**PREDICTING PULSAR STARS**

**Dankó Zsolt – DBQVH8**

A feladat az volt, hogy prediktáljuk a rendelkezésre álló adatok segítségével, hogy az adatállományban szereplő csillagok pulzárok-e (gyorsan forgó neutroncsillagok), vagy sem.

Az eredeti adatállomány az alábbi linken érhető el:  
<https://archive.ics.uci.edu/ml/machine-learning-databases/00372/HTRU2.zip>

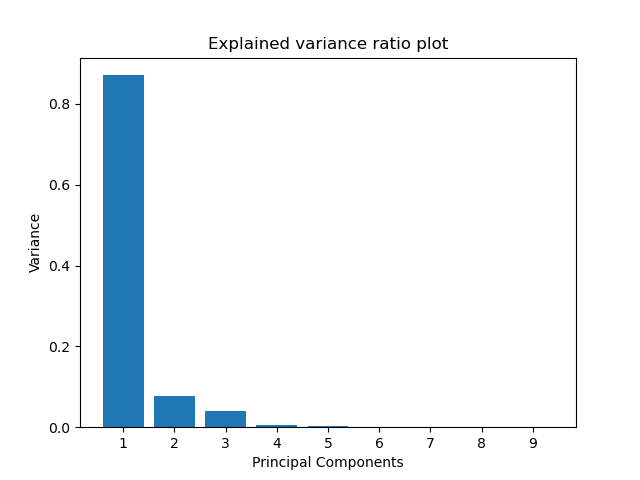
**Előkészületek**

A .csv formátumú adatállomány beolvasásához a Pandas könyvtár read\_csv metódusát használtam, majd ugyanezen könyvtár használatával feltöltöttem a NaN értékeket az adott oszlop átlagával; majd definiáltam az X és y (input és target) változókat.

**PCA**

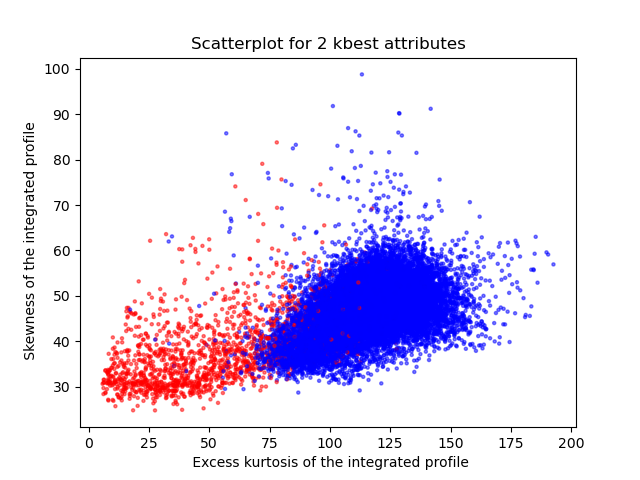
Mivel összesen 9 attribútum van az adatállományunkban, ezt le akarnánk szűkíteni, így a főkomponens analízis illesztésével megnézzük az oszlopaink varianciáját/fontosságát, majd ebből kiválasztjuk a legfontosabbakat.

Főkomponensek varianciája:



Az első ábrán észrevehető, hogy az oszlopok címkéje nem egyezik meg az oszlopaink nevével, de számunkra ez még nem is fontos, mivel a következő ábrán 2D pontdiagrammot készítünk a két legfontosabb főkomponens segítségével, és csak itt írjuk ki címkeként az oszlopok neveit.

2D pontdiagram:

