une mesure de confiance à ces prévisions.

où l'indice *k* varie pour tous les arbres individuels de la forêt.

Lors de la construction du modèle dans STATISTICA, lorsqu'une observation particulière comporte des valeurs manquantes, la prévision réalisée pour cette observation se base sur l'avant dernier noeud (non-terminal) de l'arbre respectif. Ainsi, par exemple, si à un moment donné de la séquence des arbres, une variable prédictive est sélectionnée au noeud racine (ou un autre noeud non-terminal) et que certaines observations n'ont pas de données valides, la prévision de ces observations sera simplement basée sur la moyenne globale du noeud racine (ou du noeud non-terminal).

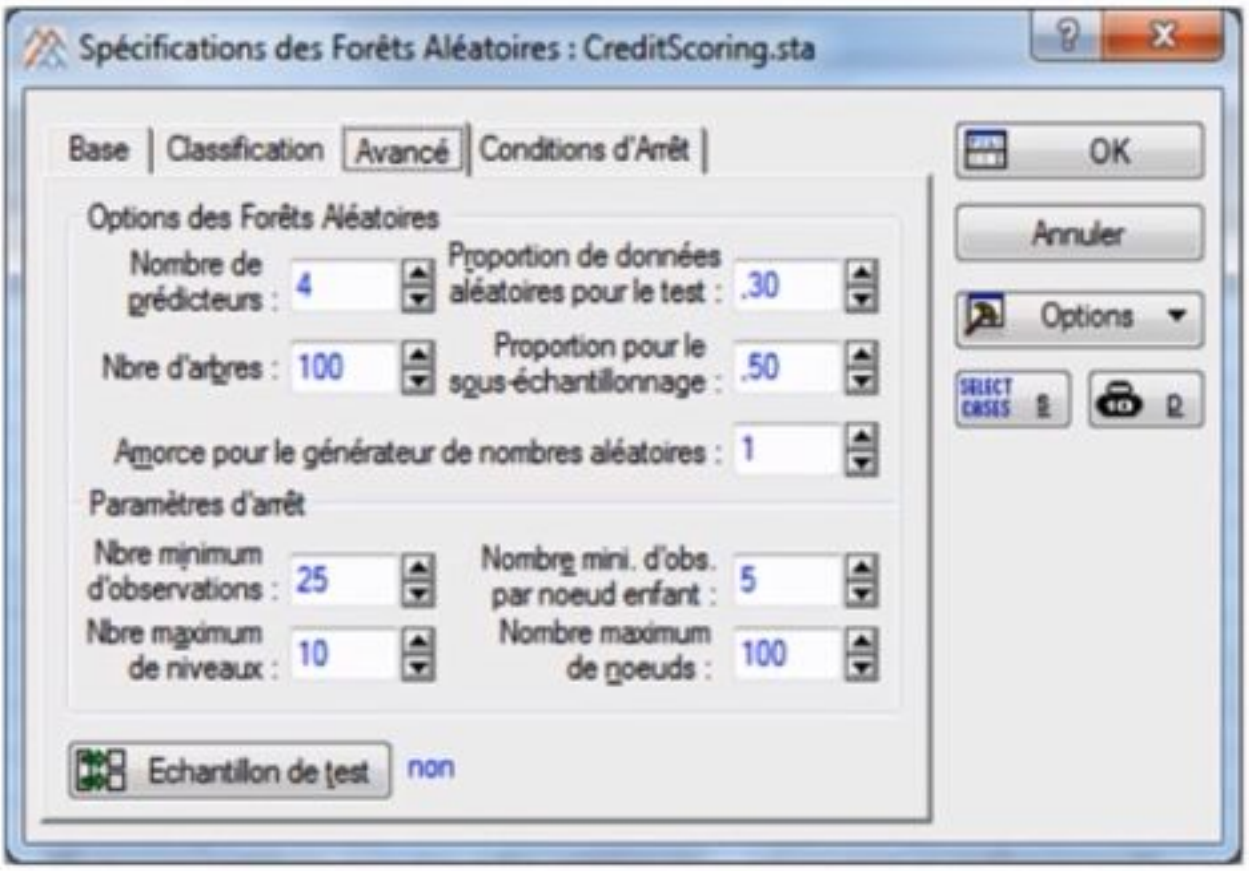
ii. Classification par les forêt aléatoire

Dans **STATISTICA** il suffit d'aller dans l'onglet **Data mining** et aller à **forêt aléatoire** ensuite choisir type de méthode **classification.**

Nous obtiendrons le tableau de configuration ci-dessous que nous pouvons renseigner manuellement.

