Data Mining Project Report

**2022 /2023**

# **1. Source of data**

Data was supplied by city of Chicago web site, which is the city of Chicago's official site, and the name of the data set is public passenger vehicle licenses. A public passenger vehicle is a vehicle used for the transportation of passengers for hire by a public chauffeur, the list of public passenger vehicles includes licensed taxicabs (medallions), liveries, ambulances, medicars, charter-sightseeing buses, and pedicabs.

The original data set contains seventeen column and eleven thousand, three hundred and forty seven record, the columns names are vehicle type, public vehicle number, status, vehicle make, vehicle model, vehicle model year, vehicle color, vehicle fuel source, wheelchair accessible, company name, address, city, state, zip code, taxi affiliation, taxi medallion license management, and record id.

# **2. Data Cleaning**

It is essentially that all data mining algorithms accept only cleaned data for applying these algorithms, and that is why we cleaned our dataset from missing data and inappropriate negative values. The columns we have cleaned them were: vehicle make, vehicle Model, vehicle Model Year, vehicle Color, city, state, but some columns were complete and cleaned from the source of the data including: vehicle Type, public vehicle number, status, vehicle fuel source, wheel chair accessible, company name, record ID. In addition there were some columns including address, Zip Code, Taxi Medallion License Management that are not important and very hard to preprocess so we delete these columns.

* **Inappropriate negative values:**

The column that contains inappropriate values was vehicle mode. We found numeric values that includes (350,650,250), but this column accept characters, so we solved this problem by reading the data set and then we found that E350 is the correct value for 350, E250 instead of 250, and F650 instead of 650 .

* **Missing values:**

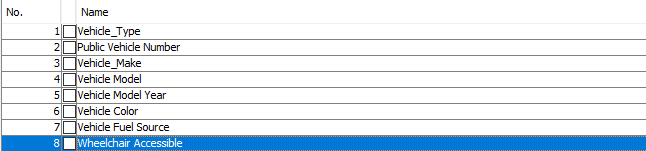
Vehicle Make, Vehicle Model, Vehicle Model Year, Vehicle Color, City, State columns contain null values.

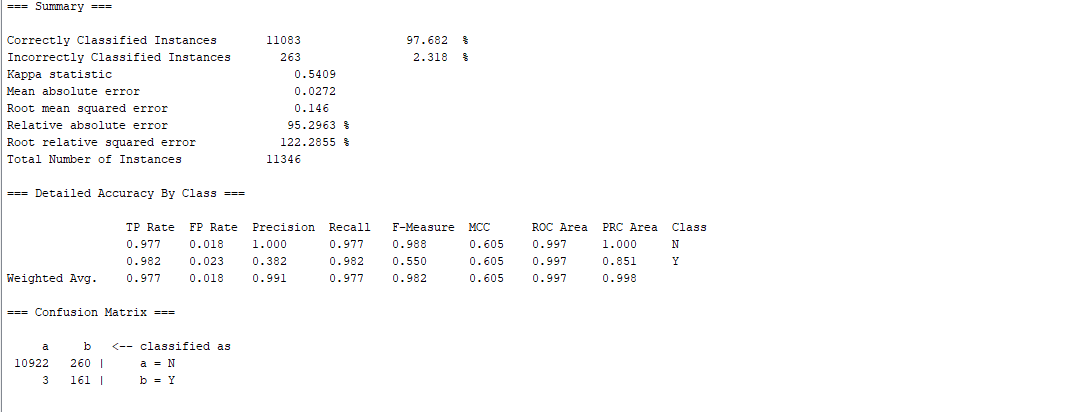
Filling in missing values in vehicle make, vehicle model columns was based on the vehicle type column for example if the vehicle type was ambulance the most common values for vehicle make, vehicle model columns were FORD, E350 in consequence, and like this way the missing values in this column were completed.

For vehicle model year column we dealt with null values by comparing them with vehicle model and then we filled them with suitable year.

