|  |
| --- |
|  |
| Communautés et criminalité non normalisées |
| **Réalisé par :** |
|  |
| **BEN AHMED Chaima - MTIBAA Ali –GASMI Amine** |
| **03/05/2018** |

**Encadré par :**

**Mr.HAMROUNI Tarek**

TABLE DES MATIERES :

[I. Introduction 1](#_Toc513102800)

[II. Source deS donnée 1](#_Toc513102801)

[III. Problématiques 6](#_Toc513102802)

[IV. Processus de travail 6](#_Toc513102803)

[1. Les principales causes qui concrétisent la criminalité 6](#_Toc513102804)

[1.1 Sélection des données 6](#_Toc513102805)

[1.2 Prétraitement des données 7](#_Toc513102806)

[1.3 Fouille deS données 9](#_Toc513102807)

[2. la fixation du nombre de policier par population 16](#_Toc513102809)

[2.1 Sélection des données 16](#_Toc513102810)

[2.2 Prétraitement des données 16](#_Toc513102811)

[2.3 Fouille deS données 17](#_Toc513102812)

[V. Conclusion 21](#_Toc513102812)

TABLE DES FIGURES :

[Figure 1:types des données 7](file:///C:\Users\PC\Downloads\Projet-data-mining.docx#_Toc513102835)

[Figure 3: Binairisation des données 8](file:///C:\Users\PC\Downloads\Projet-data-mining.docx#_Toc513102836)

[Figure 4:Base de données binairisée 9](file:///C:\Users\PC\Downloads\Projet-data-mining.docx#_Toc513102837)

[Figure 5:MySQL-connector accueil 10](#_Toc513102838)

[Figure 6:MySQL-connector paramètres 10](#_Toc513102839)

[Figure 7:SPSS configurations 11](#_Toc513102840)

[Figure 8:SPSS importation de la B.D 11](#_Toc513102841)

[Figure 9: Paramètrage des variables 12](#_Toc513102842)

[Figure 10:SPSS paramètres apriori 12](#_Toc513102843)

[Figure 11:Requéte SQL 14](#_Toc513102844)

[Figure 12:SPSS Règles d'associations paramètres 15](file:///C:\Users\PC\Downloads\Projet-data-mining.docx#_Toc513102845)

[Figure 13:BD CART 16](#_Toc513102846)

[Figure 14:BD CART Final 17](#_Toc513102847)

[Figure 15:CART importance 19](#_Toc513102848)

|  |
| --- |
|  |

1. Introduction :

* L’évolution de la criminalité et de la technologie, la croissance des banques de données, nécessitent d’adapter l’analyse avec les moyens d’aujourd’hui, tels que le data mining. Celui-ci permettrait d’aider les analystes à détecter des schémas inconnus, cachés dans des banques de données, en complément de leur expérience.

1. Source de donnée :

* Notre étude se base sur une data set (voir le lien ci-dessous) préparée à partir de données réelles tirées des données socioéconomiques du recensement américain en 1990, des données sur l'application de la loi tirées de l'enquête LEMAS aux États-Unis et des données de l'UCR (FBI) en 1995. Cet ensemble de données contient un nombre total de 147 attributs et 2216 instances.

Lien :

[https://www.kaggle.com/kkanda/analyzing-uci-crime-and-communities-dataset/notebook](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.kaggle.com%2Fkkanda%2Fanalyzing-uci-crime-and-communities-dataset%2Fnotebook&h=ATO0PGylIQCw7xvVi9nA0BMQJ6tWKnu7UEOHT-NhYFV1uo43xDTc_2vXrWPjlazF1MMZnFNP_b6v7bSXM7Plkxq1ERFp3sj7gkidPlN5OD3QS1trOplgoA)

Information sur les attributs :

http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized

-- communityname: Community name - not predictive - for information only (string)   
-- state: US state (by 2 letter postal abbreviation)(nominal)   
-- countyCode: numeric code for county - not predictive, and many missing values (numeric)   
-- communityCode: numeric code for community - not predictive and many missing values (numeric)   
-- fold: fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric - integer)   
  
