SUNS – Zadanie 3:
Súborové učenie

Petra Kirschová

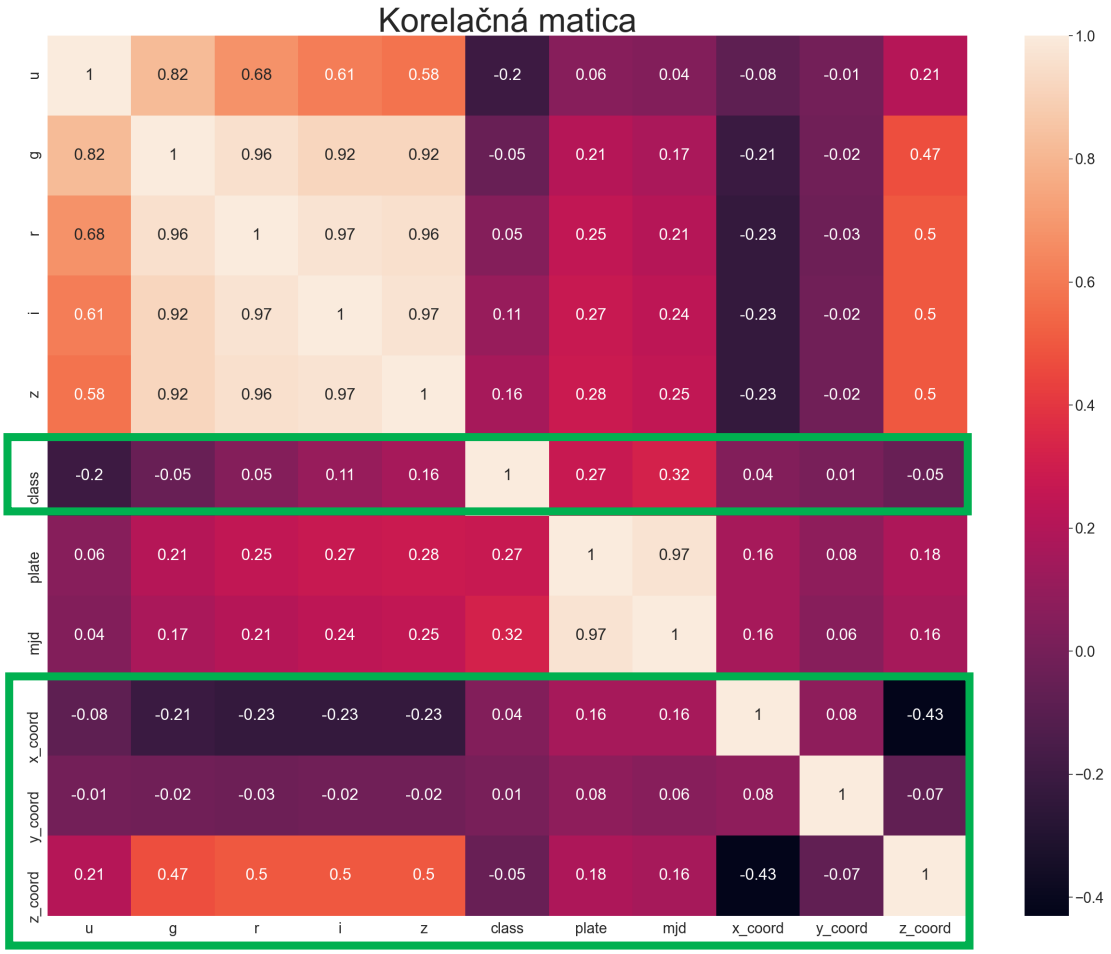
1 ANALÝZA A SPRACOVANIE DÁT

Na načítanie a prípravu dát som použila knižnicu `pandas`. Boli k dispozícii zvlášť trénovacie (11740 záznamov) a testovacie dáta (2935 záznamov).

Datasety neobsahovali žiadne null hodnoty a tiež nebolo potrebné mazať outliery, keďže strojové učenie nebolo ovplyvnené prítomnosťou outlierov.

Z datasetov som odstránila stĺpce, ktoré predstavujú identifikáciu záznamov: *objid*, *specobjid*, *run*, *rerun, camcol* a *field.*

Po odstránení stĺpcov vyzerala korelačná matica takto:



Na korelačnej matici som sledovala stĺpec *class* a súradnice *x_coord*, *y_coord* a *z_coord*.

Stĺpec *class* má pomerne nízke korelácie s ostatnými ukazovateľmi, no napriek tomu klasifikátor dosahoval veľmi dobré výsledky, keď som na trénovanie použila celý dataset (okrem *class*).

