Tóth Ferenc - gépi tanulás beadandó

A projektben használt adathalmaz: https://archive.ics.uci.edu/ml/datasets/Automobile

Ez az adathalmaz autóknak tárolja az adatait, és a Ward's Automotive Yearbook 1985-ös évkönyvének adataiból épül fel.

26 attribútummal rendelkezik, ezek a következőképpen néznek ki:

1. symboling: -3, -2, -1, 0, 1, 2, 3. (A megbízhatóságot írja le a 3 a legrosszabb, a -3 a legjobb)
2. normalized-losses: continuous from 65 to 256.
3. make:

   alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo

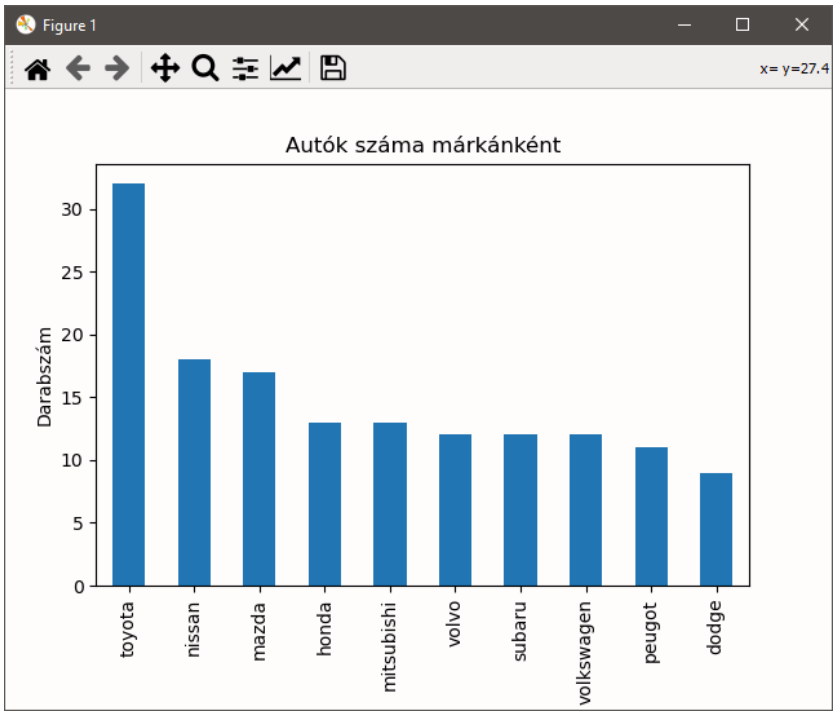
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

A fájl elején a dataset „hibáit” kezdtem el kijavítani:

- a horsepower oszlop kérdőjeleit távolítom el és helyettesítem az oszlop átlagos értékével.
- a normalized-losses oszlop NaN értékeit cserélem ki a fillna segítségével az oszlop átlagos értékével
- a num-of-doors oszlopnál a string adatokat átalakítom int típusúvá, hogy tudjam használni a későbbiekben
- a num-of-cylinders oszlopnál úgyszintén a string adatokat alakítom át int típusúvá
- a num-of-doors oszlopnál maradt 2 NaN érték, ezeket kitöltöm 4-el, mivel az autók közel 90%-a 4 ajtós