-- population: population for community: (numeric - expected to be integer)   
-- householdsize: mean people per household (numeric - decimal)   
-- racepctblack: percentage of population that is africanamerican (numeric - decimal)   
-- racePctWhite: percentage of population that is caucasian (numeric - decimal)   
-- racePctAsian: percentage of population that is of asian heritage (numeric - decimal)   
-- racePctHisp: percentage of population that is of hispanic heritage (numeric - decimal)   
-- agePct12t21: percentage of population that is 12-21 in age (numeric - decimal)   
-- agePct12t29: percentage of population that is 12-29 in age (numeric - decimal)   
-- agePct16t24: percentage of population that is 16-24 in age (numeric - decimal)   
-- agePct65up: percentage of population that is 65 and over in age (numeric - decimal)   
-- numbUrban: number of people living in areas classified as urban (numeric - expected to be integer)   
-- pctUrban: percentage of people living in areas classified as urban (numeric - decimal)   
-- medIncome: median household income (numeric - may be integer)   
-- pctWWage: percentage of households with wage or salary income in 1989 (numeric - decimal)   
-- pctWFarmSelf: percentage of households with farm or self employment income in 1989 (numeric - decimal)   
-- pctWInvInc: percentage of households with investment / rent income in 1989 (numeric - decimal)   
-- pctWSocSec: percentage of households with social security income in 1989 (numeric - decimal)   
-- pctWPubAsst: percentage of households with public assistance income in 1989 (numeric - decimal)   
-- pctWRetire: percentage of households with retirement income in 1989 (numeric - decimal)   
-- medFamInc: median family income (differs from household income for non-family households) (numeric - may be integer)   
-- perCapInc: per capita income (numeric - decimal)   
-- whitePerCap: per capita income for caucasians (numeric - decimal)   
-- blackPerCap: per capita income for africanamericans (numeric - decimal)   
-- indianPerCap: per capita income for native americans (numeric - decimal)   
-- AsianPerCap: per capita income for people with asian heritage (numeric - decimal)   
-- OtherPerCap: per capita income for people with 'other' heritage (numeric - decimal)   
-- HispPerCap: per capita income for people with hispanic heritage (numeric - decimal)   
-- NumUnderPov: number of people under the poverty level (numeric - expected to be integer)   
-- PctPopUnderPov: percentage of people under the poverty level (numeric - decimal)   
-- PctLess9thGrade: percentage of people 25 and over with less than a 9th grade education (numeric - decimal)   
-- PctNotHSGrad: percentage of people 25 and over that are not high school graduates (numeric - decimal)   
-- PctBSorMore: percentage of people 25 and over with a bachelors degree or higher education (numeric - decimal)   
-- PctUnemployed: percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)   
-- PctEmploy: percentage of people 16 and over who are employed (numeric - decimal)   
-- PctEmplManu: percentage of people 16 and over who are employed in manufacturing (numeric - decimal)   
-- PctEmplProfServ: percentage of people 16 and over who are employed in professional services (numeric - decimal)   
-- PctOccupManu: percentage of people 16 and over who are employed in manufacturing (numeric - decimal) #### No longer sure of difference from PctEmplManu - may include unemployed manufacturing workers ####   
-- PctOccupMgmtProf: percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal)   
-- MalePctDivorce: percentage of males who are divorced (numeric - decimal)   
-- MalePctNevMarr: percentage of males who have never married (numeric - decimal)   
-- FemalePctDiv: percentage of females who are divorced (numeric - decimal)   
-- TotalPctDiv: percentage of population who are divorced (numeric - decimal)   
-- PersPerFam: mean number of people per family (numeric - decimal)   
-- PctFam2Par: percentage of families (with kids) that are headed by two parents (numeric - decimal)   
-- PctKids2Par: percentage of kids in family housing with two parents (numeric - decimal)   
-- PctYoungKids2Par: percent of kids 4 and under in two parent households (numeric - decimal)   
-- PctTeen2Par: percent of kids age 12-17 in two parent households (numeric - decimal)   
-- PctWorkMomYoungKids: percentage of moms of kids 6 and under in labor force (numeric - decimal)   
-- PctWorkMom: percentage of moms of kids under 18 in labor force (numeric - decimal)   
-- NumKidsBornNeverMar: number of kids born to never married (numeric - expected to be integer)   
-- PctKidsBornNeverMar: percentage of kids born to never married (numeric - decimal)   
-- NumImmig: total number of people known to be foreign born (numeric - expected to be integer)   
-- PctImmigRecent: percentage of \_immigrants\_ who immigated within last 3 years (numeric - decimal)   
-- PctImmigRec5: percentage of \_immigrants\_ who immigated within last 5 years (numeric - decimal)   
-- PctImmigRec8: percentage of \_immigrants\_ who immigated within last 8 years (numeric - decimal)   
-- PctImmigRec10: percentage of \_immigrants\_ who immigated within last 10 years (numeric - decimal)   
-- PctRecentImmig: percent of \_population\_ who have immigrated within the last 3 years (numeric - decimal)   
-- PctRecImmig5: percent of \_population\_ who have immigrated within the last 5 years (numeric - decimal)   
-- PctRecImmig8: percent of \_population\_ who have immigrated within the last 8 years (numeric - decimal)   
-- PctRecImmig10: percent of \_population\_ who have immigrated within the last 10 years (numeric - decimal)   
-- PctSpeakEnglOnly: percent of people who speak only English (numeric - decimal)   
-- PctNotSpeakEnglWell: percent of people who do not speak English well (numeric - decimal)   
-- PctLargHouseFam: percent of family households that are large (6 or more) (numeric - decimal)   
-- PctLargHouseOccup: percent of all occupied households that are large (6 or more people) (numeric - decimal)   
-- PersPerOccupHous: mean persons per household (numeric - decimal)   
-- PersPerOwnOccHous: mean persons per owner occupied household (numeric - decimal)   
-- PersPerRentOccHous: mean persons per rental household (numeric - decimal)   
-- PctPersOwnOccup: percent of people in owner occupied households (numeric - decimal)   
-- PctPersDenseHous: percent of persons in dense housing (more than 1 person per room) (numeric - decimal)   
-- PctHousLess3BR: percent of housing units with less than 3 bedrooms (numeric - decimal)   
-- MedNumBR: median number of bedrooms (numeric - decimal)   
-- HousVacant: number of vacant households (numeric - expected to be integer)   
-- PctHousOccup: percent of housing occupied (numeric - decimal)   
-- PctHousOwnOcc: percent of households owner occupied (numeric - decimal)   
-- PctVacantBoarded: percent of vacant housing that is boarded up (numeric - decimal)   
-- PctVacMore6Mos: percent of vacant housing that has been vacant more than 6 months (numeric - decimal)   
-- MedYrHousBuilt: median year housing units built (numeric - may be integer)   
-- PctHousNoPhone: percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal)   
-- PctWOFullPlumb: percent of housing without complete plumbing facilities (numeric - decimal)   
-- OwnOccLowQuart: owner occupied housing - lower quartile value (numeric - decimal)   
-- OwnOccMedVal: owner occupied housing - median value (numeric - decimal)   
-- OwnOccHiQuart: owner occupied housing - upper quartile value (numeric - decimal)   
-- OwnOccQrange: owner occupied housing - difference between upper quartile and lower quartile values (numeric - decimal)   
-- RentLowQ: rental housing - lower quartile rent (numeric - decimal)   
-- RentMedian: rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)   
-- RentHighQ: rental housing - upper quartile rent (numeric - decimal)   
-- RentQrange: rental housing - difference between upper quartile and lower quartile rent (numeric - decimal)   
-- MedRent: median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)   
-- MedRentPctHousInc: median gross rent as a percentage of household income (numeric - decimal)   
-- MedOwnCostPctInc: median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)   
-- MedOwnCostPctIncNoMtg: median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal)   
-- NumInShelters: number of people in homeless shelters (numeric - expected to be integer)   
-- NumStreet: number of homeless people counted in the street (numeric - expected to be integer)   
-- PctForeignBorn: percent of people foreign born (numeric - decimal)   
-- PctBornSameState: percent of people born in the same state as currently living (numeric - decimal)   
-- PctSameHouse85: percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)   
-- PctSameCity85: percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)   
-- PctSameState85: percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)   
-- LemasSwornFT: number of sworn full time police officers (numeric - expected to be integer)   
-- LemasSwFTPerPop: sworn full time police officers per 100K population (numeric - decimal)   
-- LemasSwFTFieldOps: number of sworn full time police officers in field operations (on the street as opposed to administrative etc) (numeric - expected to be integer)   
-- LemasSwFTFieldPerPop: sworn full time police officers in field operations (on the street as opposed to administrative etc) per 100K population (numeric - decimal)   
-- LemasTotalReq: total requests for police (numeric - expected to be integer)   
-- LemasTotReqPerPop: total requests for police per 100K popuation (numeric - decimal)   
-- PolicReqPerOffic: total requests for police per police officer (numeric - decimal)   
-- PolicPerPop: police officers per 100K population (numeric - decimal)   
-- RacialMatchCommPol: a measure of the racial match between the community and the police force. High values indicate proportions in community and police force are similar (numeric - decimal)   
-- PctPolicWhite: percent of police that are caucasian (numeric - decimal)   
-- PctPolicBlack: percent of police that are africanamerican (numeric - decimal)   
-- PctPolicHisp: percent of police that are hispanic (numeric - decimal)   
-- PctPolicAsian: percent of police that are asian (numeric - decimal)   
-- PctPolicMinor: percent of police that are minority of any kind (numeric - decimal)   
-- OfficAssgnDrugUnits: number of officers assigned to special drug units (numeric - expected to be integer)   
-- NumKindsDrugsSeiz: number of different kinds of drugs seized (numeric - expected to be integer)   
-- PolicAveOTWorked: police average overtime worked (numeric - decimal)   
-- LandArea: land area in square miles (numeric - decimal)   
-- PopDens: population density in persons per square mile (numeric - decimal)   
-- PctUsePubTrans: percent of people using public transit for commuting (numeric - decimal)   
-- PolicCars: number of police cars (numeric - expected to be integer)   
-- PolicOperBudg: police operating budget (numeric - may be integer)   
-- LemasPctPolicOnPatr: percent of sworn full time police officers on patrol (numeric - decimal)   
-- LemasGangUnitDeploy: gang unit deployed (numeric - integer - but really nominal - 0 means NO, 10 means YES, 5 means Part Time)   
-- LemasPctOfficDrugUn: percent of officers assigned to drug units (numeric - decimal)   
-- PolicBudgPerPop: police operating budget per population (numeric - decimal)   
  
-- murders: number of murders in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- murdPerPop: number of murders per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- rapes: number of rapes in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- rapesPerPop: number of rapes per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- robberies: number of robberies in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- robbbPerPop: number of robberies per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- assaults: number of assaults in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- assaultPerPop: number of assaults per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- burglaries: number of burglaries in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- burglPerPop: number of burglaries per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- larcenies: number of larcenies in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- larcPerPop: number of larcenies per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- autoTheft: number of auto thefts in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- autoTheftPerPop: number of auto thefts per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- arsons: number of arsons in 1995 (numeric - expected to be integer) potential GOAL attribute (to be predicted)   
-- arsonsPerPop: number of arsons per 100K population (numeric - decimal) potential GOAL attribute (to be predicted)   
-- ViolentCrimesPerPop: total number of violent crimes per 100K popuation (numeric - decimal) GOAL attribute (to be predicted)   
-- nonViolPerPop: total number of non-violent crimes per 100K popuation (numeric - decimal) potential GOAL attribute (to be predicted)

1. Problématiques:

* Ce travail met au centre de son propos les techniques de fouille de données comme moyens de maîtriser la masse documentaire, identifie deux problématiques purement décisionnelles distinctes qu’on va traiter à l’aide des deux outils ‘SPSS’ et ‘R’ :
* Les principales causes qui concrétisent la criminalité.
* L’aide à la prise des décisions en tout ce qui concerne les protocoles policières (la fixation du nombre de policier par population).