Zo súradníc *x_coord, y_coord* a *z_coord* majú *z_coord* vysoké korelácie takmer so všetkými ostatnými parametrami, hlavne s meraniami *u*, *g*, *r*, *i*, *z.*

2 KLASIFIKÁCIA

Zo súborových klasifikátorov som si vybrala random forest klasifikátor a ako ďalší som zvolila MLP klasifikátor. Na ich trénovanie som použila knižnicu `sklearn`, na vykreslenie grafov `seaborn` a `matplotlib` a na vykreslenie a exportovanie stromu `graphviz`.

2.1 RANDOM FOREST KLASIFIKÁTOR

Pri random foreste nebolo potrebné dáta normalizovať ani konvertovať na číselné reprezentácie.

Vstupné parametre trénovania sú stĺpce: *u, g, r, i, z, plate, mjd, x_coord, y_coord* a *z_coord*.

Výstupným parametrom je stĺpec *class*, ktorý obsahuje 3 rôzne triedy objektov: STAR, GALAXY a QSO.

Pre klasifikátor som nastavila:

```
classifier = RandomForestClassifier(criterion='entropy',
                                    max_leaf_nodes=40,min_samples_leaf=10)
```

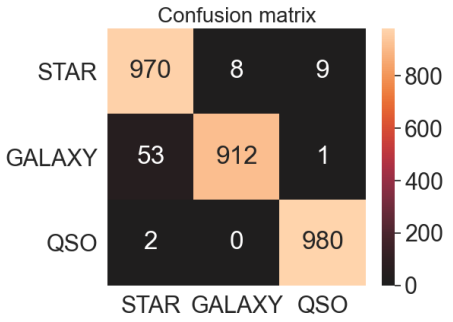
- **criterion='entropy'** – miera správnosti/úspešnosti rozdeľovania dát
- **max_leaf_nodes=40** - maximálny počet listov v stromoch
- **min_samples_leaf=10** minimálny počet vzoriek v liste

2.1.1 Výsledky

Pri trénovaní random forest klasifikátora som dosiahla veľmi dobré výsledky, ktoré sa pohybovali medzi 97-98%. Na classification report sú uvedené úspešnosti utriedenia do jednotlivých kategórií a tiež celková presnosť klasifikácie.

```
**************** RANDOM FOREST CLASSIFIER ******************
              precision    recall  f1-score   support

      GALAXY      0.946     0.983     0.964       987
         QSO      0.991     0.944     0.967       966
        STAR      0.998     0.998     0.994       982

    accuracy                          0.975      2935
   macro avg      0.976     0.975     0.975      2935
weighted avg      0.976     0.975     0.975      2935
```

Na confusion matrix je tiež vidno, že trénovanie prebehlo veľmi úspešne. Na diagonále sa nachádzajú vzorky, ktoré boli správne zatriedené do jednotlivých kategórií. Nesprávne zatriedených vzoriek je v porovnaní s veľkosťou datasetu zanedbateľné množstvo.



2.2 SLABÝ KLASIFIKÁTOR

Slabým klasifikátorom v random forest modeli je 1 strom. V modeli som nechala defaultný počet slabých klasifikátorov – 100, t.j. celkovo sa vyhodnocuje 100 rozhodovacích stromov.

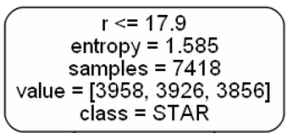
Jeden uzol stromu obsahuje:

- *Rozhodovaciu podmienku*
- *Entropy* – miera rozdielnosti dát
- *Samples* - počet vzoriek
- *Value* – koľko vzoriek prislúcha jednotlivým triedam
- *Class* – výsledok zatriedenia v danom uzli stromu

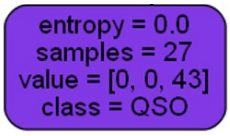
Strom na základe rozhodovacej podmienky rozdelí dáta na 2 časti – vľavo pôjdu tie, pre ktoré podmienka platí a vpravo ostatné. Tento krok sa opakuje, kým nie sú splnené zastavovacie podmienky trénovania.

Entropia vyjadruje, ako veľmi sú dáta v danej vzorke navzájom rozdielne. V roote stromu je entropia vysoká, pretože všetky triedy sú takmer rovnako zastúpené. Klesaním nižšie v strome sa entropia znižuje a v niektorých listoch dosiahla aj nulu.