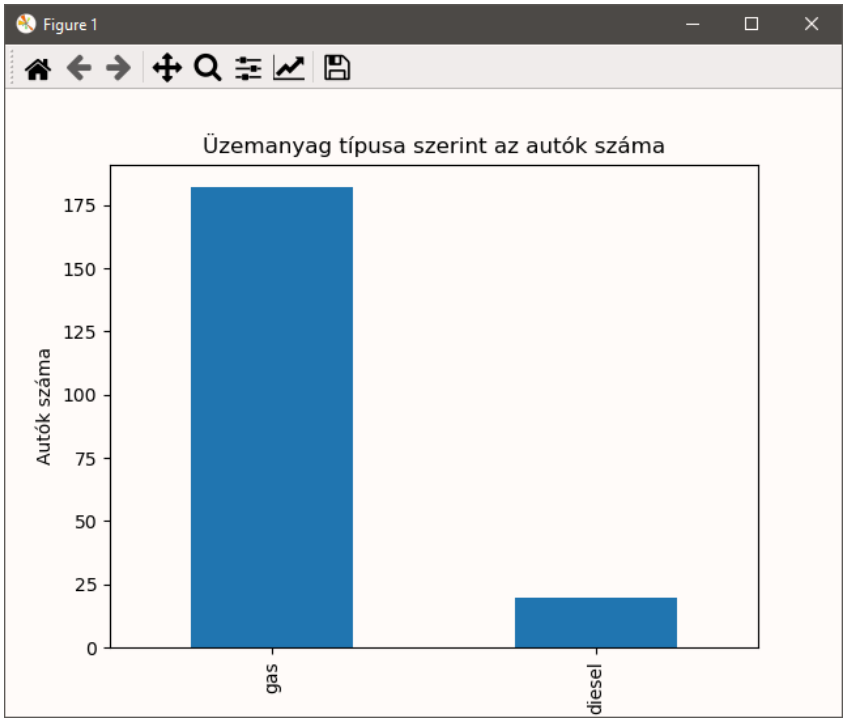
Ezután készítettem különböző érdekes diagramokat:

1.



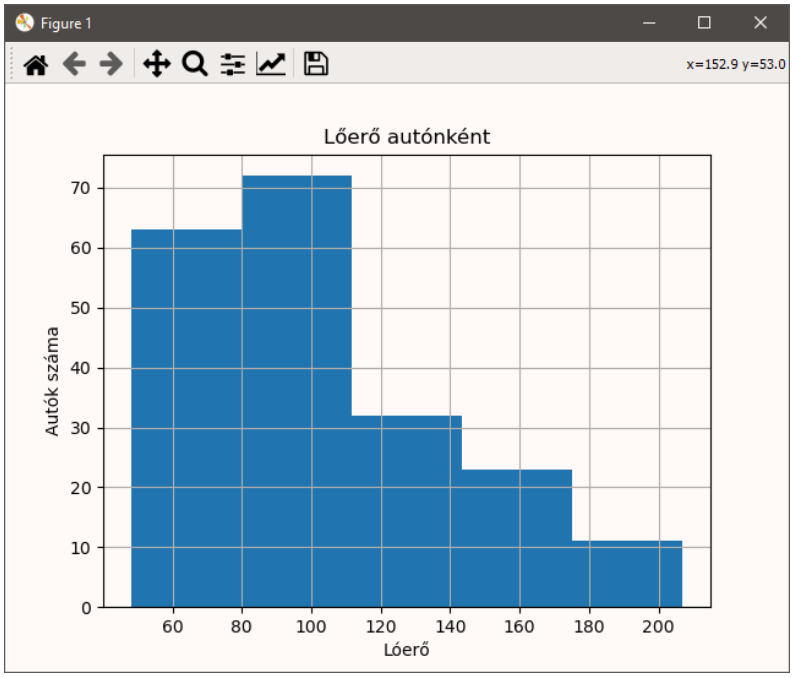
Jól látható ezen az oszlopdiagramon, hogy itt a különböző autómárkáknak van meghatározva a darabszáma, ebből következik, hogy toyotából van a legtöbb, dodgeból pedig a legkevesebb.

2.



