

Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00

Štefan Dudaško,  
Fakulta Informatiky a informačných technológií

Umelá inteligencia - Zadanie 4A - klasifikácia

## Zadanie

Máme 2D priestor, ktorý má rozmery  $X$  a  $Y$ , v intervaloch od  $-5000$  do  $+5000$ . V tomto priestore sa môžu nachádzať body, pričom každý bod má určenú polohu pomocou súradníc  $X$  a  $Y$ . Každý bod má unikátne súradnice (t.j. nemalo by byť viac bodov na presne tom istom mieste). Každý bod patrí do jednej zo 4 tried, pričom tieto triedy sú: red (R), green (G), blue (B) a purple (P). Na začiatku sa v priestore nachádza 5 bodov pre každú triedu (dokopy teda 20 bodov). Súradnice počiatočných bodov sú:

R:  $[-4500, -4400]$ ,  $[-4100, -3000]$ ,  $[-1800, -2400]$ ,  $[-2500, -3400]$  a  $[-2000, -1400]$

G:  $[+4500, -4400]$ ,  $[+4100, -3000]$ ,  $[+1800, -2400]$ ,  $[+2500, -3400]$  a  $[+2000, -1400]$

B:  $[-4500, +4400]$ ,  $[-4100, +3000]$ ,  $[-1800, +2400]$ ,  $[-2500, +3400]$  a  $[-2000, +1400]$

P:  $[+4500, +4400]$ ,  $[+4100, +3000]$ ,  $[+1800, +2400]$ ,  $[+2500, +3400]$  a  $[+2000, +1400]$

Vašou úlohou je naprogramovať klasifikátor pre nové body – v podobe funkcie `classify(int X, int Y, int k)`, ktorá klasifikuje nový bod so súradnicami  $X$  a  $Y$ , pridá tento bod do nášho 2D priestoru a vráti triedu, ktorú pridelila pre tento bod. Na klasifikáciu použite  $k$ -NN algoritmus, pričom  $k$  môže byť 1, 3, 7 alebo 15.

Na demonštráciu Vášho klasifikátora vytvorte testovacie prostredie, v rámci ktorého budete postupne generovať nové body a klasifikovať ich (volaním funkcie `classify`). Celkovo vygenerujte 40000 nových bodov (10000 z každej triedy). Súradnice nových bodov generujte náhodne, pričom nový bod by mal mať zakaždým inú triedu (dva body vygenerované po sebe by nemali byť rovnakej triedy):

- R body by mali byť generované s 99% pravdepodobnosťou s  $X < +500$  a  $Y < +500$
- G body by mali byť generované s 99% pravdepodobnosťou s  $X > -500$  a  $Y < +500$
- B body by mali byť generované s 99% pravdepodobnosťou s  $X < +500$  a  $Y > -500$
- P body by mali byť generované s 99% pravdepodobnosťou s  $X > -500$  a  $Y > -500$

Návratovú hodnotu funkcie `classify` porovnávajte s triedou vygenerovaného bodu. Na základe týchto porovnaní vyhodnoťte úspešnosť Vášho klasifikátora pre daný experiment.

Experiment vykonajte 4-krát, pričom zakaždým Váš klasifikátor použije iný parameter  $k$  (pre  $k = 1, 3, 7$  alebo 15) a vygenerované body budú pre každý experiment rovnaké.

Vizualizácia: pre každý z týchto experimentov vykreslite výslednú 2D plochu tak, že vyfarbíte túto plochu celú. Prázdne miesta v 2D ploche vyfarbíte podľa Vášho klasifikátora.

V závere zhodnoťte dosiahnuté výsledky ich porovnaním.

## Úvod

V tomto zadaní riešime program, ktorý dokáže pridávať body do tried podľa určitých pravidiel – v našom prípade len postupné pridávanie striedavo do každej triedy po jednom, a klasifikovať body podľa knn algoritmu. Popis úlohy je spomenutý už vyššie, a preto prejdeme hneď na technickú špecifikáciu a ďalej už na naše riešenie.

## Technická špecifikácia

Problém sme riešili, kvôli jednoduchosti implementácie vizualizácie v Pythone, vizualizáciu sme riešili cez knižnicu matplotlib. Programovali sme objektovo. Časy spracovania sme nemerali, keďže neboli predmetom tohoto zadania a technické prostriedky, použité na tento problém boli nižšie.