Root stromu



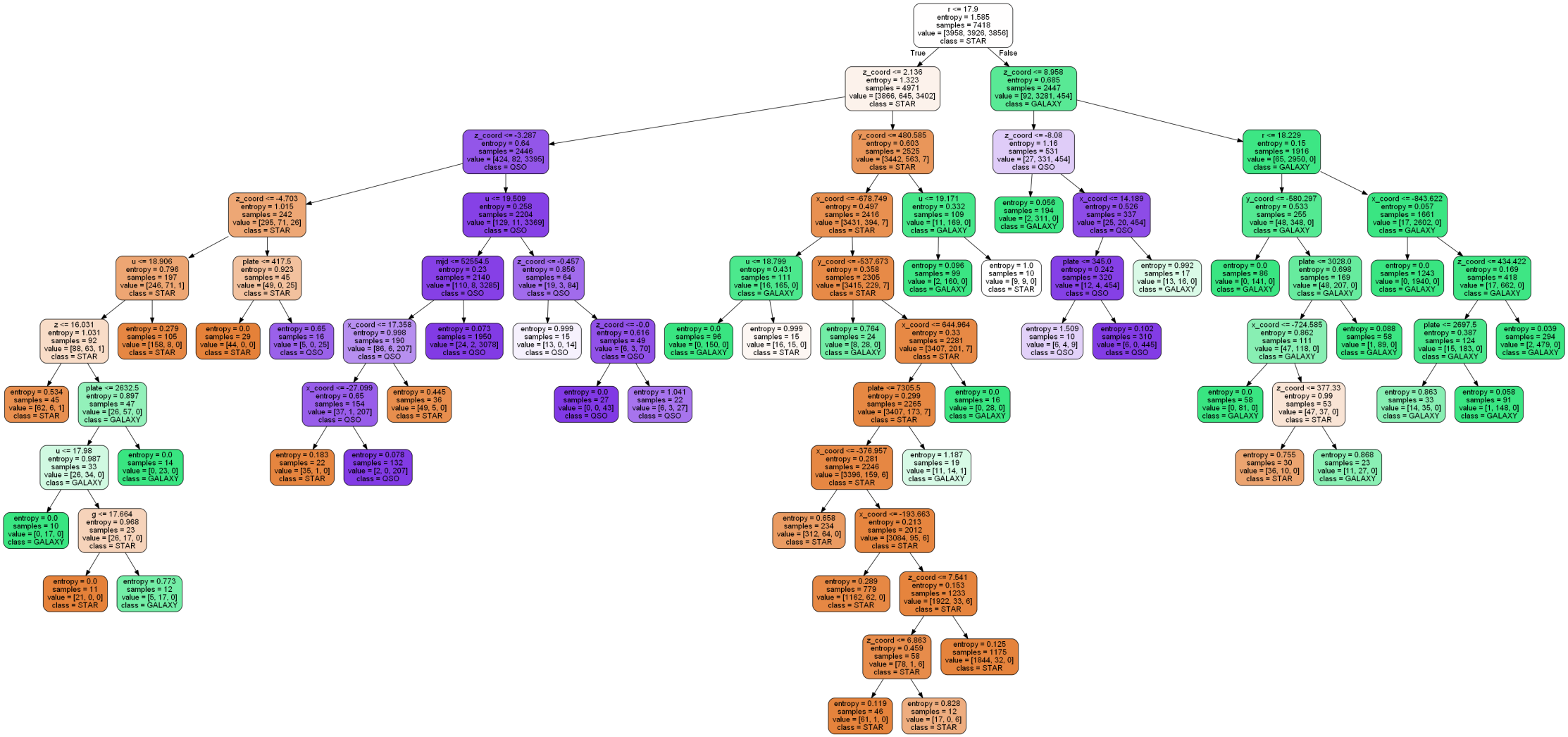
List stromu s nulovou entropiou



Výslednou triedou v danom uzle stromu je tá trieda, ktorá má vo vzorke najväčšie zastúpenie (najväčšie číslo v poli *value*).

Vykreslený strom je výsledkom pruningu, kvôli tomu, že nenastavenie zastavovacej podmienky viedlo k pretrénovaniu modelu. Pôvodný strom bol tiež príliš veľký a koncové uzly obsahovali málo dát na to, aby boli smerodajné. Pri veľkom strome vo väčšine vetiev nastala situácia, kedy sa ďalším delením stromu nezískala žiadna výpovedná hodnota, keďže sa vzorky v oboch častiach datasetu zaradili do rovnakej triedy.