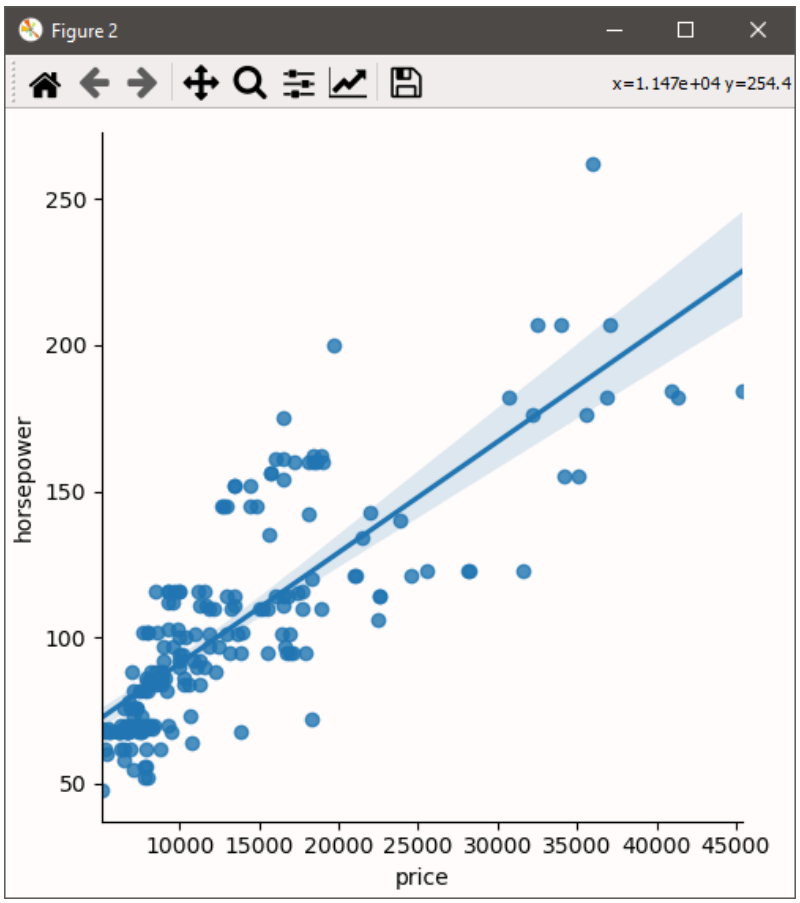
Kirajzoltam az üzemanyag típusa szerinti autók számát, ebből leolvasható, hogy a dataset lényegesen több benzines autót tartalmaz.

3.



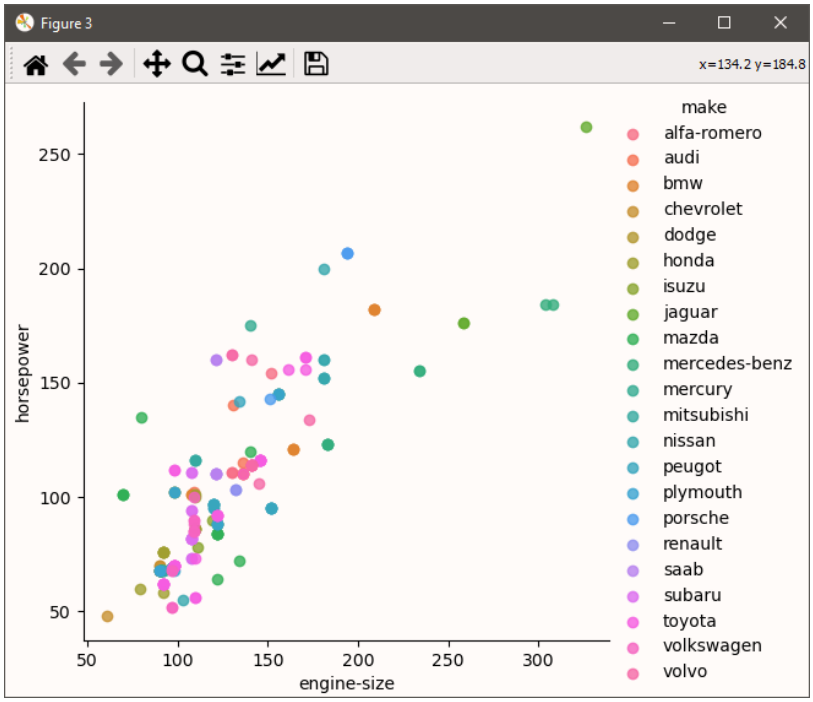
Ezen a diagramon az autók teljesítménye (lóerőben) és azoknak száma olvasható le, a diagram szerint a ~100 lóerős autóból van a legtöbb.