1. Processus de travail:
2. Les principales causes qui concrétisent la criminalité:
   1. Sélection des données:

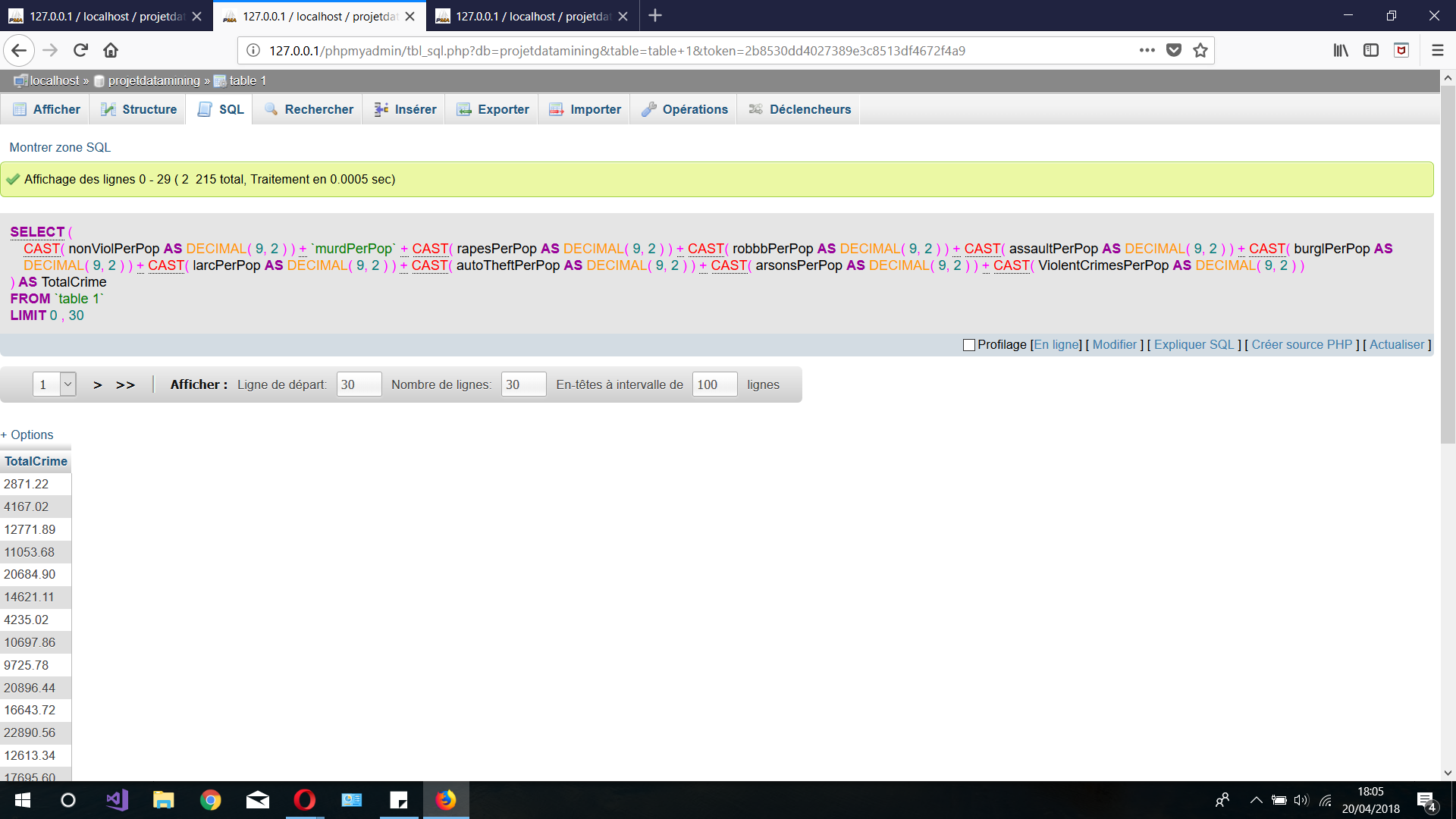
* Dans cette première phase, On va sélectionner les données sur lesquelles va porter notre travail, c’est-à-dire choisir les colonnes qui semblent intéressantes pour l’exécution des différents algorithmes de fouille des données…
* Cette tâche a été réalisée à l’aide de l’exécution de différentes requêtes SQL, qui nous ont permis finalement de garder que les colonnes suivantes :
* Communityname
* Population
* Householdsize
* Racepctblack
* racePctWhite
* racePctAsian
* racePctHisp
* agePct12t21
* agePct12t29
* agePct16t24
* agePct65up
* PctLess9thGrade
* PctNotHSGrad
* PctEmploy
* PolicPerPop
* murdPerPop
* rapesPerPop
* robbbPerPop
* assaultPerPop
* burglPerPop
* larcPerPop
* autoTheftPerPop
* arsonsPerPop
* ViolentCrimesPerPop
  1. Prétraitement des données :
* Pour assurer la cohérence des données, nous avons eu recours à la technique de formatage des données pour qu’elles soient traitables par les différents algorithmes de fouilles des données.
* D’abord, On a commencé par changer les types des données qui seront traitées, donc on a utilisé la fonction CAST (de ‘varchar’ vers ‘decimal’).

Figure 1:types des données

En second lieu, On s’est retrouvé dans l’obligation de réduire le nombre des colonnes donc on a fusionné certaines variables.

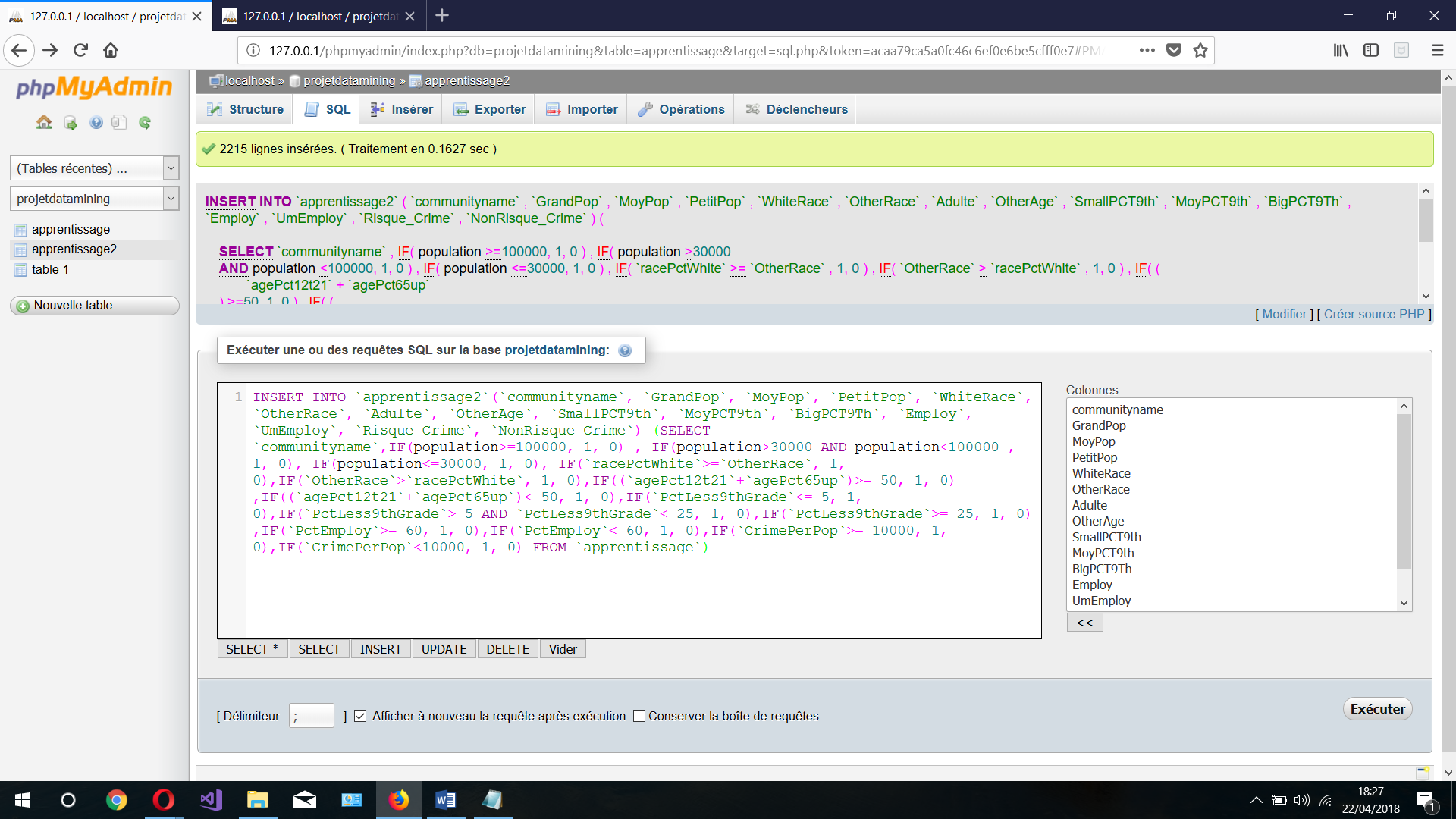
* En dernier lieu, on a anticipé la phase de ‘binairisation’ des données, dans cette phase on a remplacé les valeurs de différentes colonnes par des valeurs binaires (0-1) fixés grâce à des comparaisons faites par rapport à des valeurs bien étudiées qu’on a fixé. 