Dans ce qui suit nous allons essayer d'afficher et d'expliquer les résultat que nous obtenons a la suite de nos paramètres fourni dans le tableau.

❑ Options d'Analyse et Description des paramètres



➔Options des Forêts Aléatoires

- **Nombre de prédicteur :** détermine le nombre de variables indépendante à considérer à chaque noeud de chaque arbres. sa valeur optimale est donnée par $\log_2 M+1$, où M représente le nombre de variables prédictive en entrée. STATISTICA utilise donc cette formule pour déterminer ce paramètre.

Nous limitons le nombre de prédicteur a chaque noeud afin de minimiser la corrélation entre les différent arbres de la forêt et par conséquent sa nous permettra de réduire le taux d'erreur de la forêt.

- **nombre d'arbres:** permettent de définir le nombres d'arbres donc nous voulons dans notre cas nous en avons choisi de construire 100
- **proportion de données aléatoires pour le test:** permet de régler la proportion de l'échantillonnage aléatoire dans notre cas nous avons choisi une proportion de 0.30

- **Proportion pour le sous-échantillonnage**
- **Amorce pour le générateur de nombres aléatoires**

➔Paramètres d'arrêt

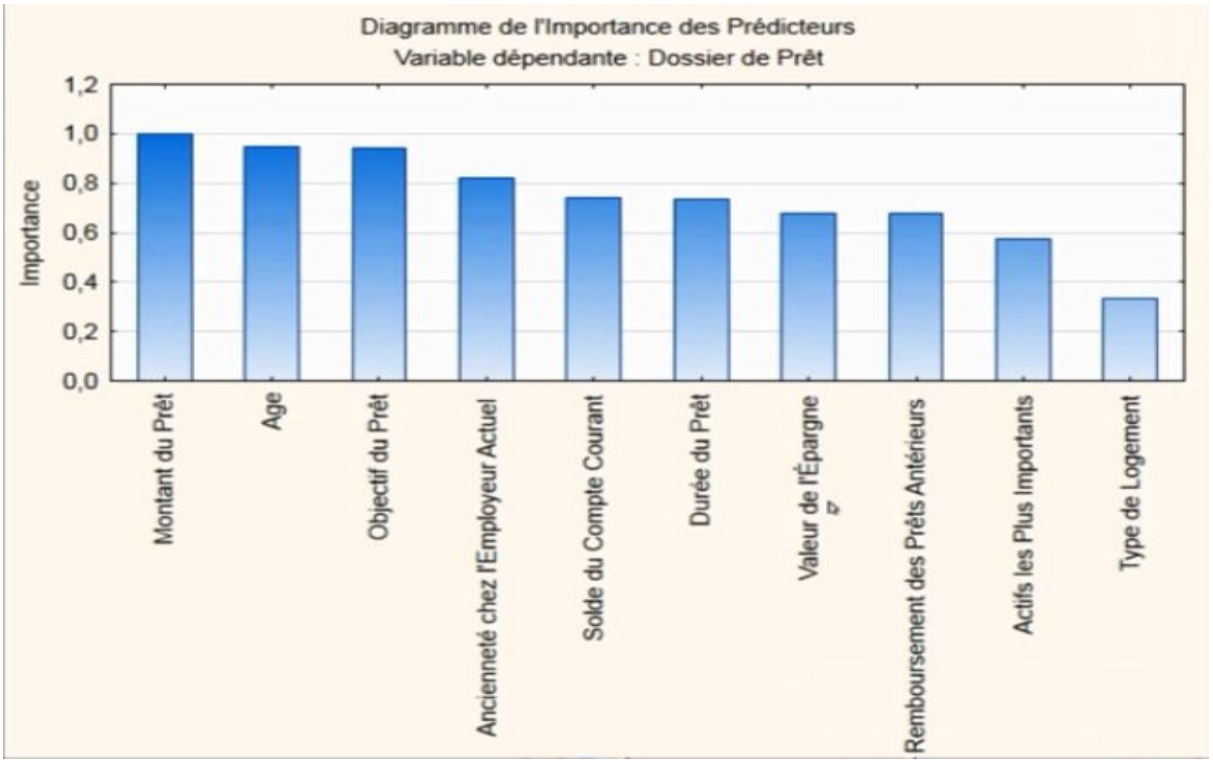
- **nombre minimum d'observation**
- **nombres maximum de niveaux:** représentent pour sa part le nombre minimum de noeud enfant à obtenir par niveau.
- **nombre minimum d'observation par noeud enfant**
- **nombre maximum de noeuds**

Tous ses paramètres d'arrêts permettent de contrôler la complexité des arbres qui constituent la forêt.