4.



A diagramon az autó árához viszonyított teljesítmény olvasható le.

5.

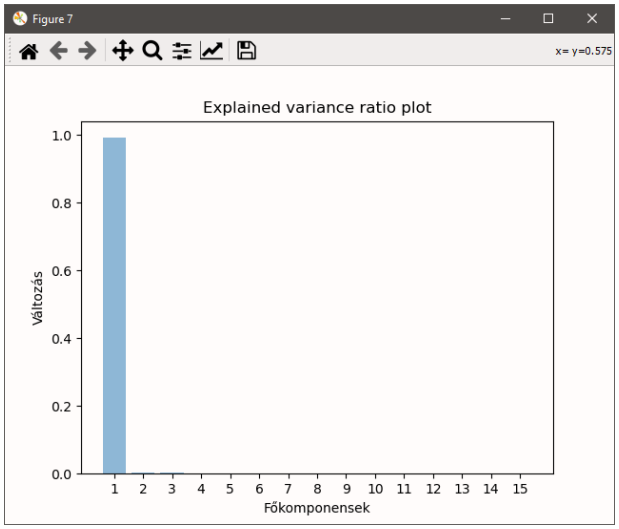


Hasonlít az előzőhöz, csak itt a motor „mérete” (hengerűrtartalma) és a lóerő viszonya olvasható le. Logikusnak tűnik, hogy minél nagyobb egy autó motorja, azzal arányosan nő az ereje is, ez főként igaz is.

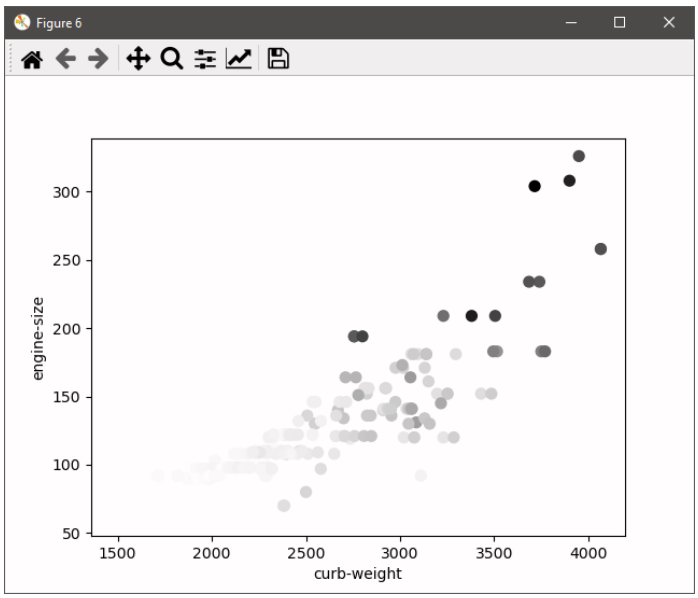
Létrehoztam egy másolatot a datasetről, amelyben a márkák neveit átalakítottam int típusúvá, ezt is azért, hogy tudjam használni ezeket az adatokat is a későbbiekben

Ezután szétszedtem, és készítettem belőle egy data és egy target állományt. A target az autók árát tartalmazza.