Figure 2: Binairisation des données

* Finalement, nous avons abouti à un résultat comme le montre la figure ci-dessous, cette base de données va être introduite comme étant un flux d’entré pour l’algorithme Aprioriqui est un algorithme d'[exploration de données](https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es), se basant sur les règles d’association.
* La force d’une règle d’association peut être mesurée en utilisant son support et sa confiance.

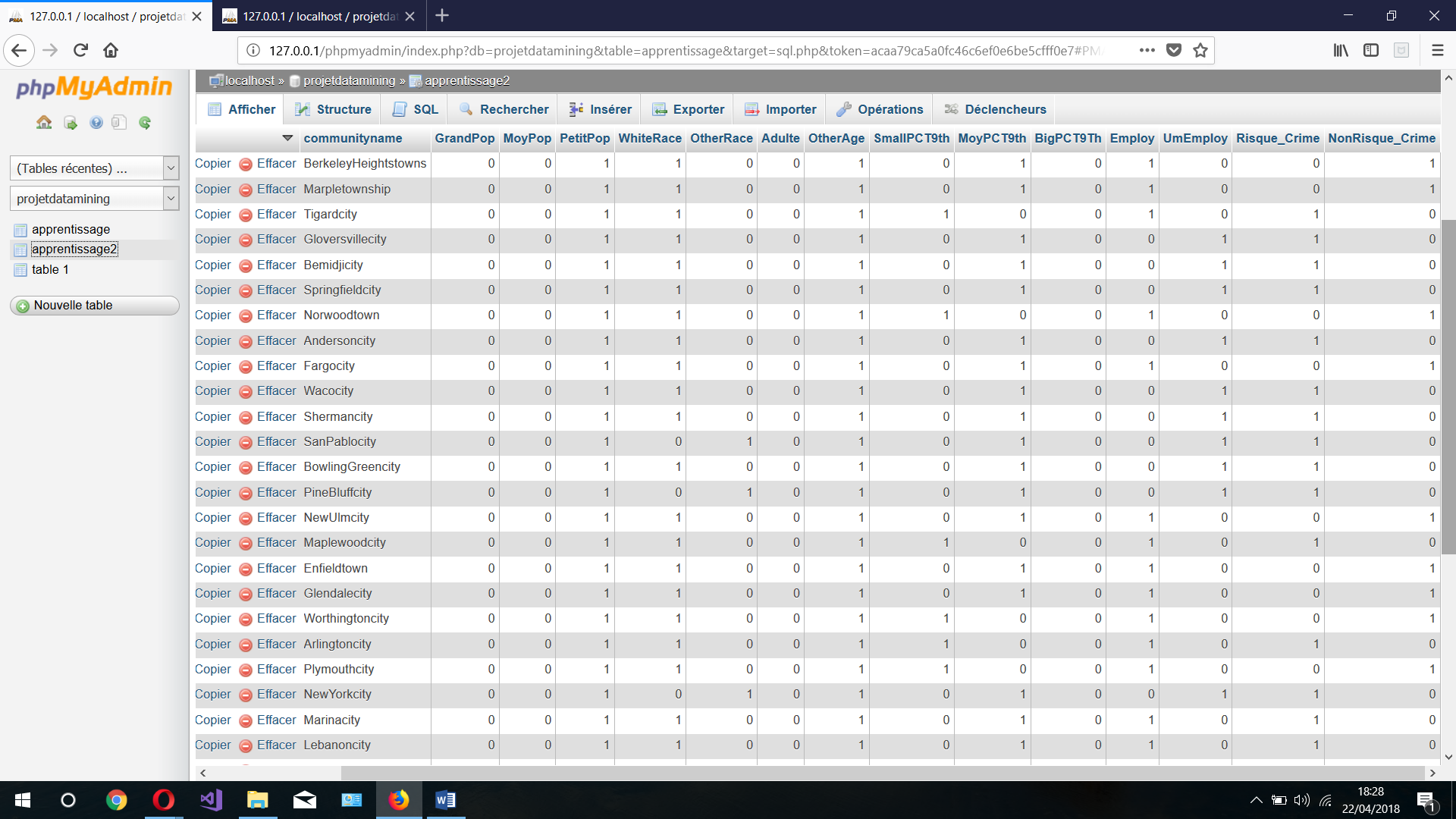


Figure 3:Base de données binairisée

* 1. Fouille de données :

Implémentation Apriori à l’aide de l’outil SPSS :

* IBM SPSS Modeler est une plateforme d'analyse prédictive qui vous aide à créer rapidement des modèles d'analyse prédictive et intégrer les éléments prédictifs aux décisions des individus, des équipes, des systèmes et l'entreprise. Il fournit toute une gamme d'algorithmes et de techniques d'analyse avancés, dont l'analytique textuelle, l'analyse d'entité, la gestion et l'optimisation des décisions, et fournit des connaissances quasiment en temps réel.
* Pour éviter toutes erreurs relatives à la compatibilité, ainsi que les problèmes des mises à jour de la base des données nous avons eu l’occasion d’utiliser MySQL-Connector,Un aspect important qui concerne presque toutes les bases de données est le pilote ODBC (également connu sous le nom Open DatabaseConnectivity) destiné à se connecter aux serveurs de base de données SQL. Il définit un ensemble d'appels de fonctions, de codes d'erreur et de types de données pouvant être utilisés pour développer des programmes indépendants de la base de données.