Po pruningu sa v strome nachádza tiež zopár takých vetiev, ktorých rozdelením dostaneme rovnaký výsledok, no vzhľadom na dosiahnutú úspešnosť je tento model relatívne dobre rozložený. Sprísňovanie zastavovacích podmienok sa potom odzrkadľovalo na celkovej úspešnosti klasifikátora.


r <= 17.9
entropy = 1.585
samples = 7418
value = [3958, 3926, 3856]
class = STAR
True
False
z_coord <= 2.136
entropy = 1.323
samples = 4971
value = [3866, 645, 3402]
class = STAR
z_coord <= 8.958
entropy = 0.685
samples = 2447
value = [92, 3281, 454]
class = GALAXY
z_coord <= -3.287
entropy = 0.64
samples = 2446
value = [424, 82, 3395]
class = QSO
y_coord <= 480.585
entropy = 0.603
samples = 2525
value = [3442, 563, 7]
class = STAR
z_coord <= -8.08
entropy = 1.16
samples = 531
value = [27, 331, 454]
class = QSO
r <= 18.229
entropy = 0.15
samples = 1916
value = [65, 2950, 0]
class = GALAXY
z_coord <= -4.703
entropy = 1.015
samples = 242
value = [295, 71, 26]
class = STAR
u <= 19.509
entropy = 0.258
samples = 2204
value = [129, 11, 3369]
class = QSO
x_coord <= -678.749
entropy = 0.497
samples = 2416
value = [3431, 394, 7]
class = STAR
u <= 19.171
entropy = 0.332
samples = 109
value = [11, 169, 0]
class = GALAXY
entropy = 0.056
samples = 194
value = [2, 311, 0]
class = GALAXY
x_coord <= 14.189
entropy = 0.526
samples = 337
value = [25, 20, 454]
class = QSO
y_coord <= -580.297
entropy = 0.533
samples = 255
value = [48, 348, 0]
class = GALAXY
x_coord <= -843.622
entropy = 0.057
samples = 1661
value = [17, 2602, 0]
class = GALAXY
u <= 18.906
entropy = 0.796
samples = 197
value = [246, 71, 1]
class = STAR
plate <= 417.5
entropy = 0.923
samples = 45
value = [49, 0, 25]
class = STAR
mjd <= 52554.5
entropy = 0.23
samples = 2140
value = [110, 8, 3285]
class = QSO
z_coord <= -0.457
entropy = 0.856
samples = 64
value = [19, 3, 84]
class = QSO
u <= 18.799
entropy = 0.431
samples = 111
value = [16, 165, 0]
class = GALAXY
y_coord <= -537.673
entropy = 0.358
samples = 2305
value = [3415, 229, 7]
class = STAR
entropy = 0.096
samples = 99
value = [2, 160, 0]
class = GALAXY
entropy = 1.0
samples = 10
value = [9, 9, 0]
class = STAR
plate <= 345.0
entropy = 0.242
samples = 320
value = [12, 4, 454]
class = QSO
entropy = 0.992
samples = 17
value = [13, 16, 0]
class = GALAXY
entropy = 0.0
samples = 86
value = [0, 141, 0]
class = GALAXY
plate <= 3028.0
entropy = 0.698
samples = 169
value = [48, 207, 0]
class = GALAXY
entropy = 0.0
samples = 1243
value = [0, 1940, 0]
class = GALAXY
z_coord <= 434.422
entropy = 0.169
samples = 418
value = [17, 662, 0]
class = GALAXY
z <= 16.031
entropy = 1.031
samples = 92
value = [88, 63, 1]
class = STAR
entropy = 0.279
samples = 105
value = [158, 8, 0]
class = STAR
entropy = 0.0
samples = 29
value = [44, 0, 0]
class = STAR
entropy = 0.65
samples = 16
value = [5, 0, 25]
class = QSO
x_coord <= 17.358
entropy = 0.998
samples = 190
value = [86, 6, 207]
class = QSO
entropy = 0.073
samples = 1950
value = [24, 2, 3078]
class = QSO
entropy = 0.999
samples = 15
value = [13, 0, 14]
class = QSO
z_coord <= -0.0
entropy = 0.616
samples = 49
value = [6, 3, 70]
class = QSO
entropy = 0.0
samples = 96
value = [0, 150, 0]
class = GALAXY
entropy = 0.999
samples = 15
value = [16, 15, 0]
class = STAR
entropy = 0.764
samples = 24
value = [8, 28, 0]
class = GALAXY
x_coord <= 644.964
entropy = 0.33
samples = 2281
value = [3407, 201, 7]
class = STAR
entropy = 1.509
samples = 10
value = [6, 4, 9]
class = QSO
entropy = 0.102
samples = 310
value = [6, 0, 445]
class = QSO
x_coord <= -724.585
entropy = 0.862
samples = 111
value = [47, 118, 0]
class = GALAXY
entropy = 0.088
samples = 58
value = [1, 89, 0]
class = GALAXY
plate <= 2697.5
entropy = 0.387
samples = 124
value = [15, 183, 0]
class = GALAXY
entropy = 0.039
samples = 294
value = [2, 479, 0]
class = GALAXY
entropy = 0.534
samples = 45
value = [62, 6, 1]
class = STAR
plate <= 2632.5
entropy = 0.897
samples = 47
value = [26, 57, 0]
class = GALAXY
x_coord <= -27.099
entropy = 0.65
samples = 154
value = [37, 1, 207]
class = QSO
entropy = 0.445
samples = 36
value = [49, 5, 0]
class = STAR
entropy = 0.0
samples = 27
value = [0, 0, 43]
class = QSO
entropy = 1.041
samples = 22
value = [6, 3, 27]
class = QSO
plate <= 7305.5
entropy = 0.299
samples = 2265
value = [3407, 173, 7]
class = STAR
entropy = 0.0
samples = 16
value = [0, 28, 0]
class = GALAXY
entropy = 0.0
samples = 58
value = [0, 81, 0]
class = GALAXY
z_coord <= 377.33
entropy = 0.99
samples = 53
value = [47, 37, 0]
class = STAR
entropy = 0.863
samples = 33
value = [14, 35, 0]
class = GALAXY
entropy = 0.058
samples = 91
value = [1, 148, 0]
class = GALAXY
u <= 17.98
entropy = 0.987
samples = 33
value = [26, 34, 0]
class = GALAXY
entropy = 0.0
samples = 14
value = [0, 23, 0]
class = GALAXY
entropy = 0.183
samples = 22
value = [35, 1, 0]
class = STAR
entropy = 0.078
samples = 132
value = [2, 0, 207]
class = QSO
x_coord <= -376.957
entropy = 0.281
samples = 2246
value = [3396, 159, 6]
class = STAR
entropy = 1.187
samples = 19
value = [11, 14, 1]
class = GALAXY
entropy = 0.755
samples = 30
value = [36, 10, 0]
class = STAR
entropy = 0.868
samples = 23
value = [11, 27, 0]
class = GALAXY
entropy = 0.0
samples = 10
value = [0, 17, 0]
class = GALAXY
g <= 17.664
entropy = 0.968
samples = 23
value = [26, 17, 0]
class = STAR
entropy = 0.658
samples = 234
value = [312, 64, 0]
class = STAR
x_coord <= -193.663
entropy = 0.213
samples = 2012
value = [3084, 95, 6]
class = STAR
entropy = 0.0
samples = 11
value = [21, 0, 0]
class = STAR
entropy = 0.773
samples = 12
value = [5, 17, 0]
class = GALAXY
entropy = 0.289
samples = 779
value = [1162, 62, 0]
class = STAR
z_coord <= 7.541
entropy = 0.153
samples = 1233
value = [1922, 33, 6]
class = STAR
z_coord <= 6.863
entropy = 0.459
samples = 58
value = [78, 1, 6]
class = STAR
entropy = 0.125
samples = 1175
value = [1844, 32, 0]
class = STAR
entropy = 0.119
samples = 46
value = [61, 1, 0]
class = STAR
entropy = 0.828
samples = 12
value = [17, 0, 6]
class = STAR