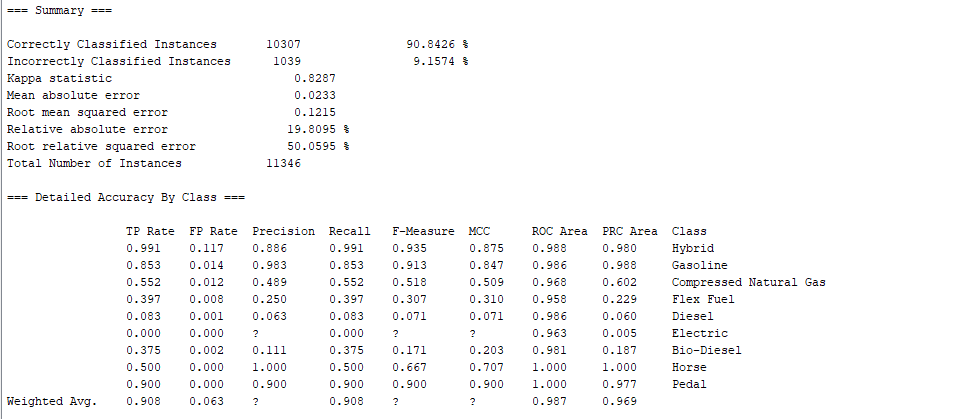
We complete the null values in Color column with black because it is the most repeated value in the dataset. And for state columns all the values was IL so we filled the null values with IL.

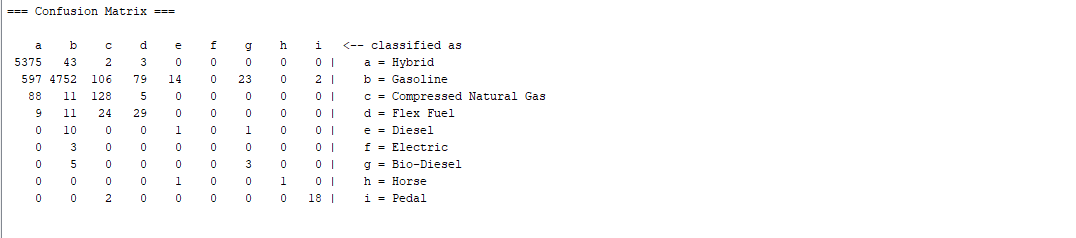
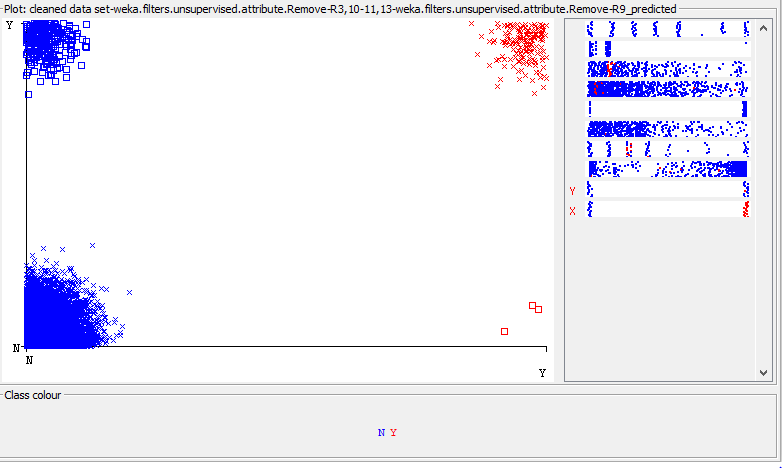
**3.Naive Bayes-type algorithm**

✔**we selected the important columns because the size of data is large.by ( 1-Wheelchair Accessible , 2- Vehicle Fuel Source)**



* **نسبة التنبؤ الصحيحة = 97.682% - نسبة التنبؤ الخطاء=2.318% .**
* **قسم التصنيف الى 2 Y.Nاعتماد على عمود Wheelchair Accessible .**
* **القيم الصحيحة هي التي تنبأ بها في القطر الرئيسي .**
* **القيم الخطأ هي التي تنبأ بها في القطر ثانوي** .