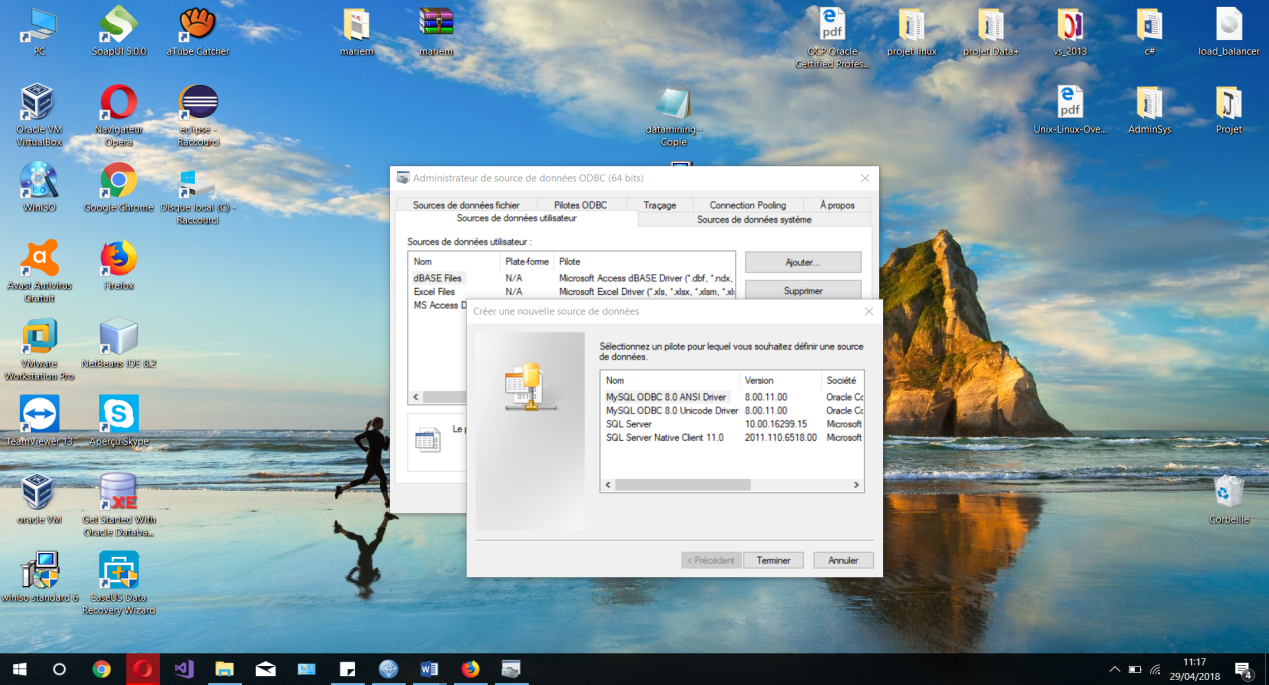


Figure 4:MySQL-connector accueil

* Dans cette interface, nous avons introduit les paramètres de notre base de données

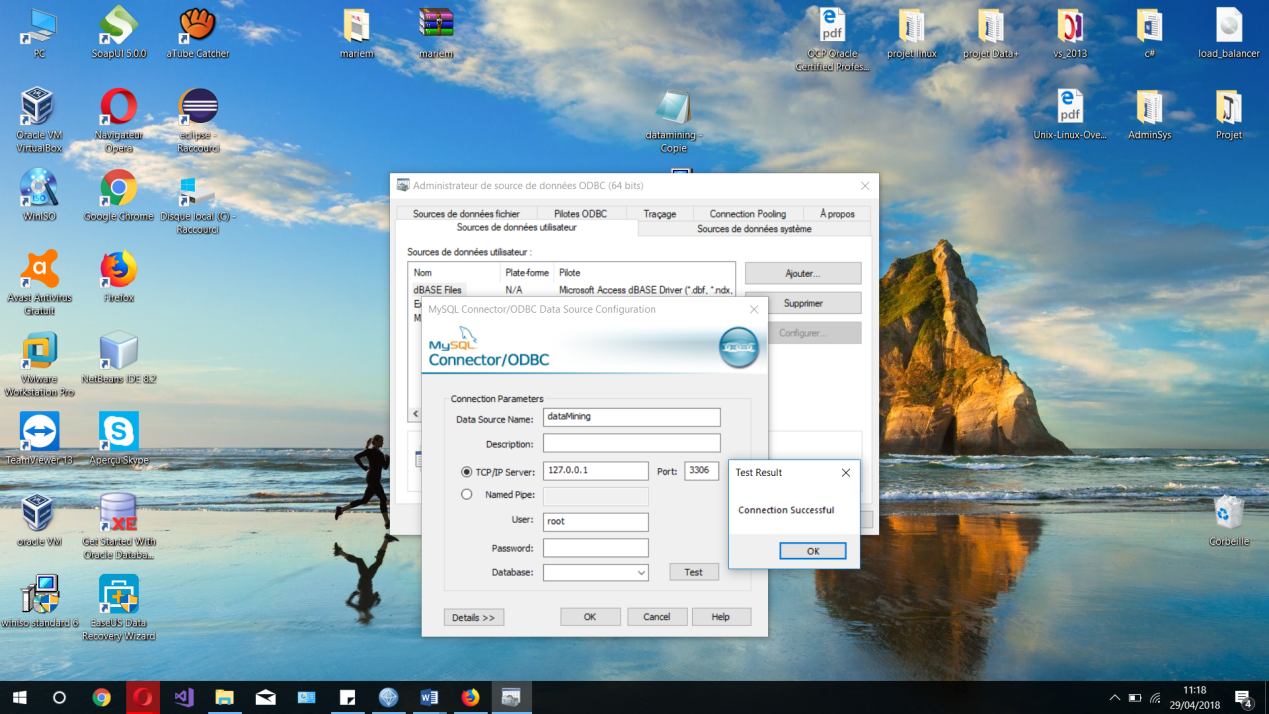


Figure 5:MySQL-connector paramètres

* Dans l’étape suivante, on a fait la liaison entre l’outil SPSS et MY-SQL-connector

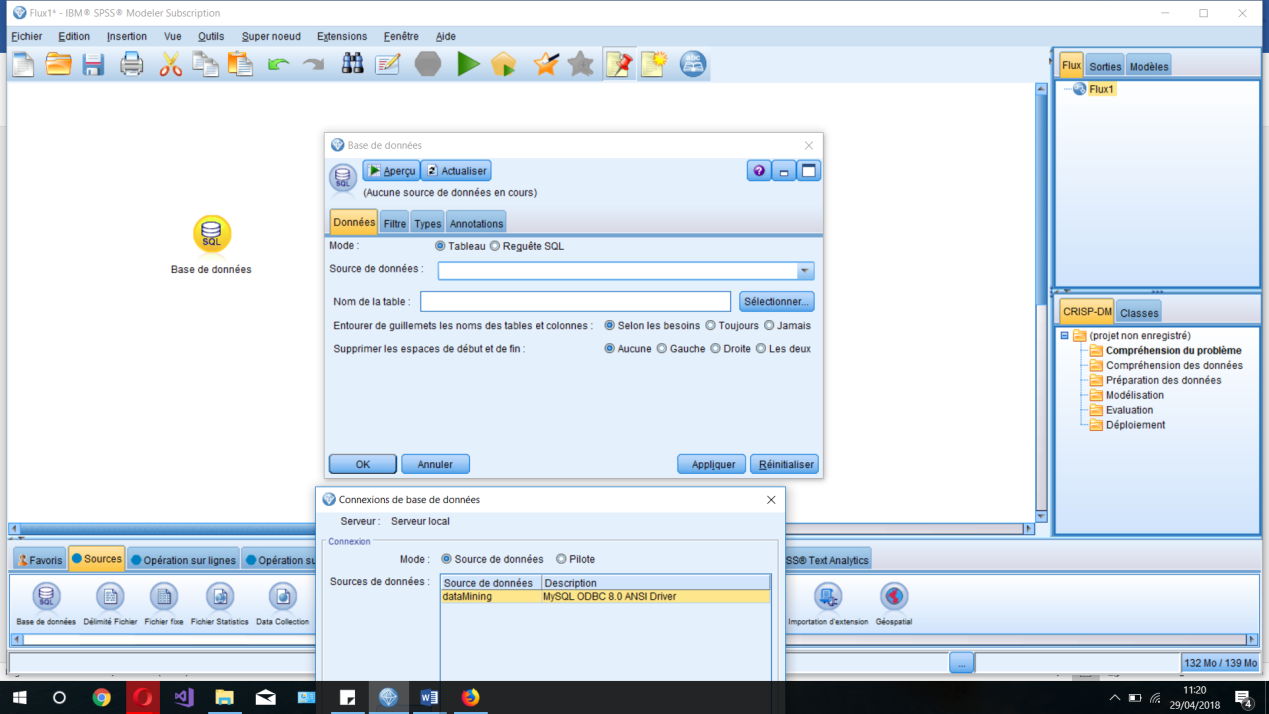


Figure 6:SPSS configurations

* Apréson a importé les données à partir de la session ouverte dans MySQL-connector