➔ Description des résultat obtenue

- **affichage du diagramme d'importance des prédicteurs**

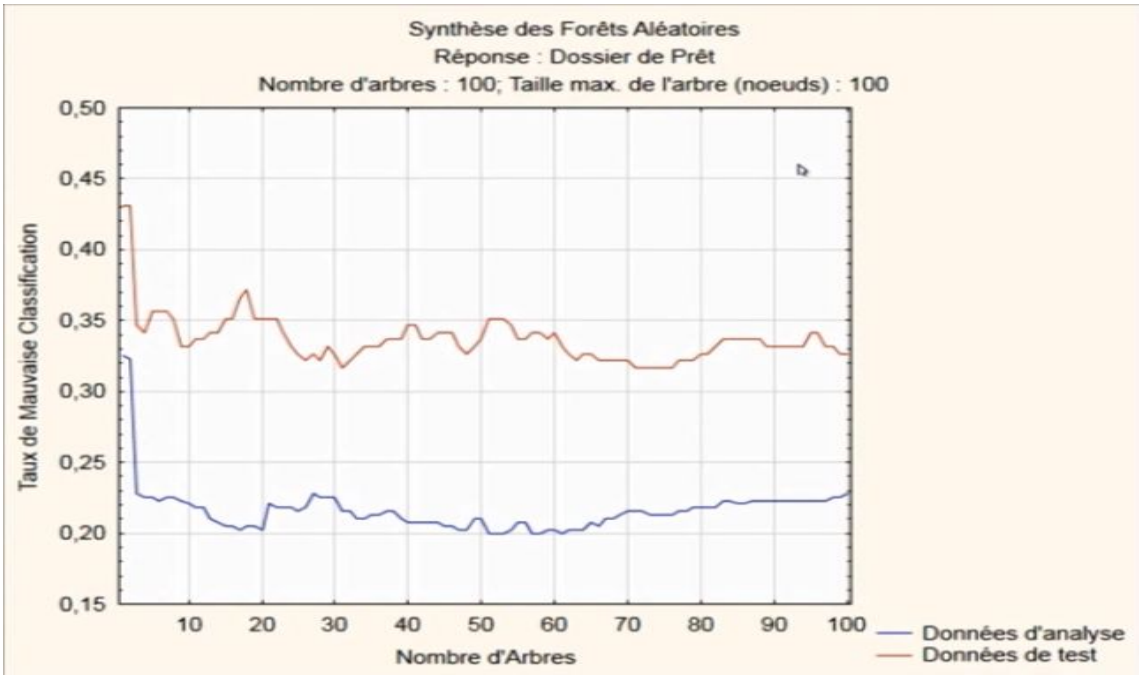
Nous pouvons représenter le degré d'importance de meilleurs prédicteurs sous la forme de diagramme et de tableau de valeurs comme nous pouvons le constater dans les deux figure ci dessous:



Cet histogramme laisse ressortir le degré d'importance des différentes variables pour les forêts aléatoires lors de l'entraînement du modèle.

	Importance des prédicteurs (CreditScoring.sta) Réponse : Dossier de Prêt	
	Rang de var.	Importance
Montant du Prêt	100	1,000000
Age	95	0,946922
Objectif du Prêt	94	0,942309
Ancienneté chez l'Employeur Actuel	82	0,822893
Solde du Compte Courant	74	0,739656
Durée du Prêt	73	0,733348
Valeur de l'Épargne	68	0,677599
Remboursement des Prêts Antérieurs	68	0,676960
Actifs les Plus Importants	57	0,572993
Type de Logement	33	0,334019