Ezután egy pca-val próbálkoztam, a főkomponensek variációjának oszlopdiagramja a következő:



Kiderült az is, hogy az autó árát leginkább a motor mérete és az autó súlya határozza meg, erről is készült egy diagram:



Tanítás:

LogisticRegression:

r2_score is : 0.6173372956191914 → 61,7%

LinearRegression:

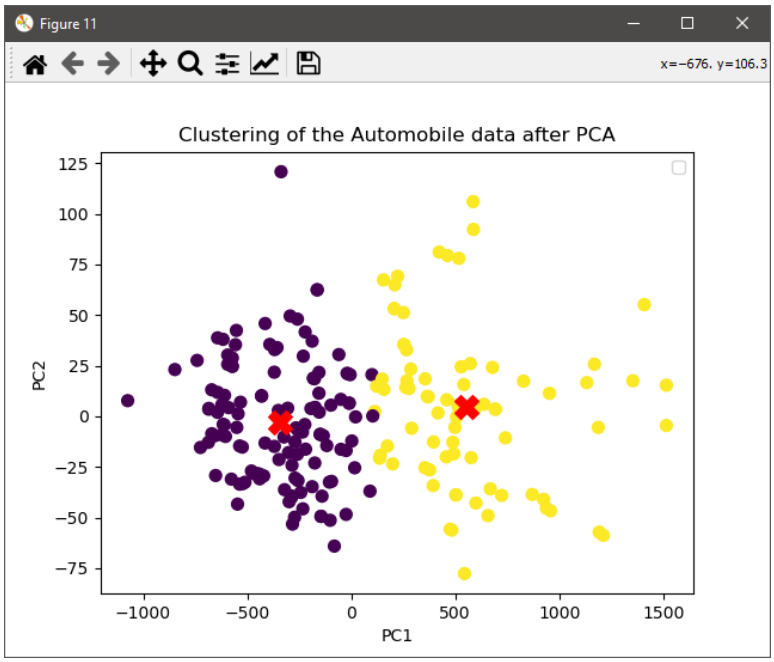
r2_score is : 0.795423348943254 → 79,5%

GradientBoostingRegressor:

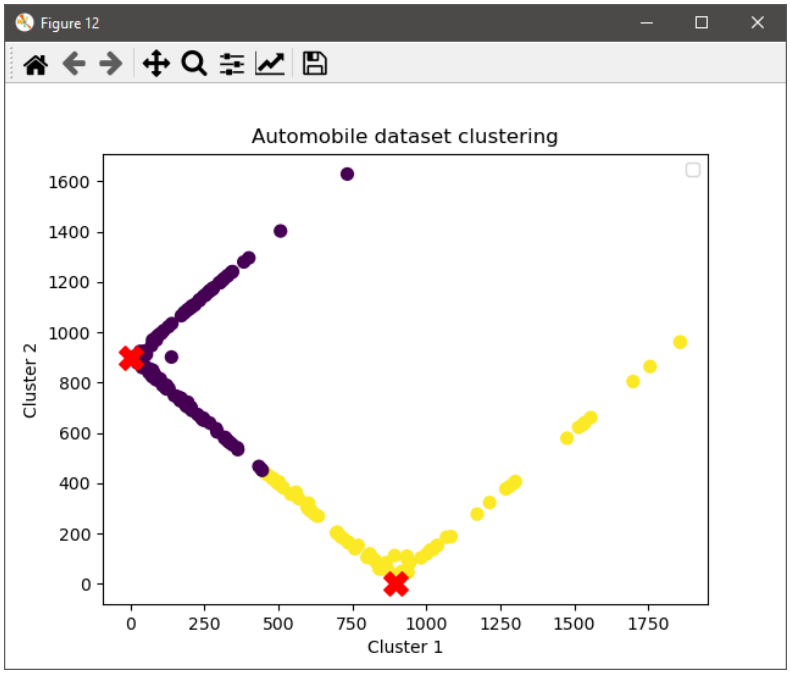
r2_score is : 0.9520775779605176 → 95,2%