2.3 MLP KLASIFIKÁTOR

Vstup a výstup MLP klasifikátora bol rovnaký, ako pri random foreste, vstupné dáta sa na rozdiel od random forestu museli normalizovať. Výstupnú množinu, t.j. stĺpec *class* som nechala bez úprav v pôvodnom tvare – trénovanie fungovalo, aj keď som vo výstupnej množine nechala triedy vo formáte string.

Nastavenie klasifikátora:

```
classifier = MLPClassifier(max_iter=600, alpha=0.01,
                           hidden_layer_sizes=(20,))
```

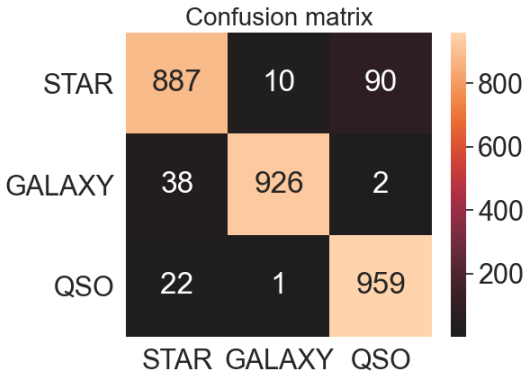
- aktivačná funkcia=**relu** (default)
- solver=**adam** (default)
- **max_iter=600** - maximálny počet iterácii
- **alpha= 0.01** – regularizačná penalta
- **hidden_layer_sizes=(20,)** - 1 skrytá vrstva s 20 neurónmi

2.3.1 Výsledky

MLP klasifikátor dosiahol trochu nižšie úspešnosti, ako random forest, ale stále sú veľmi vysoké. Celkovo sa presnosti MLP klasifikátora pohybovali okolo 93-94%.