**Classifier errors**

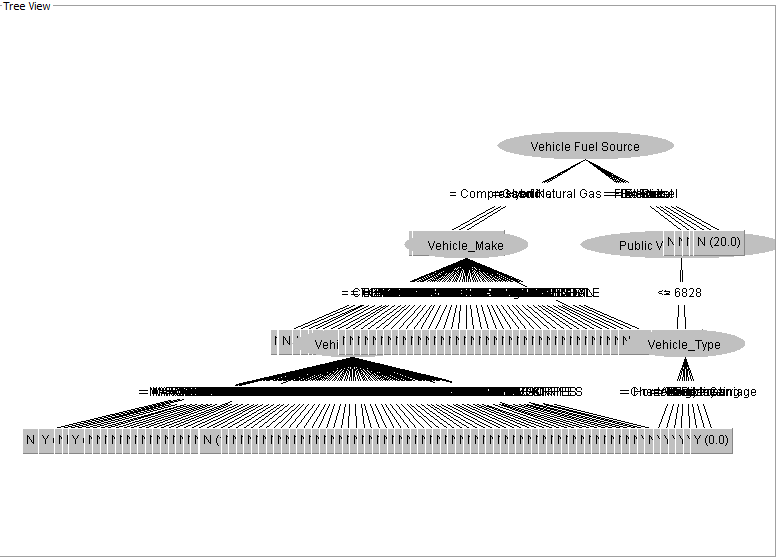
# **4. j48-type algorithm**

* **نسبة التنبؤ الصحيحة = 99.3654% - نسبة التنبؤ الخطاء=0.6346% .**
* **قسم التصنيف الى 2 Y.Nاعتماد على عمود Wheelchair Accessible .**
* **القيم الصحيحة هي التي تنبأ بها في القطر الرئيسي .**
* **القيم الخطأ هي التي تنبأ بها في القطر ثانوي.**
* **حجم الشجرة =147 .**
* **عدد العقد في الشجرة=****142**
* **العدد الكلي instances = 11346 .**

# 

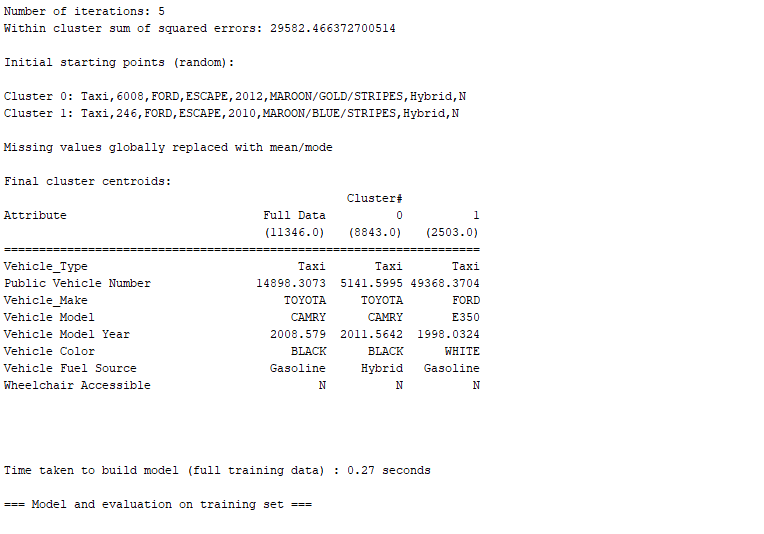
**✔The right values are the ones in the main diameter.**