- Pamäť - 8 GB 1600 MHz DDR3
- CPU - 1,3 GHz Intel Core i5

## Popis algoritmu a fungovanie programu

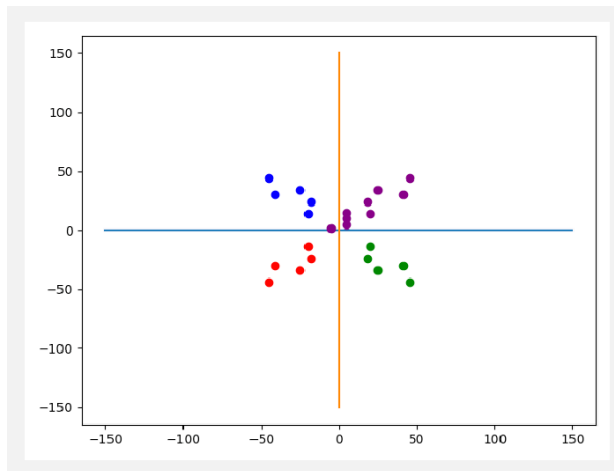
Program je veľmi jednoduchý, a preto ho popíšeme takto za sebou, ako sa vykonáva:

1. Zavoláme funkciu na vytvorenie bodov, táto funkcia spraví (pre jedno k) – prečistenie mapky, kvôli predošlej iterácii, vytvorí si svoj vlastný klasifikátor, aby sme nepoužívali (mutable) predošlý a nemali tam už nasetované nejaké dáta
2. Ten si v konštruktori nastaví 20 základných trénovacích bodov a pridá ich hneď aj na mapku
3. Postupne klasifikujeme po jednom všetky body, v poradí RGBP
4. Klasifikuje sa tak, že (uvedieme napríklad pre červený bod, obdobne potom aj pre všetky ostatné) vyberieme bod zo zadaného rozmedzia, pri 1% šanci ho hodíme náhodne hocikde na mapku a zisťujeme jeho klasifikáciu, pozeráme sa do množiny všetkých bodov, ktoré sú práve na mapke, vyberieme k najbližších, a na základe toho mu priradíme farbu. Túto vzdialenosť počítame pytagorovou vetou. V tejto funkcii máme aj možnosť použiť len už pridané body, aby sme potom vedeli ešte vyplniť celú mapku, avšak, po pridaní 40k bodov z pôvodného zadania, to vôbec nie je treba, preto to nepoužívame, môžeme to samozrejme použiť, len je to veľmi náročná hardvérová operácia. Potom vyhodnotíme, či náš klasifikátor klasifikoval bod rovnako, to potom ukážeme na konci pri štatistikách.
5. Na záver, po klasifikovaní všetkých bodov vidíme vykreslenú mapku a štatistiky
6. Úpravu bodov, a zopár základných vlastností vieme modifikovať v súbore [config.py](#)

Program sa skladá len zo súborov main, kde sa robia základné volania a classifier, ktorý je samodeskriptívny a nie je ho treba zrejme ďalej veľmi popisovať. Pre trochu komplexnejšie alebo otázne časti algoritmu, sú tam pridané komentáre.

## Testovanie

Pre tento program sme nerobili žiadne zložitejšie integračné testy, vychádzali sme len z vizuálnej kontroly pridaných bodov, či algoritmus funguje tak, ako očakávame. Uvedieme preto v tejto sekcii len jeden dôležitejší test, ktorý sme vykonali a ďalej už len výsledky pokusov.



*Veľmi jednoduchý, ale podstatný test – bod – klasifikovanie pridania modrého bodu ako fialového, pretože fialové body sú jeho najbližší susedia.*

## Vizualizácie a výsledky pokusov

Pre začiatok je ešte dobré uviesť, že sme najprv vytvorili algoritmus, ktorý úspešne zbehol aj na 40 020 bodoch pre všetky k, ale nakoľko spustenie simulácie na pozadí nie je zložité, uviedli sme experiment aj pre viac pokusov, inicializačné atribúty programu sú ľahko upraviteľné v configu.

Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00

Test 1:

Úspešnosť klasifikácie:

**Body:** 2520

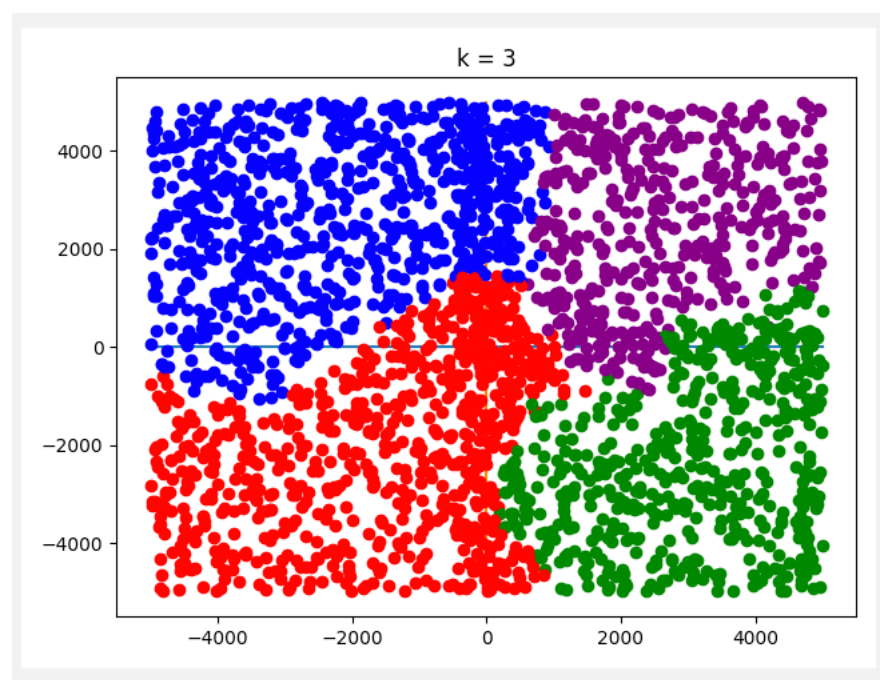
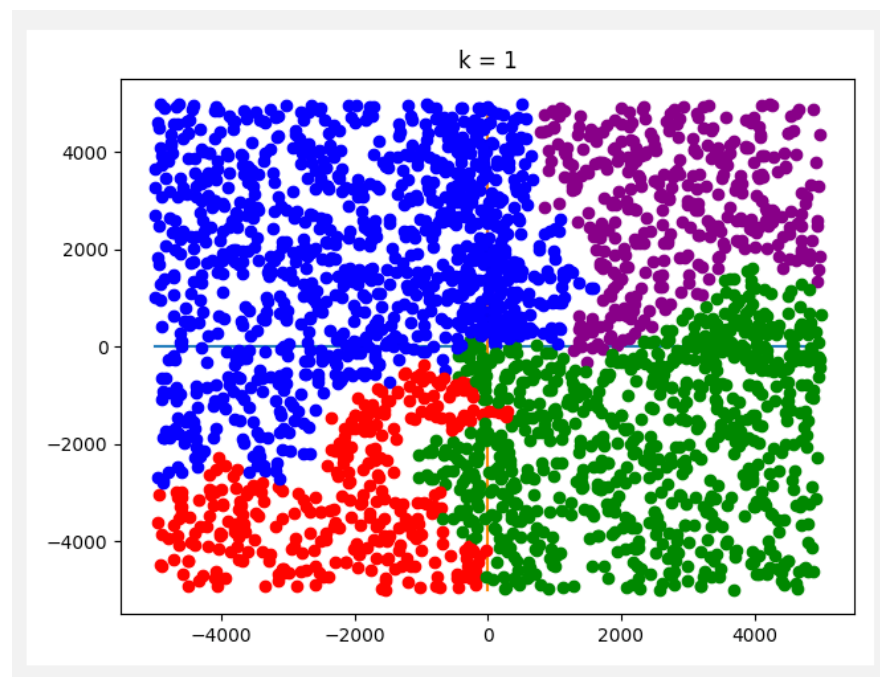
K = 1: 74.03999%