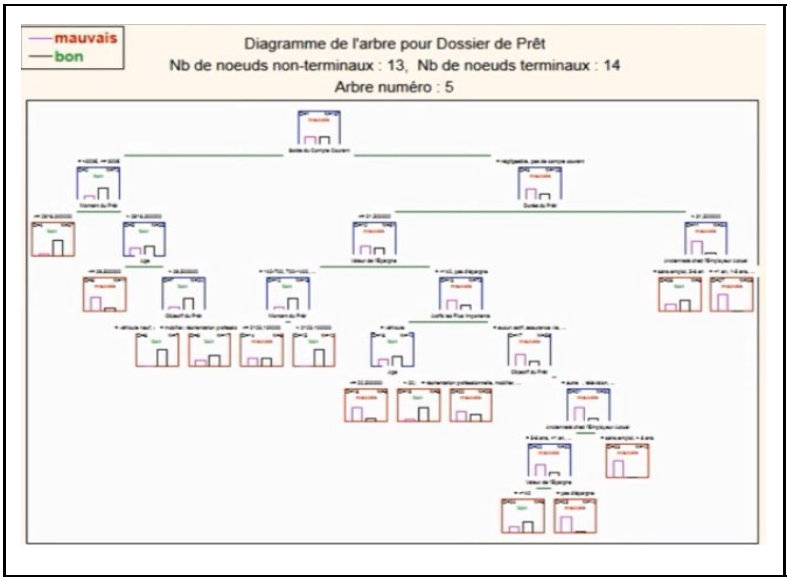
- **synthèse de l'échantillon de test**

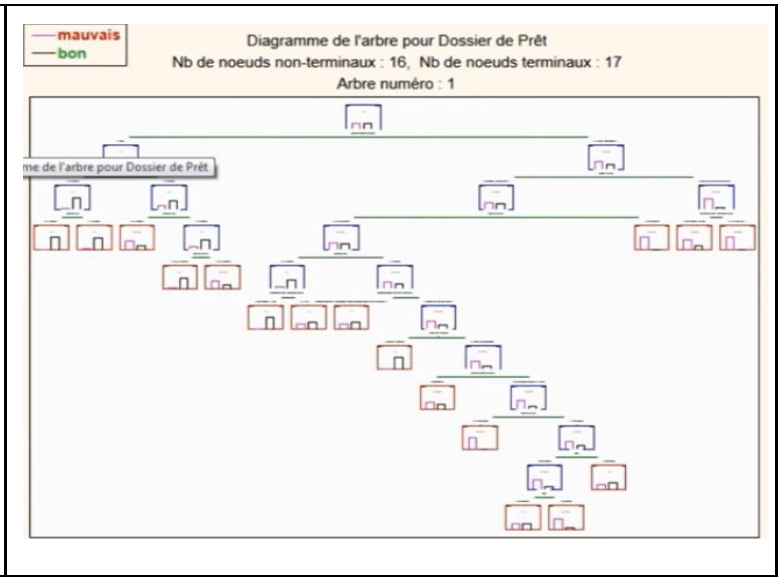


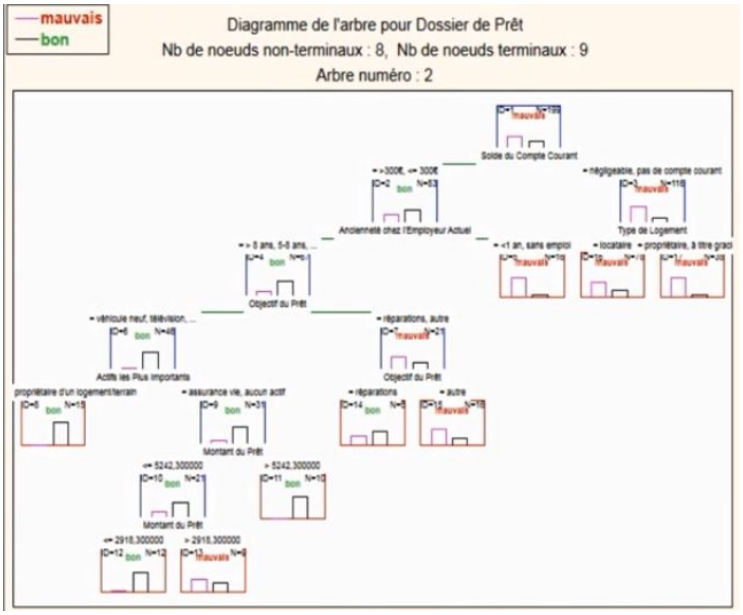
On remarque d'après le graphe obtenue que le pourcentage de l'échantillon de **test** est de l'ordre de **0.32** tandis que celui de l'apprentissage est de l'ordre de **0.23.**

- **représentation de quelques arbres de la forêt**

Il faut noter que chacun de ces arbres représentent un vote pour le résultat final ici nous avons juste extrait 3 arbres de l'ensembles des 100 arbres construit.