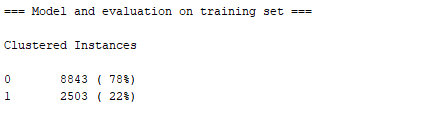
# **Classifier errors**

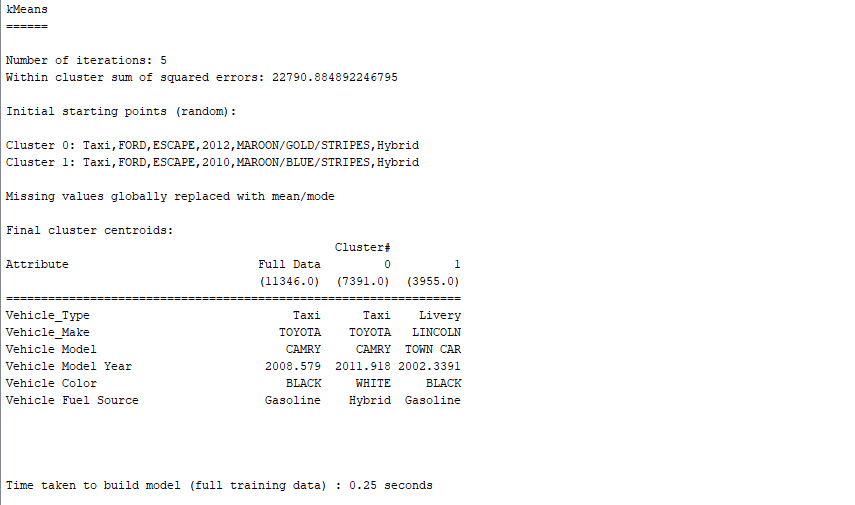
**Tree view**

**طبعا الشجرة كبيرة لهذا ظهرت بهذا الشكل.**

# **5.Simplekmeans-type algorithm**



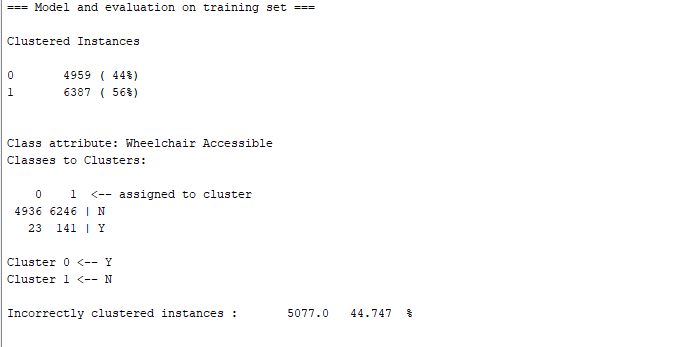




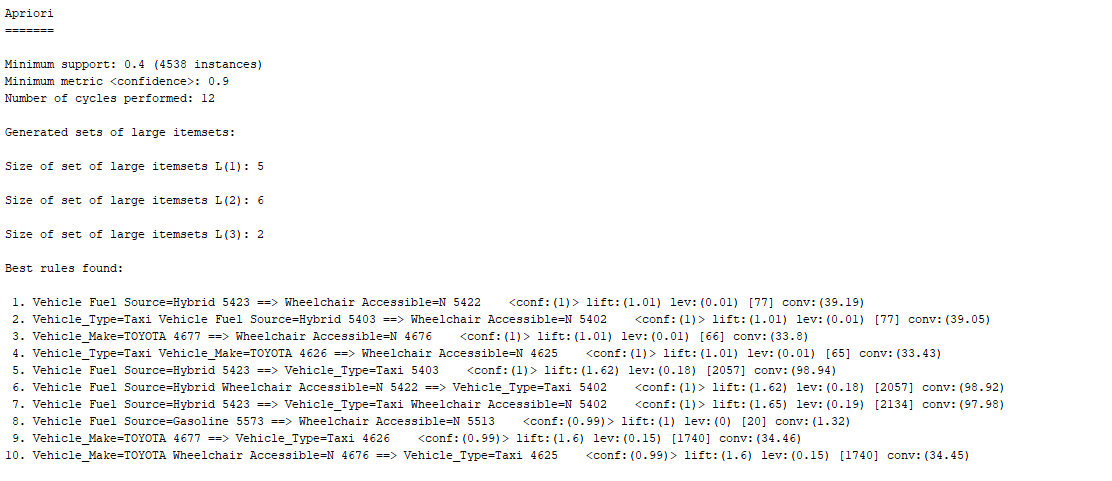
* **عدد الدورات =5 , طبعا يتوقف عند حالة الاستقرار.**
* **قسم التصنيف الى 2 Y.Nاعتماد على عمود Wheelchair Accessible .**
* **الوقت المستغرق = 0.27 من الثانية.**
* **التصنيف الأول = 78% .**
* **التصنيف الثاني = 22%**

# **Cluster assignments**

# 

**evaluation**

# **6.Apriori-type algorithm**

**✔After converting a colunn Vehicle Model Year from numeric to nominal.**

* **Minmux rupport =0.4 means 40%**
* **confidence=0.9 means 90%**
* **عدد الدورات =12**
* **Frequent itemset = 5+6+2=13**
* **Left>1 than it is positive dependent**
* **Left=0 than it is independent**
* **Left<0 than it is negative independent**
* **العشر القواعد تعتبر قوية لان con اكبر او يساوي 90%**

# **7. Conclusion**

To conclude, this data set includes information about vehicles hired by public chauffeur and by analyzing this data it may help in determining the most widely used Wheelchair Accessible and Vehicle Fuel Source with more details about vehicles and results of analysis well help in making decisions to improve the work in this field.