K = 3: 78.60001%

K = 7: 74.48000%

K = 15: 59.19999%

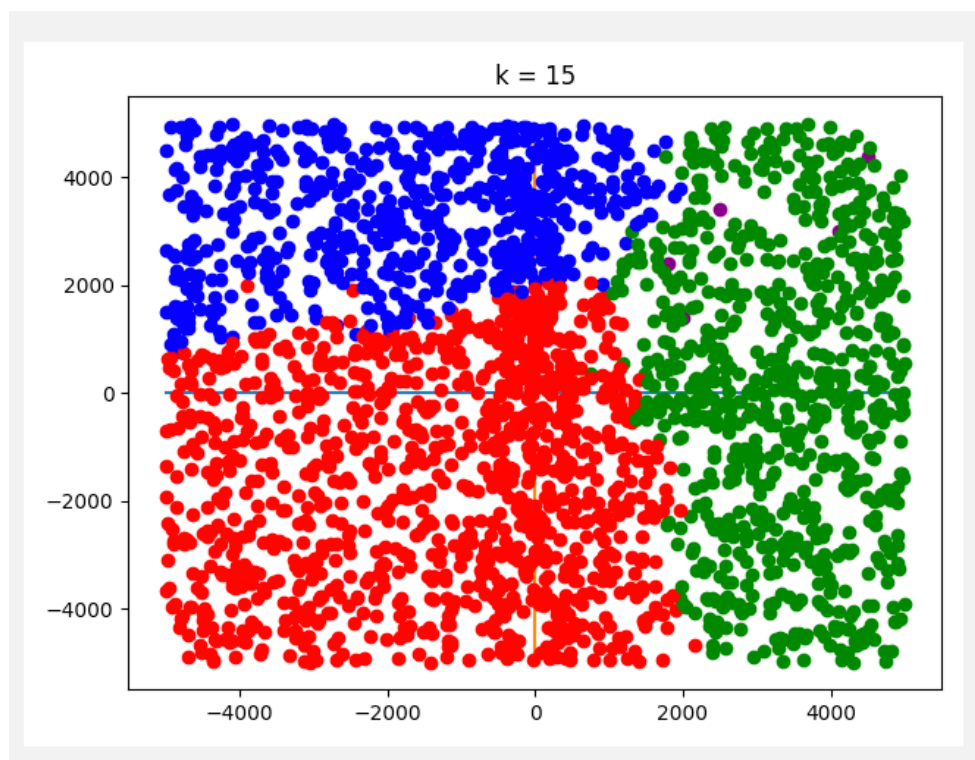
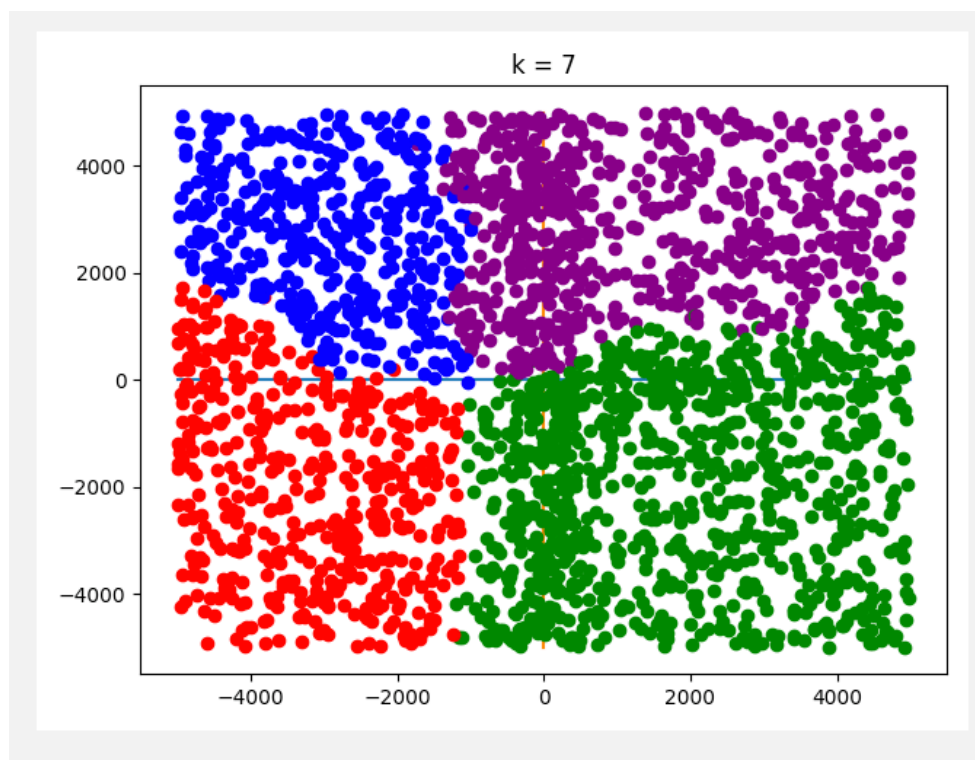
Vizualizácia:



Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00



Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00

## Test 2.

Úspešnosť klasifikácie:

**Body:** 5020

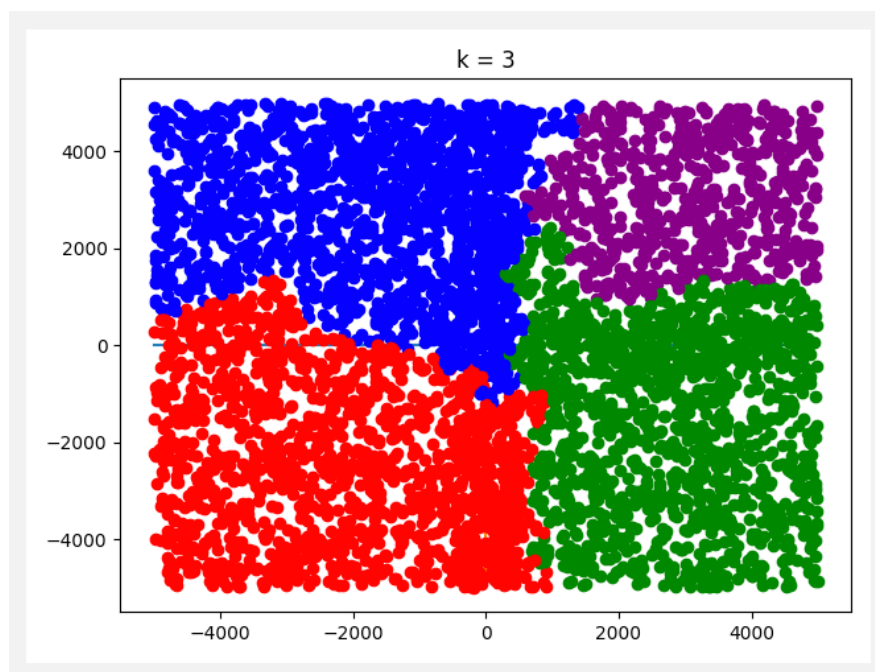
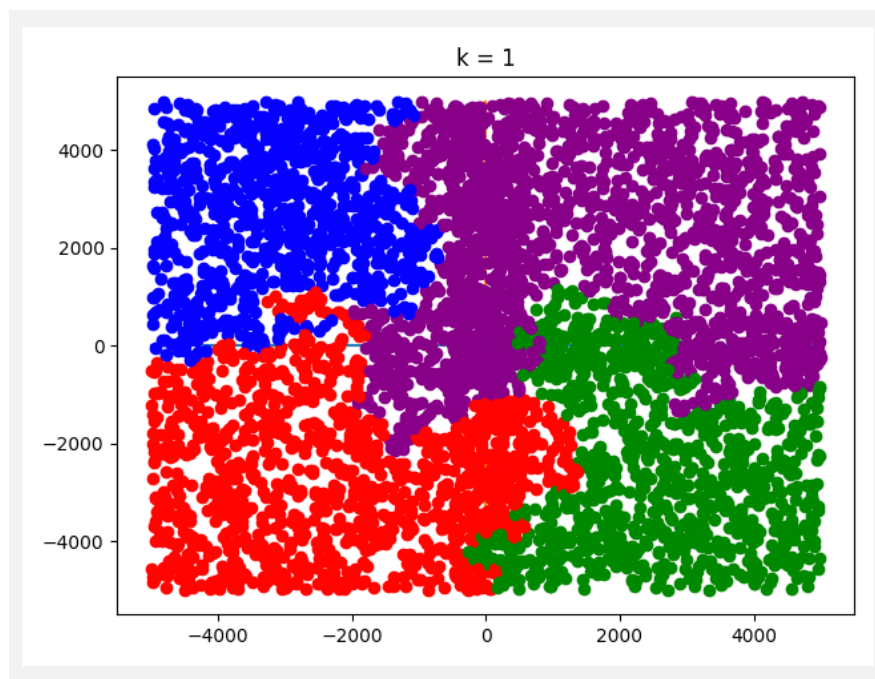
K = 1: 74.11999%