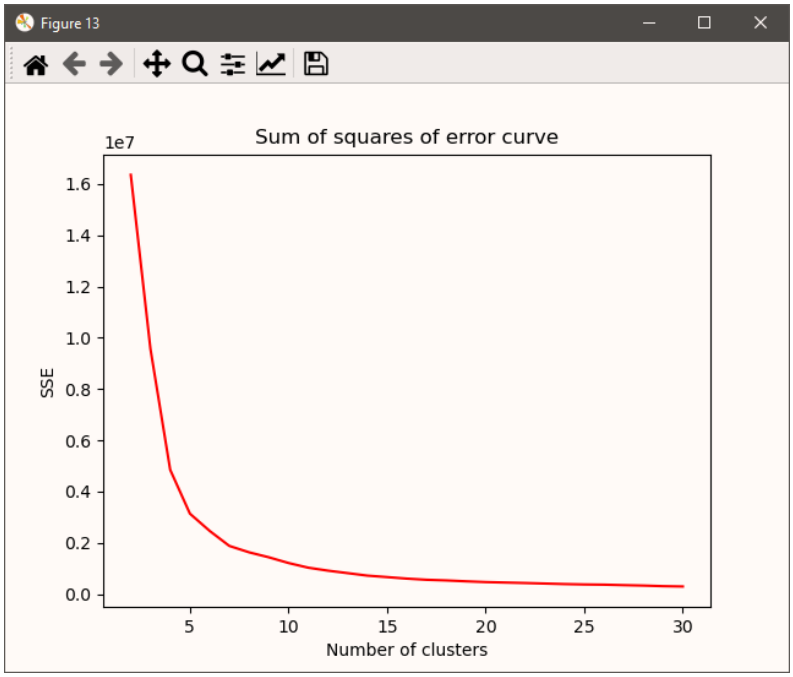
Három módszert próbáltam ki a tanításra, a LogisticRegression volt a legrosszabb, a GradientBoostingRegressor pedig a legjobb

Klaszterezés:

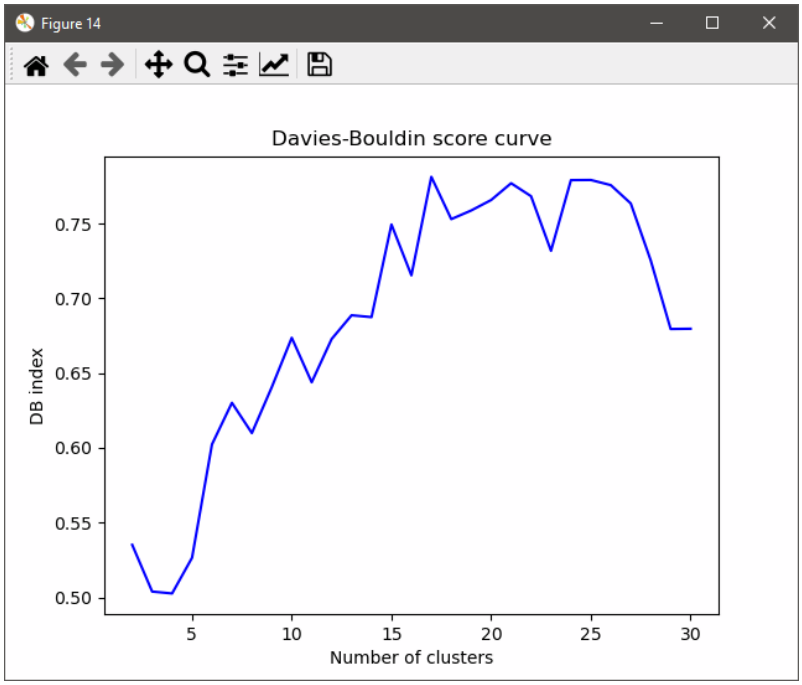


PCA kimenete, jól választja szét az adatokat.





Az SSE értékek vizualizálása



A diagramon leolvasható a lokális minimum helyén az optimális klaszter-szám, ez a mi esetünkben a 4.