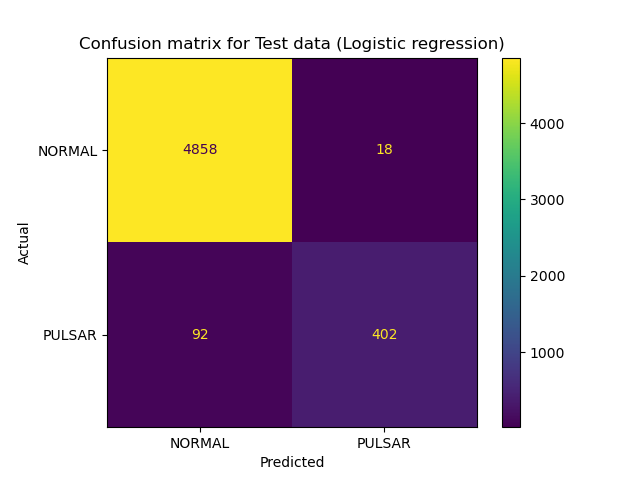


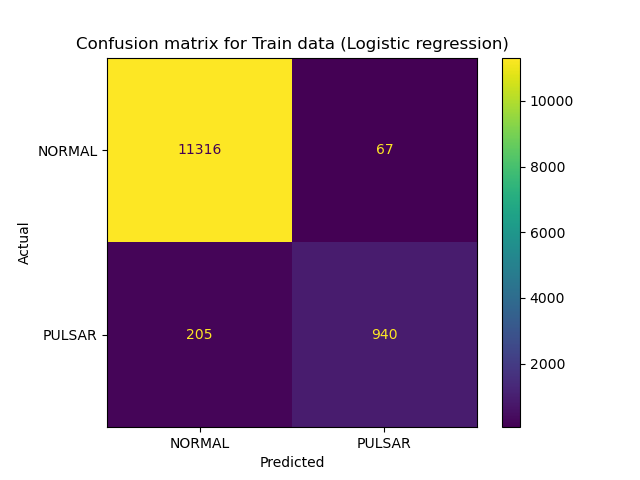
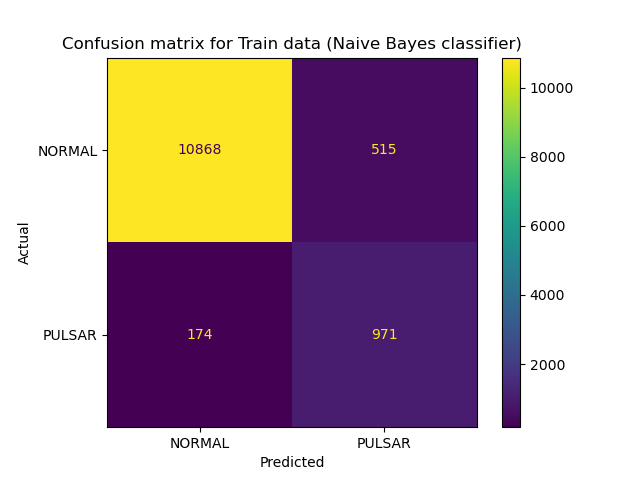
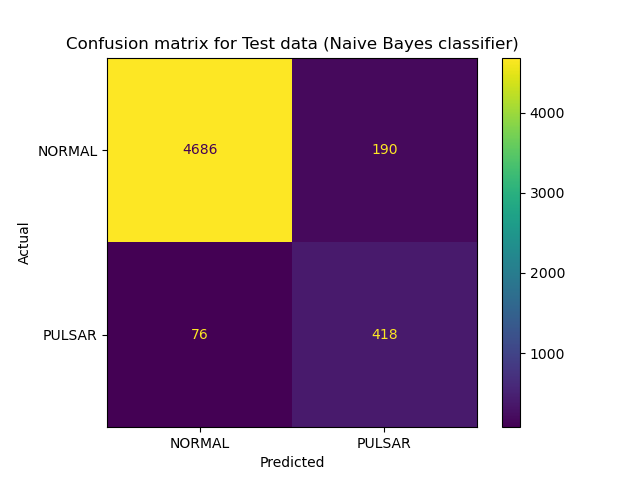
A fenti ábrán a piros pontok jelzik a pulzárokat, míg a kékek a „rendes” csillagokat, és mint látjuk, nem könnyű elhatárolni őket még a két legfontosabb főkomponens segítségével sem, ez később meg fog látszani a tanításon is. Az ábráról az olvasható, hogy minél kisebb az integrált profil ferdesége és fölösleg lapultsága, annál valószínűbb, hogy pulzárról van szó.