K = 3: 78.52000%

K = 7: 77.98000%

K = 15: 63.94000%

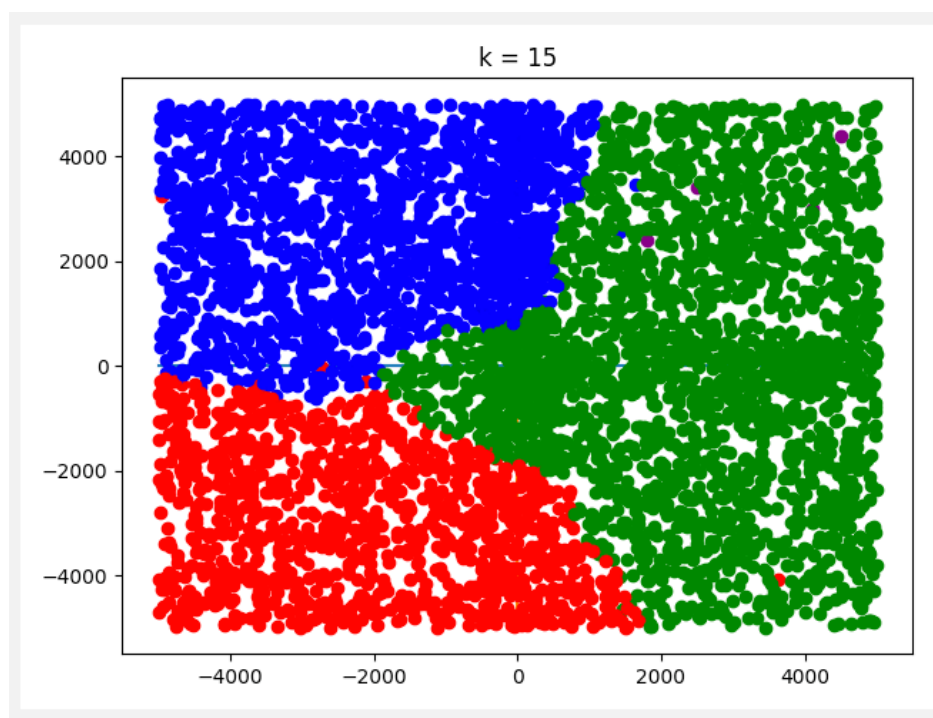
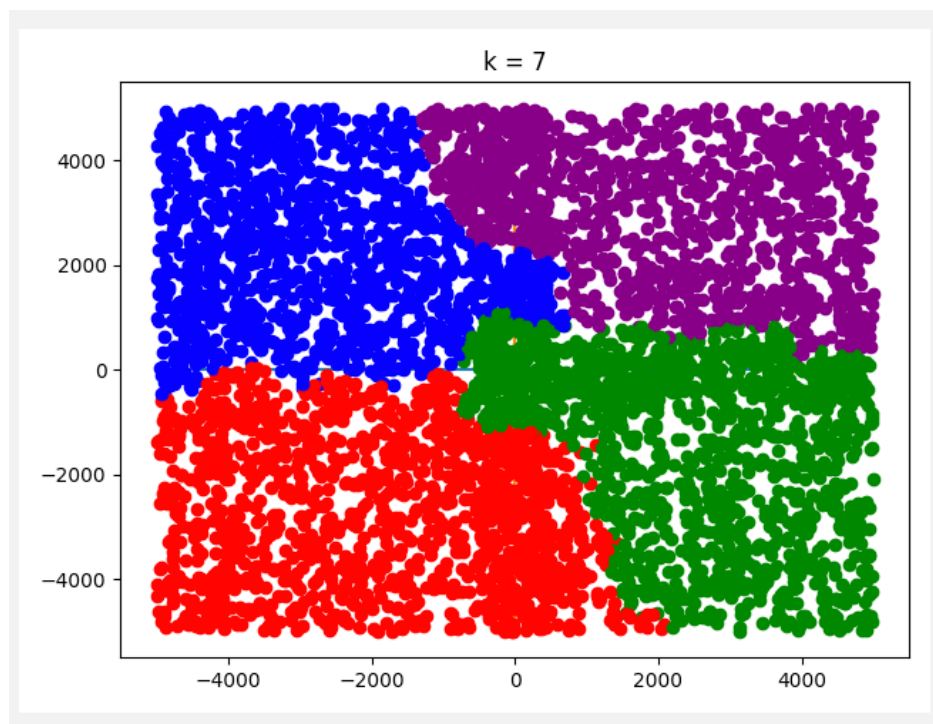
**Vizualizácia:**