********** MLP CLASSIFIER **********

	precision	recall	f1-score	support
GALAXY	0.937	0.899	0.917	987
QSO	0.988	0.959	0.973	966
STAR	0.912	0.977	0.943	982
accuracy			0.944	2935
macro avg	0.946	0.945	0.945	2935
weighted avg	0.946	0.944	0.944	2935



3 REGRESIA

Spomedzi súborových klasifikátorov som na regresiu použila random forest regresor a ako ďalší regresor som si vybrala k-nearest neighbors regresor. Regresory som implementovala pomocou knižnice `sklearn` a na vykreslenie grafov som použila `seaborn`, `matplotlib` a `plotly`.

3.1 RANDOM FOREST REGRESOR

Pri random forest regresore nebolo potrebné dáta normalizovať, ale vstupný stĺpec class bolo treba konvertovať na číselné reprezentácie tried.

Vstupné parametre trénovania sú stĺpce: *u, g, r, i, z, plate, mjd* a *class*.

Výstupnými parametrami sú stĺpce so súradnicami *x_coord, y_coord* a *z_coord*, ktoré určujú polohu sledovaného objektu vo vesmíre.

Nastavenie regresora:

```
regressor = RandomForestRegressor(criterion="mse", max_depth=15)
```

- **criterion="mse"** – miera správnosti rozdelenia dát = mean squared error
- **max_depth=15** – maximálna hĺbka stromu, zastavovacia podmienka

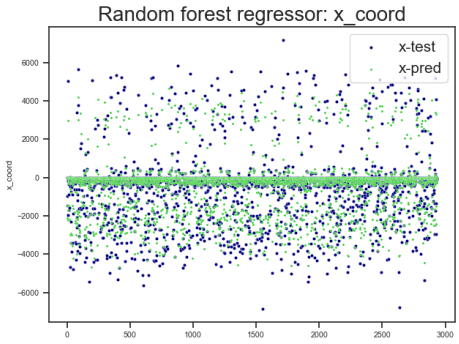
3.1.1 Výsledky

Random forest regresor dosahoval dosť dobré výsledky, dosiahol R2 score pre celý výstup okolo 73% a mean absolute error cca 280. Pre jednotlivé súradnice bola pre *x_coord* najvyššia chyba a pre *z_coord* najnižšia. R2 score bolo najvyššie pre *x_coord* a najnižšie pre *y_coord*.

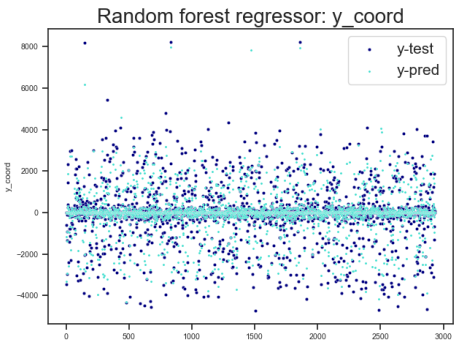
```
**********  RANDOM FOREST REGRESSOR  **********
[x_coord,y_coord,z_coord]: MAE = 280.0764288871531 R2 = 0.7265198924894873

x_coord: MAE= 324.1265951496431 R2= 0.7641190247821393
y_coord: MAE= 291.8999896621916 R2= 0.6537591774454663
z_coord: MAE= 224.20270184962345 R2= 0.7616814752408566
```

V 2D a 3D grafoch som porovnala výstup testovacích a predikovaných dát. Na grafoch je vidno, že regresor priblížil výstupné hodnoty relatívne dobre, pozície predpovedaných bodov sú veľmi blízko k reálnym bodom. Na 2D grafoch sú porovnané predpovedané a očakávané výstupy (v závislosti od id alebo pri residual plote priamo závislosť medzi predikovaným a očakávaným výstupom) pre každú súradnicu zvlášť. 3D graf obsahuje body *[x_coord, y_coord, z_coord]* v trojrozmernom priestore.