- **matrice de classification**

	Matrice de classification (CreditScoring.sta) Réponse : Dossier de Prêt Tous les échantillons; Nombre d'arbres : 100	
	Classe Prévue mauvais	Classe Prévue bon
Classe observée mauvais	239,0000	57,0000
Classe observée bon	100,0000	196,0000

En observant la matrice de répartition de bon contre de mauvais dossier de prêt, on constate que notre algorithme a prédit 100 dossiers bon comme étant mauvais et de même a prédit 57 dossier mauvais comme étant bon dossier. cela signifie que notre modèle prédit plus de bon résultat qu'il ne se trompe.

- **valeurs prévues et valeurs observé**

Valeurs prévues (CreditScoring.sta)
Réponse : Dossier de Prêt
Tous les échantillons; Nombre d'arbres : 100

	Classes observées	Classes prévues	Probas de mauvais	Probas de bon
1	mauvais	mauvais	0,858586	0,141414
2	mauvais	mauvais	0,653061	0,346939
3	bon	mauvais	0,595960	0,404040
4	bon	mauvais	0,673469	0,326531
5	mau	mauvais	0,680000	0,320000
6	bon	bon	0,110000	0,890000
7	mauvais	bon	0,450000	0,550000
8	mauvais	mauvais	0,747475	0,252525
9	mauvais	mauvais	0,886598	0,113402
10	mauvais	mauvais	0,858586	0,141414
11	bon	bon	0,265306	0,734694
12	bon	bon	0,220000	0,780000
13	bon	mauvais	0,770000	0,230000
14	mauvais	mauvais	0,888889	0,111111
15	bon	mauvais	0,560000	0,440000
16	bon	mauvais	0,537634	0,462366
17	mauvais	bon	0,300000	0,700000
18	mauvais	mauvais	0,585859	0,414141
19	mauvais	mauvais	0,670000	0,330000
20	bon	bon	0,210000	0,790000
21	mauvais	mauvais	0,846939	0,153061
22	mauvais	mauvais	0,640000	0,360000
23	mauvais	mauvais	0,741573	0,258427
24	bon	bon	0,220000	0,780000
25	mauvais	mauvais	0,530000	0,470000
26	bon	bon	0,030000	0,970000

- **Classes Observées :** elles représentent la valeurs de la classe prévu par la classe d'origine.
- **Classes prévues :** elles représentent la valeurs de la classe prévu et prédit par le modèle.
- **Probabilité de mauvais** : représentent la probabilité d'obtention de la classe mauvais dossier de prêt correspondant.
- **Probabilité de bon** : représentent la probabilité d'obtention de la classe bon dossier de prêt correspondant.

9. Comparaison de modèles et sélection du meilleure modèle

Nous allons dans cette section comparer nos différents modèle que nous avons construit afin de choisir le modèle qui résout efficacement notre problématiques. nous évaluerons particulièrement notre code a l'aide des

courbes de ***lift et de gain*** qui permettent de synthétiser graphiquement la performance de different model.

Dans STATISTICA, après avoir déployé les différents modèle, et déployer les code PMML issue de nos 4 apprentissage précédent, nous observons dans statistica les information ci-dessous :

- **Tableau de synthese apres déploiement de différents modèle**

Synthèse du Déploiement (CreditScoring.sta)
FileNames: Boosting.xml|CHAID.xml|CRT.xml|Forets_Aleatoires.xml