Štefan Dudaško, AIS ID: 96835. Cvičiacci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00





Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00

### Test 3.

Úspešnosť klasifikácie:

**Body:** 10 020

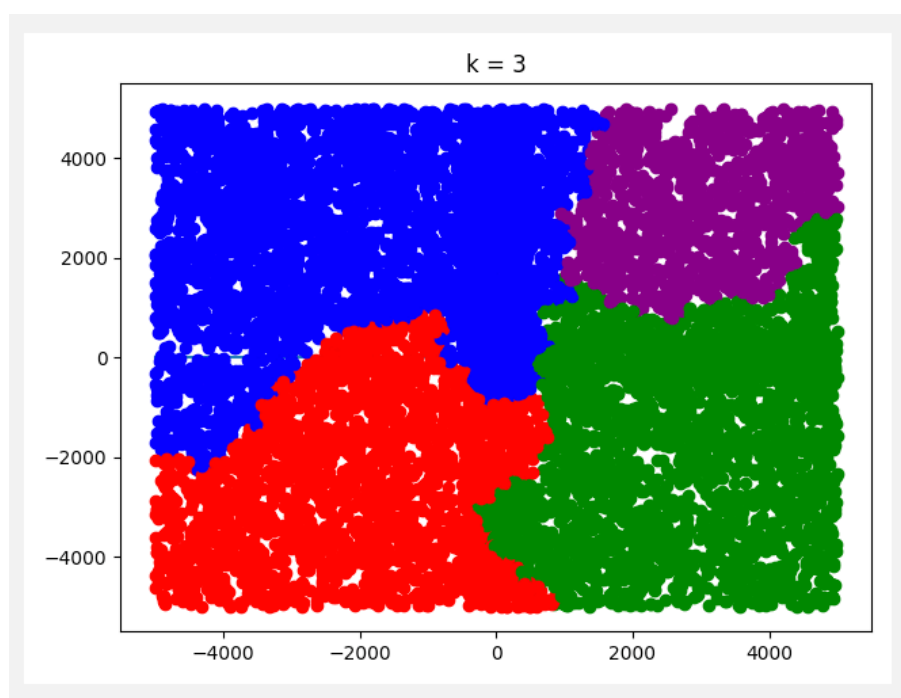
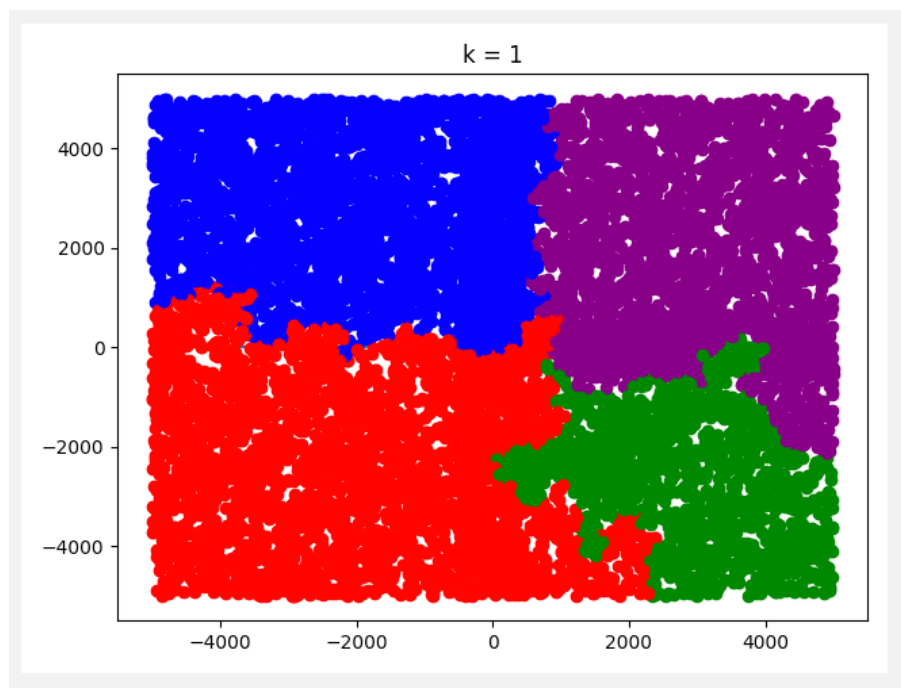
K = 1: 76.00000%