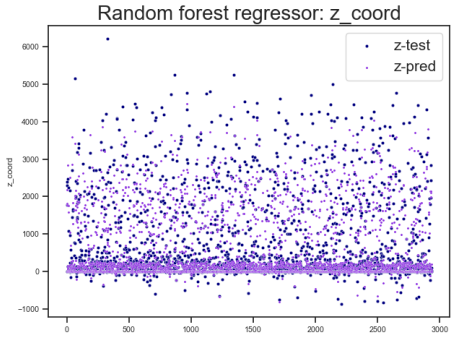
x_coord: porovnanie očakávaných hodnôt s predpovedanými + residual plot

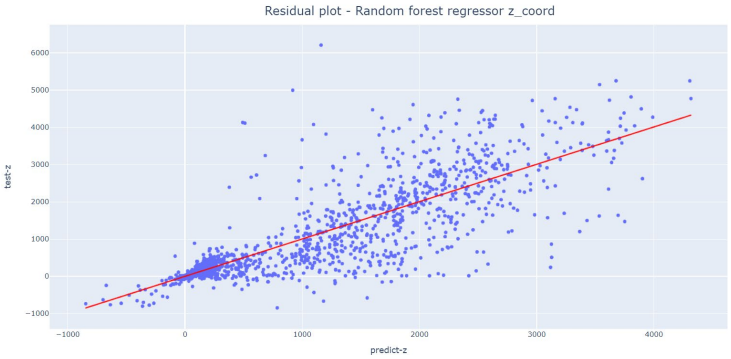


y_coord: porovnanie očakávaných hodnôt s predpovedanými + residual plot

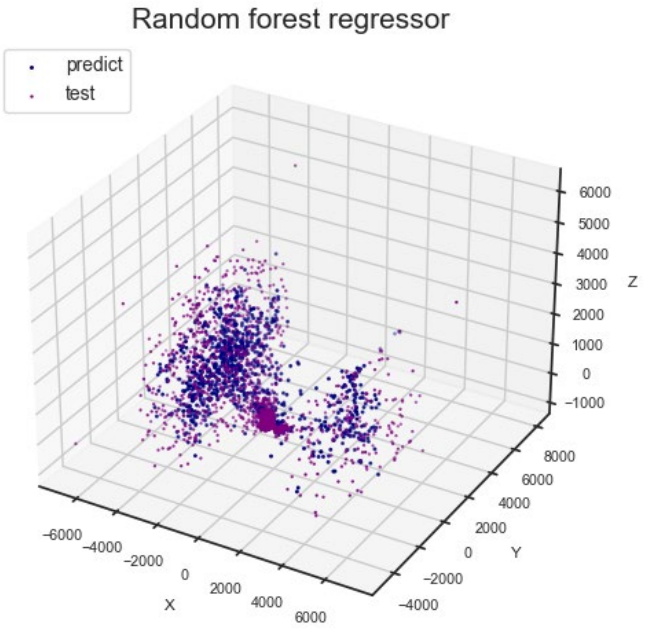


z_coord: porovnanie očakávaných hodnôt s predpovedanými + residual plot





[x_coord, y_coord, z_coord]: porovnanie očakávaných hodnôt s predpovedanými



3.2 K-NEAREST NEIGHBORS REGRESOR

K-nearest neighbors regresor predpovedá výsledné hodnoty na základe toho, aké hodnoty dosahujú iné dáta v jeho okolí (neighbors). Dáta tiež nebolo treba normalizovať, ale vstupný stĺpec *class* sa konvertoval na číselné reprezentácie tried.

Nastavenie regresora:

```
regressor = KNeighborsRegressor(n_neighbors=8, weights='distance')
```

- **n_neighbors=8** – počet „susedov“, podľa ktorých sa vyhodnocuje výsledok
- **weights='distance'** – váhy susedov sa určujú podľa ich vzdialenosti k sledovanému bodu, t.j. body, ktoré sú bližšie viac ovplyvnia výsledok, ako tie, ktoré sú ďalej.

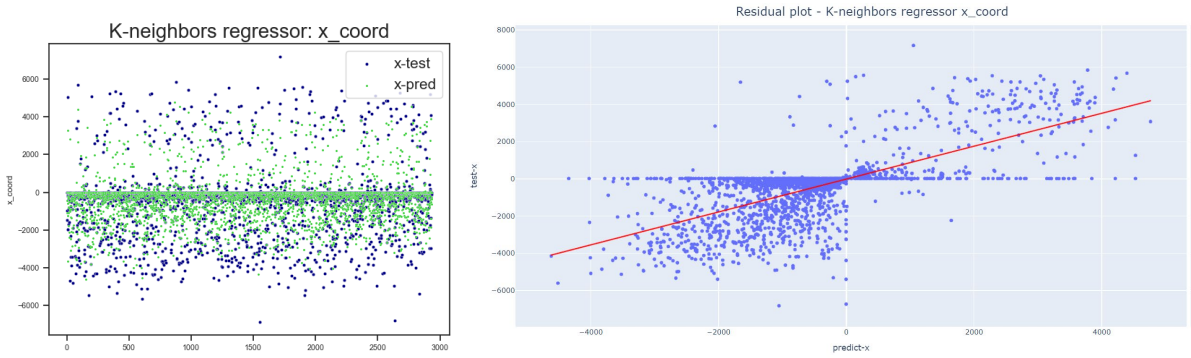
3.2.1 Výsledky

Výsledky KNN regresora boli horšie, ako pri random forest, R2 bolo celkovo okolo 0.5, MAE cca 490, pre jednotlivé súradnice sa chyby pohybovali približne medzi 400-700, R2 score medzi 0.4-0.6.