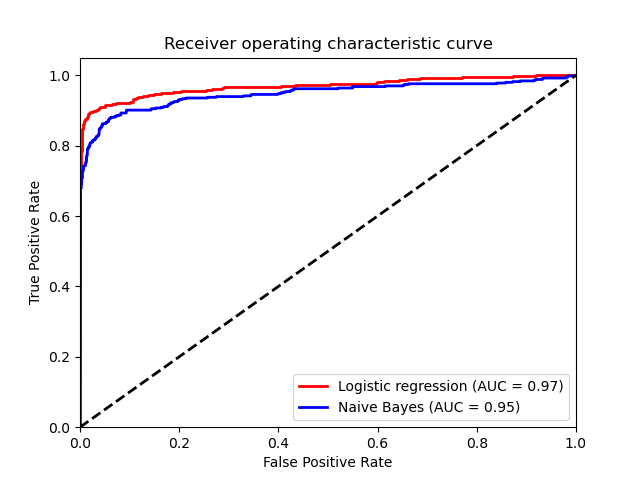
**Tanítás**

A következő teendő maga a tanítás volt, ehhez kétféle felügyelt tanítást használtam, a Logisztikus Regressziót, és a Naív Bayes módszert.

Először is felosztottam az adatokat tanító és teszt adatokra 70-30%-ban, majd illesztést végeztem, végül összehasonlítottam őket tévesztési mátrix, valamint ROC görbék segítségével.

Logisztikus Regresszió: Test Accuracy of Logistic Regression: **97.95%**

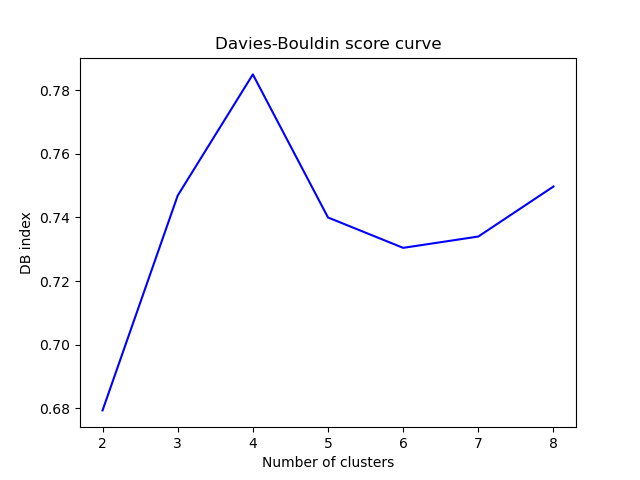
Naív Bayes: Test Accuracy of Naive Bayes: **95.05%**

ROC Görbe:

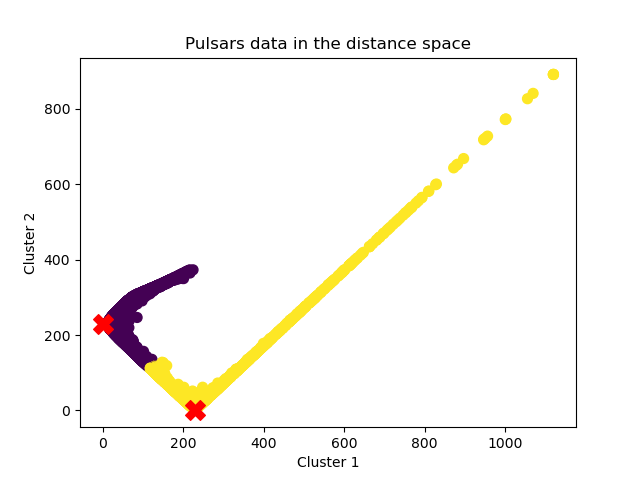
Ahogy azt a grafikonokból is láthatjuk, a Logisztikus Regresszió pontosabb eredményt produkál, mint a Naív Bayes, de így is látszik, hogy sok pulzár annyira hasonlít egy normál csillagra, hogy normál csillagként ismeri fel az algoritmus, ez látszódott a 2D pontdiagrammunkon is.

**Klaszterezés**

Mielőtt klaszterekre bontanánk az adatállományt, meg kell néznünk, mi lenne az optimális klaszterszámunk.

Erre Davies-Bouldin indexet használunk, és megnézzük egy grafikonon, hogy adott klaszterszámnál mennyi a goodness-of-fit score. Minél alacsonyabb ez az érték, annál jobb számunkra. Az alábbi ábrán láthatjuk, hogy 2-es klaszterszámnál a legkisebb, ezért 2 klaszterre fogjuk bontani az adatállományunkat.

Nem maradt más hátra, minthogy klaszterezzünk.

Illesztjük a modellt az adatokra a két klaszter-középpont meghatározásával, majd   
transzformáljuk az inputot a klaszter-középpontoktól való távolságok terébe:

**A projekthez használt könyvtárak:**

* **numpy:**
* **matplotlib.pyplot**
* **pandas**
* **sklearn**