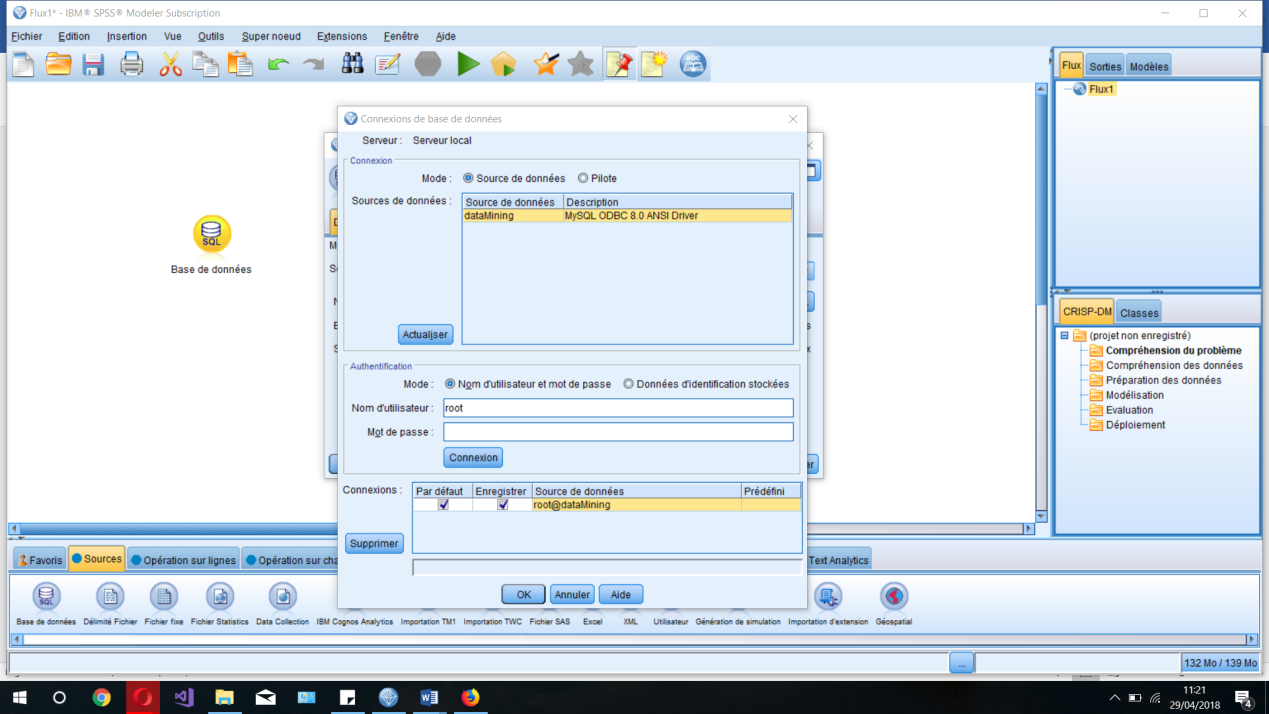


Figure 7:SPSS importation de la B.D

* Avant le lancement de l’algorithme Apriori, une petite configuration est requise, il suffit de distinguer les variables d’entrées de celle qui est notre cible, ainsi qu’attribuer les différents mesures qui vont être utilisées (indicateur dans notre cas)

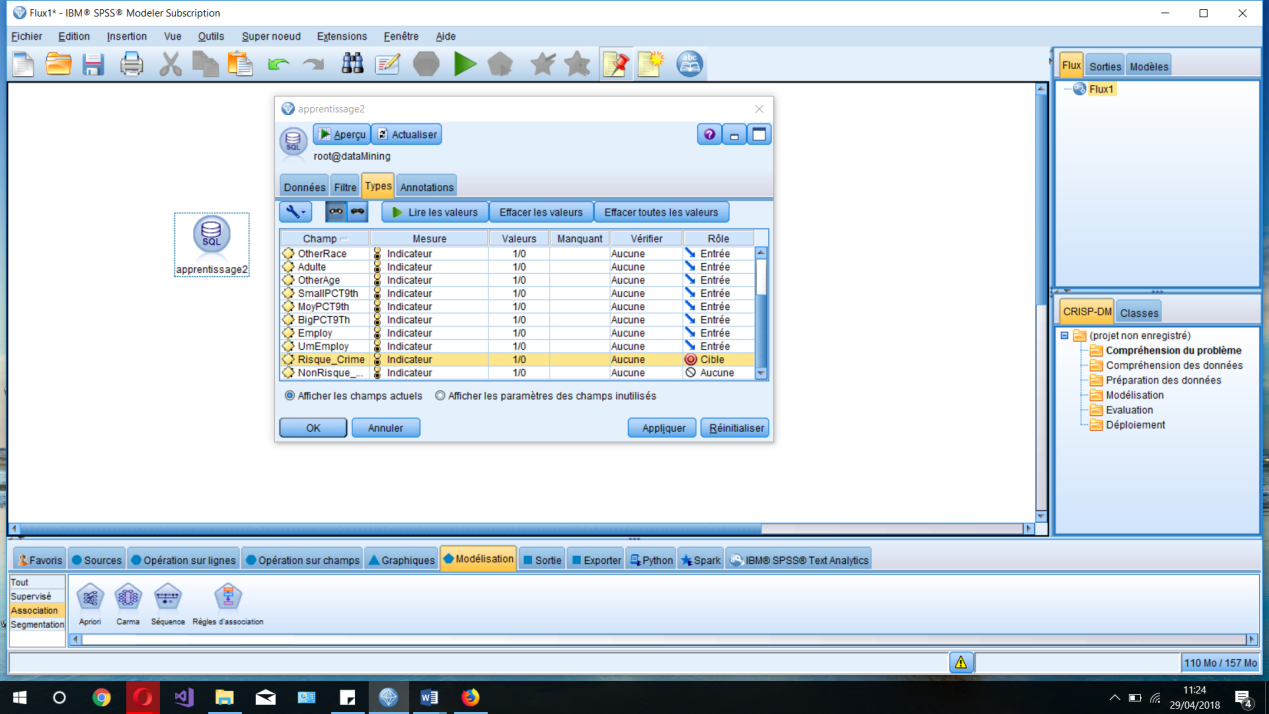


Figure 8: Paramètrage des variables

* Nous avons aussi le choix pour fixer différents paramètres qui concernent la génération du modèle, et pour ce faire on a attribué les valeurs suivantes