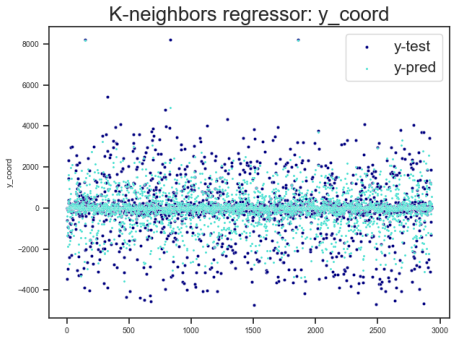
```
**********  K-NEIGHBORS REGRESSOR  **********
[x_coord,y_coord,z_coord]: MAE = 491.929882806797 R2 = 0.5040752458655859

x_coord: MAE= 656.3278901376939 R2= 0.4997842644419148
y_coord: MAE= 414.38610080972256 R2= 0.5721132709708907
z_coord: MAE= 405.07565747297576 R2= 0.4403282021839513
```

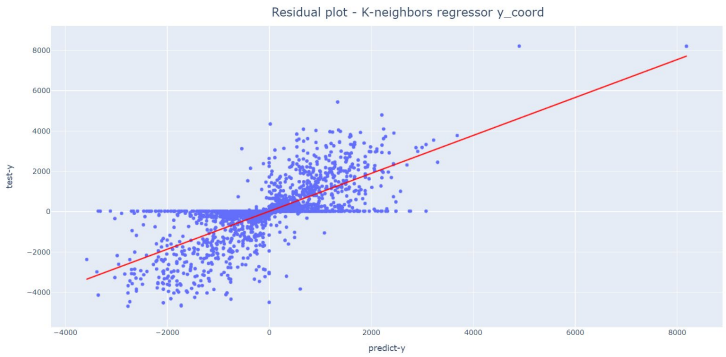
Grafy, na ktorých sú vykreslené pôvodné a predpovedané výstupy sú na prvý pohľad veľmi podobné s grafmi random forest regresora, rozdiely sú pozorovateľné na residual grafoch. Pri tomto regresore sú predpovedané body rozptýlené viac, ako pri random foreste a najviac sa ich zoskupuje pozdĺž nuly.

x_coord: porovnanie očakávaných hodnôt s predpovedanými + residual plot

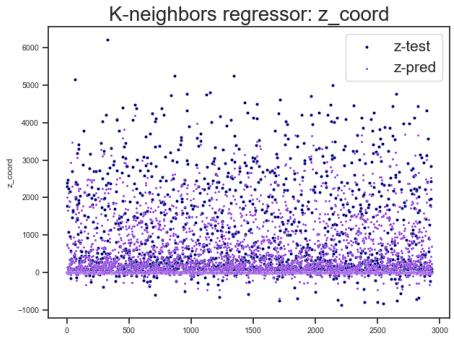


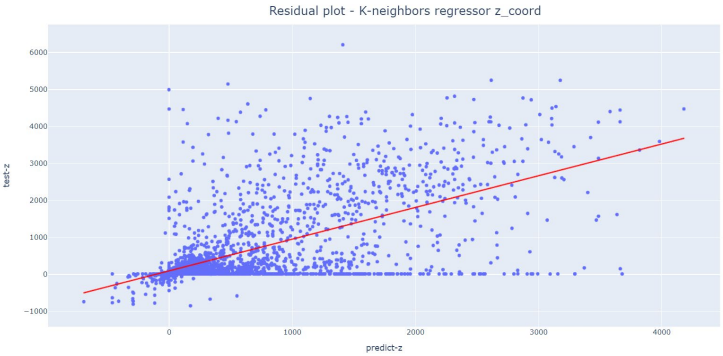
y_coord: porovnanie očakávaných hodnôt s predpovedanými + residual plot





z_coord: porovnanie očakávaných hodnôt s predpovedanými + residual plot





[x_coord, y_coord, z_coord]: porovnanie očakávaných hodnôt s predpovedanými