	Dossier de Prêt	BoostTreeModelPrév	BoostTreeModelRés	BoostTreeModel mauvais	BoostTreeModel bon	CHAIDModelPrév	CHAIDModelRés	CHAIDModel mauvais	CHAIDMod bon
1	mauvais	mauvais	Correct	0,721396	0,278604	mauvais	Correct	0,750000	0,250
2	mauvais	mauvais	Correct	0,706373	0,293627	bon	Incorrect	0,336538	0,663
3	bon	bon	Correct	0,499253	0,500747	bon	Correct	0,473684	0,526
4	bon	bon	Correct	0,309639	0,690361	mauvais	Incorrect	0,633540	0,366
5	mauvais	mauvais	Correct	0,718739	0,281261	mauvais	Correct	0,750000	0,250
6	bon	bon	Correct	0,192784	0,807216	bon	Correct	0,336538	0,663
7	mauvais	bon	Incorrect	0,414552	0,585448	bon	Incorrect	0,340000	0,660
8	mauvais	mauvais	Correct	0,504519	0,495481	mauvais	Correct	0,633540	0,366
9	mauvais	mauvais	Correct	0,802081	0,197919	mauvais	Correct	0,750000	0,250
10	mauvais	mauvais	Correct	0,745788	0,254212	mauvais	Correct	0,633540	0,366
11	bon	bon	Correct	0,142561	0,857439	bon	Correct	0,336538	0,663
12	bon	bon	Correct	0,169047	0,830953	bon	Correct	0,114943	0,885
13	bon	mauvais	Incorrect	0,535803	0,464197	mauvais	Incorrect	0,750000	0,250
14	mauvais	mauvais	Correct	0,728846	0,271154	mauvais	Correct	0,633540	0,366
15	bon	mauvais	Incorrect	0,572398	0,427602	mauvais	Incorrect	0,633540	0,366
16	bon	mauvais	Incorrect	0,520746	0,479254	bon	Correct	0,340000	0,660
17	mauvais	bon	Incorrect	0,204692	0,795308	bon	Incorrect	0,336538	0,663
18	mauvais	mauvais	Correct	0,624692	0,375308	mauvais	Correct	0,633540	0,366
19	mauvais	mauvais	Correct	0,573555	0,426445	bon	Incorrect	0,336538	0,663
20	bon	bon	Correct	0,169277	0,830723	bon	Correct	0,473684	0,526
21	mauvais	mauvais	Correct	0,792215	0,207785	mauvais	Correct	0,633540	0,366
22	mauvais	mauvais	Correct	0,696262	0,303738	bon	Incorrect	0,473684	0,526
23	mauvais	mauvais	Correct	0,848386	0,151614	mauvais	Correct	0,750000	0,250
24	bon	bon	Correct	0,230191	0,769809	bon	Correct	0,336538	0,663
25	mauvais	mauvais	Correct	0,571237	0,428763	bon	Incorrect	0,336538	0,663
26	bon	bon	Correct	0,094971	0,905029	bon	Correct	0,336538	0,663
27	mauvais	mauvais	Correct	0,772513	0,227487	mauvais	Correct	0,750000	0,250

Nous remarquons qu'elle contient les valeurs observées de la variable de sortie dossier de prêt ainsi que les valeurs prédite et les probabilité associé pour les quatres modèles**.**

- **Synthèse du déploiement des Taux d'erreur associé à nos différents données d'apprentissage**

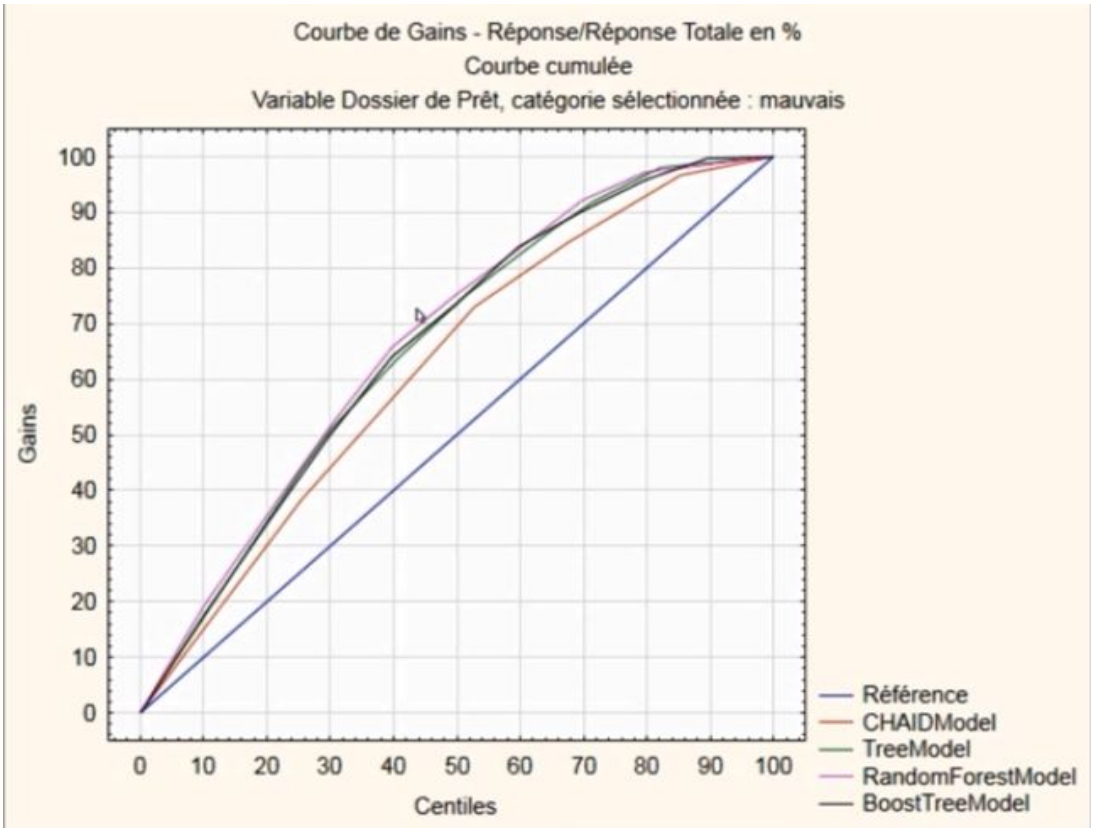
Synthèse du Déploiement (Taux d'erreur) (CreditScoring.sta)