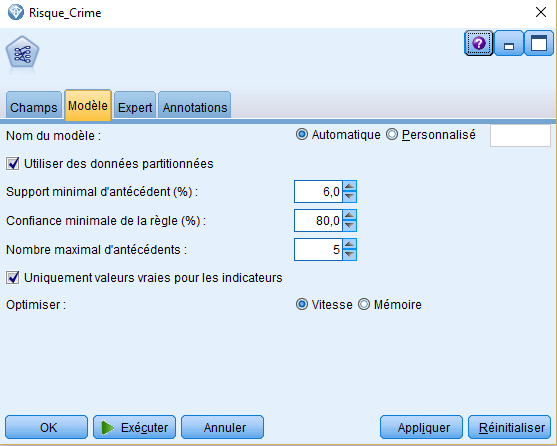
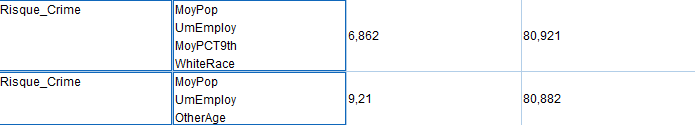
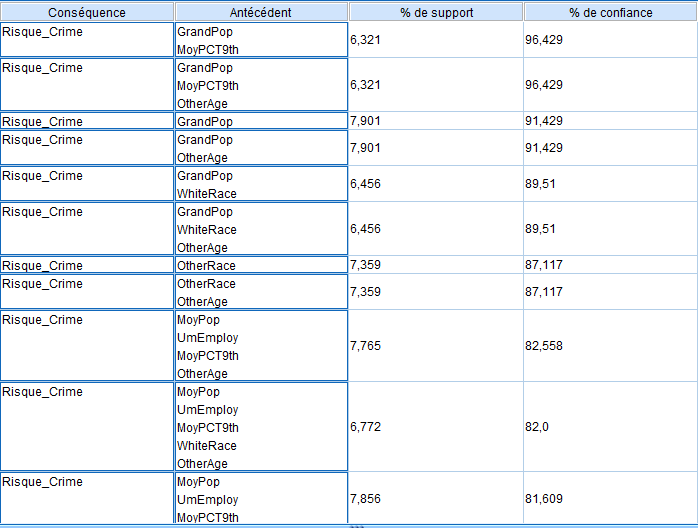


Figure 9:SPSS paramètres apriori

* Finalement on a abouti à ces résultats.



* On constate que les principales causes de la criminalité sont ces différentes combinaisons d’antécédents, en effet par exemple prenons ce cas

MoyPop,UmEmploy,OtherAge->Risque\_Crime, cette règle a comme valeur de support 9.21 % et une valeur de confiance 80.882 %, par ce fait on constate que la présence de ces différents facteurs favorise le risque de crime, donc il existe une forte relation qui lie ces différents causes à cette conséquence, la crédibilité de cette relation peut être jugé grâce aux différents valeurs de support et de confiance.

Implémentation Apriori à l’aide de l’outil R :

* R est un logiciel permettant de faire des analyses statistiques et de produire des graphiques. Mais R est également un langage de programmation complet, c’est cet aspect qui fait que R est différent des autres logiciels statistiques.
* R étant un logiciel qui sait lire les fichiers .csv sous une certaine forme particulière, pour ce faire on a eu l’idée de lancer certaines requêtes SQL pour migrer vers la structure adéquate.

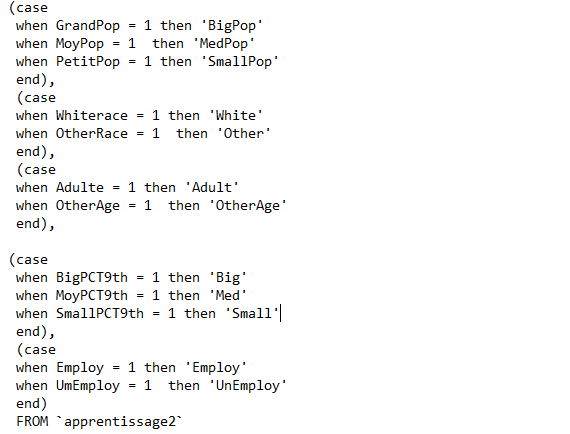


Figure 10:Requéte SQL

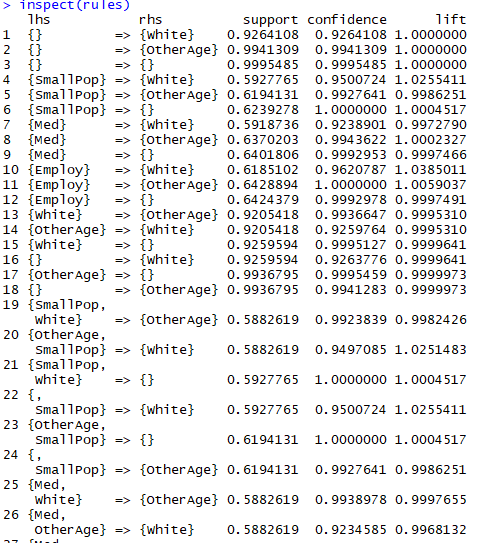
* On a commencé par lire le fichier et le visualisé



* Ensuite, grâce à cette commande on a exécuté l’algorithme apriori en passant en paramètre la valeur du support ainsi que celle de la confiance



* Finalement, on a pu degager ces différents régles d’associations suivantes :



Dégagements des règles d’association à l’aide de l’outil SPSS :

* Dans cette phase on a essayé de tester l’algorithme de dégagement des règles d’associations mis en place par l’outil SPSS, On a fixé les paramètres comme la figure ci-dessous l’indique :

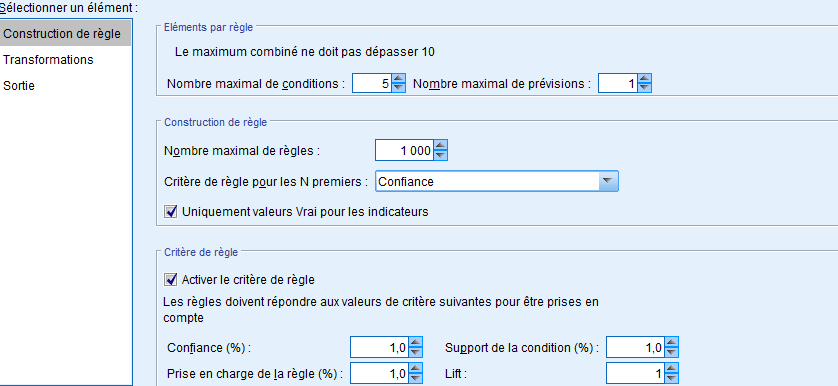
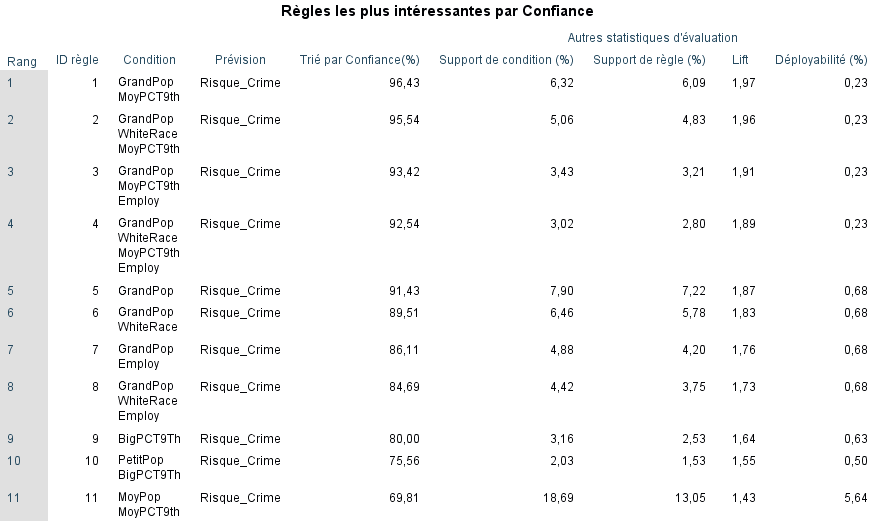


Figure 11:SPSS Règles d'associations paramètres

* Par la suite, on a réussi à dégager les règles suivantes :



1. la fixation du nombre de policier par population
   1. Sélection des données

* Dans cette partie on va prédire l’intervalle adéquat de policier à fixer pour une zone quelconque.
* Cette partie est fortement dépendante de la partie précédente puisque c’est grâce à elle qu’on a identifié les différentes causes de criminalité, donc sur la base de différentes valeurs aperçus on va mettre en œuvre notre base de données.