	BoostTreeModel	CHAIDModel	TreeModel	RandomForestModel
Taux d'erreur	0,261824	0,298986	0,258446	0,265203

Cette feuille contient le taux d'erreur de classement des 4 modèles nous pouvons constater à ce sujet que l'arbre C&RT possède le taux d'erreur le plus faible.

a. Courbe de LIFT pour la catégorie mauvais dossier

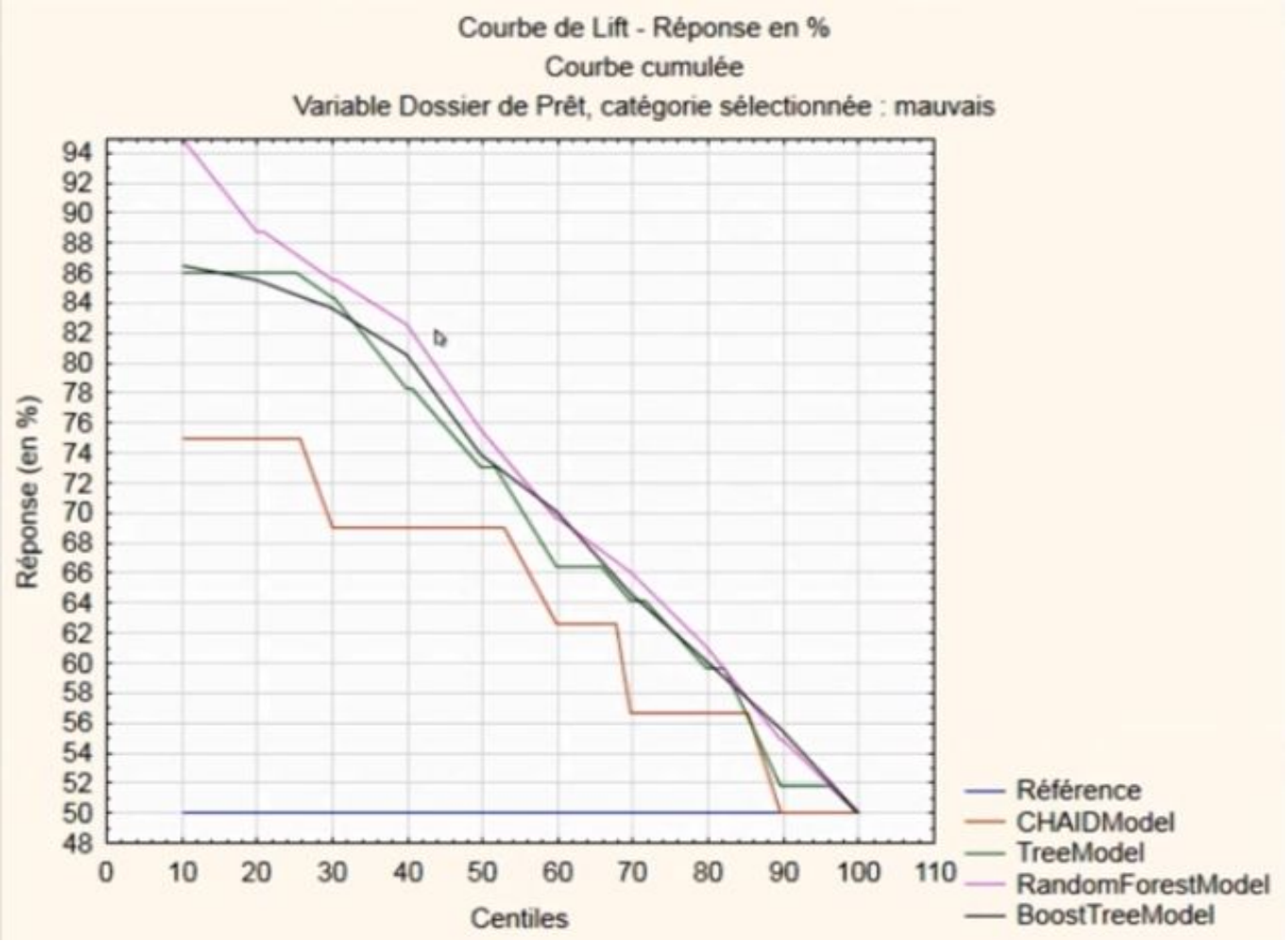
Elle permet de comparer l'efficacité du modèle par rapport à un modèle de choix aléatoire.



Nous constatons dans sa courbe représentative que c'est l'arbre des forêt aléatoire qui possède la plus grande aire(représenté en rose).

b. Courbe de GAIN pour la catégorie mauvais dossier

Elle indique la part d'observations correctement classées pour une classe donnée.



Cette courbe informe que c'est la méthode de **forêt aléatoire** qui permet de représenter le mieux les dossier à risque, puis suis le modèle de **Boosting** puis le modèle **CHAID**

conclusion sur la sélection du modèle:

Après avoir effectué toutes les opération de traitement nettoyages égalisation des variables traitements des valeurs manquantes aberrantes, atypiques…

nous avons dans cette partie construit des arbres de classification pour notre jeux de donnée de creditScoring qui consiste pour une institution financière de mesurer le risque de crédit associé à ses différent clients.

Après analyse des résultats obtenue nous pouvons tirer les conclusion ci dessous.

- Nous avons pu observer à partir de la courbes de **GAIN** appliquée au mauvais dossier de prêt qui permet d'indiquer la part d'observations correctement classées pour une classe donnée que **la méthode de classification par les Forêts aléatoire** étais la plus approprié pour représenter les dossiers à risque devant les autres méthode de prédiction comme C&RT, CHAID…
- De même, avec la courbe de **LIFT**, qui de sa part permet de comparer l'efficacité du modèle par rapport à un modèle de choix aléatoire que la méthode de **l'arbre des forêt aléatoire** qui possède la plus grande aire(représenté en rose).
- Nous sommes donc en face d'un modèle qui est efficace et donc la part d'observation correctement classées affiche de meilleur résultat par rapport a d'autre modèle de classification.

De tout ce qui précède, il révèle le modèle conçu sur les forêt aléatoire est la plus adapter pour reconnaître les mauvais dossier dans notre car

Conclusion

Dans ce travail pratique nous avons effectué un très long processus d'analyse de données. Commençant par l'acquisition et préparation de données, puis continuer avec les analyses. Cela nous a permis de faire une mise en évidence de l'analyse de données dans la fouille de données qu'on a appris en classe. Ainsi, nous avons rencontré beaucoup de difficulté au niveau de la compréhension des donnée et du choix final du jeux de donnée ce qui nous a permis de faire un grand pas dans l'analyse de donnée surtout avec ce nouvel outils STATISTICA qui est celui d'un des partenaire de l'IFI cela nous préparera certainement aussi pour les stages et les entreprises.

Références

1. ***STATISTICA* Statistiques Avancées + Solutions Industrielles**

   http://www.statsoft.fr/logiciels/statistiques-avancees-et-lean-six-sigma.php
2. **Forêts** Aléatoires (ou **Forêts** Décisionnelles)

   http://www.statsoft.fr/concepts-statistiques/concepts-statistiques/forets-decisionnelles/forets-aleatoires.php#.XSBKXSYxVpg
3. http://www.statsoft.fr/search/search.html?cx=016251123175171892936%3Apxvam5rw-zi&cof=FORID%3A9&ie=latin1&oe=latin1&q=for%C3%AAt+&sa.x=0&sa.y=0&siteurl=https%3A%2F%2Fwww.google.com%2F
4. https://www.tibco.com/fr/node/19696