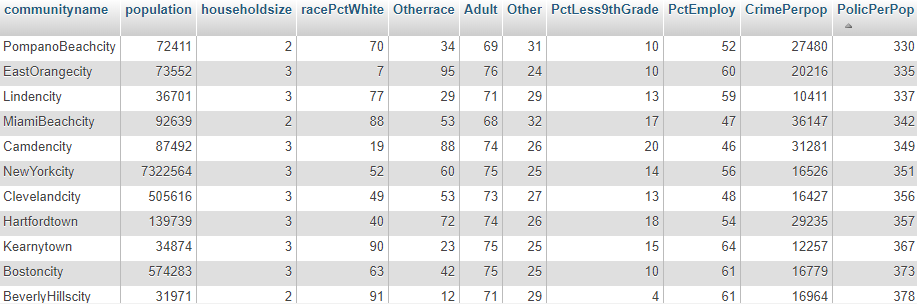


Figure 12:BD CART

* 1. Prétraitement des données
* Lors de cette étape, On a attribué les mêmes valeurs des colonnes précédemment utilisé, et on a remplacé les valeurs qui peuvent être présente pour PolicePerPop par quatre intervalles (A,B,C,D) où :
* 0<A=<160 160<B=<210 210<C=<500 500<D
* Voici donc un échantillon de la base :

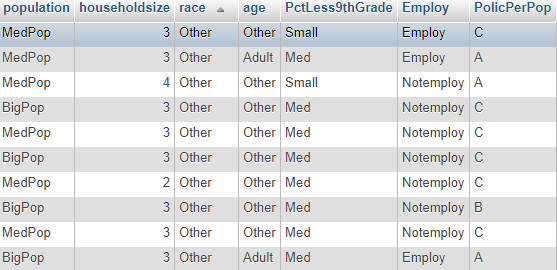
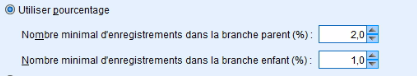


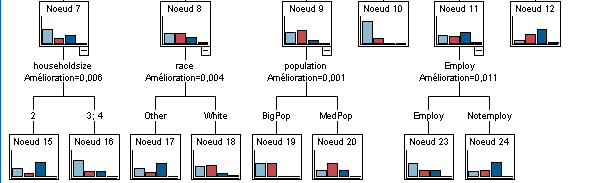
Figure 13:BD CART Final

* 1. Fouille des données

Implémentation CART à l’aide de l’outil SPSS :



* CART utilise une construction gloutonne de l'arbre de décision avec un critère de gain basé sur la fonction de Gini avec un élagage effectué.
* Une base de test sera mise en place automatiquement par l’outil, à fin de vérifier la crédibilité des règles généré par l’arbre dans notre cas 99 parmi 340 lignes sont réservées pour la base de test le reste sera considérés comme étant une base d’apprentissage.



* Si age=adult, pctless9thGrade=med,population=bigpop et householdsize=2donc on doit fixer 210<C=<500
* Si age=adult, pctless9thGrade=med,population=bigpop et householdsize=3:4 donc on doit fixer 0<A=<160
* Si age=adult, pctless9thGrade=med,population=medpop et race=otherdonc on doit fixer 160<B=<210
* Cette figure nous illustre l’importance des prédicteurs
* On constate que la population et l’âge agissent directement sur la fixation du nombre de policier.

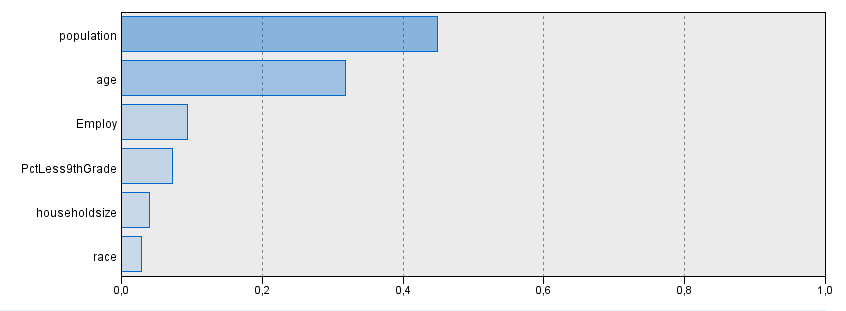


Figure 14:CART importance

Implémentation CART à l’aide de l’outil R :

* Dans cette partie on va utiliser la bibliothèque ‘rpart’, Celle-ci contient les fonctions permettant de construire et exploiter un arbre de décision :
* Chargement du package « rpart ».
* Chargement des données : **read.csv.**

* Utilisation de la fonction rpart (), en passant en paramètres les configurations adéquates.

Vu qu’on dispose d’un nombre de ligne qui n’est pas assez grand (340 lignes), et qu’on a traité cette problématique avec l’outil SPSS en travaillant avec une base de test, cependant, on a choisi de garder toutes les lignes pour la base d’apprentissage mais il existe une fonction qui traite ce processus de découpage de base (CreateDataPartition()).

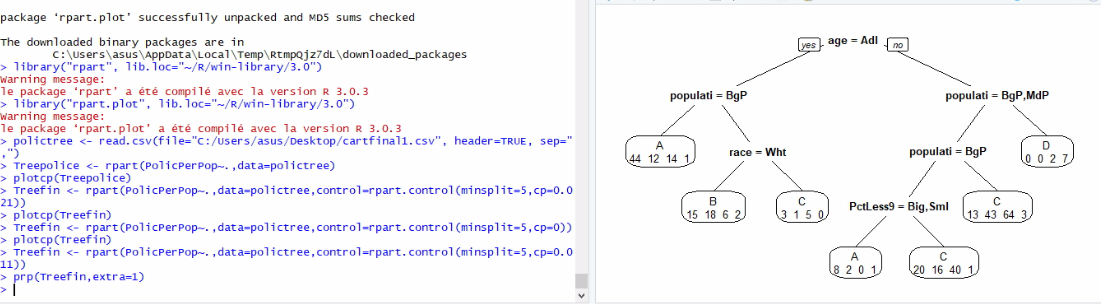
Le résultat observé est donc comme le montre la figure ci-dessous :

Figure :R CART

* On Constate qu’il ne s’agit pas exactement du même résultat observé lors de l’exécution de cet algorithme à l’aide de SPSS, ça c’est due à la différence entre les paramètres passé en argument, aussi le nombre des niveaux qui n’est pas le même.

Ici, Par exemple, si âge=adulte et population=BGp alors le nombre de policier par 100k citoyens doit être 0<A=<160.

Si age=other et population=BG,MdP alors le nombre de policier par 100k citoyens doit être 500<D.

1. Conclusion :

Ce projet s’est révélé très enrichissant dans la mesure où il a consisté en une approche concrète du métier d’ingénieur. En effet, la prise d’initiative, le respect des délais et le travail en équipe seront des aspects essentiels de notre futur métier.

De plus, il nous a permis d’appliquer nos connaissances en fouille de donnée à un domaine pratique, qui se révèle aujourd’hui d’intérêt général au vu de l’augmentation de quantité de données, ainsi le nombre important des problèmes décisionnels à traiter.

 Ce projet avait pour but de nous familiariser avec les divers outils de fouille de données, Du coup, il nous a appris beaucoup de nouvelles choses.

Bibliographie :

# <http://www-01.ibm.com/support/docview.wss?uid=swg27046871>

# <http://www.rdatamining.com/>

# <https://www.kaggle.com/>

# <https://fr.wikipedia.org/wiki/Algorithme_APriori>

# <https://github.com/readyready15728/communities-and-crime>

# <https://www.youtube.com/watch?v=HYV2aPHhmVg>

# <http://www.base-plus.fr/data-mining/>

# <https://www.youtube.com/watch?v=Q4fpJsZ9veY>

# <https://www.mysql.com/fr/products/connector/>

# [towardsdatascience.com/predict-customer-churn-with-r-9e62357d47b4](http://towardsdatascience.com/predict-customer-churn-with-r-9e62357d47b4)