K = 3: 75.21000%

K = 7: 75.98000%

K = 15: 78.30000%

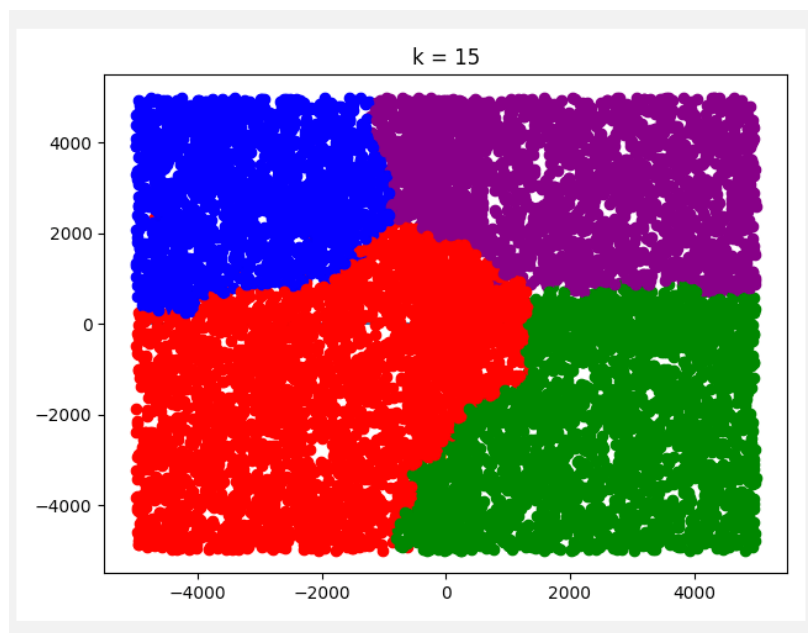
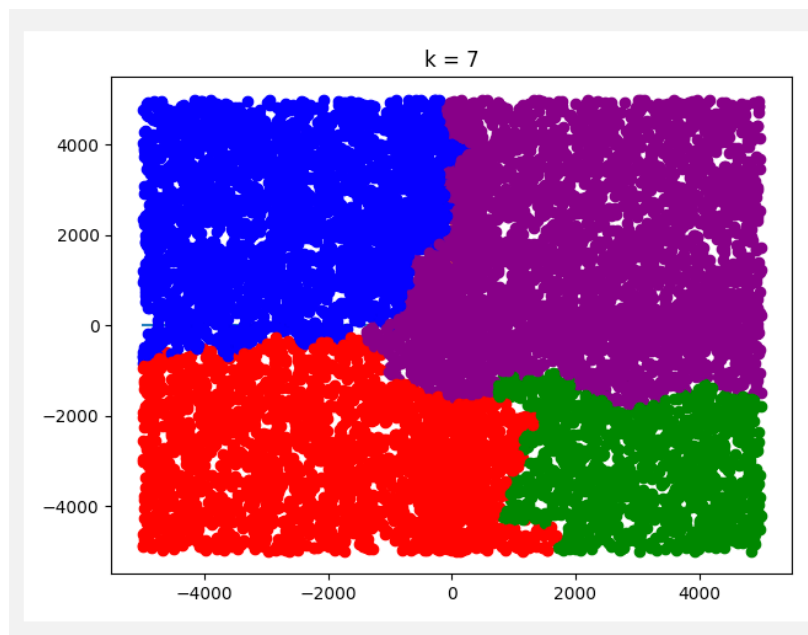
**Vizualizácia:**



Štefan Dudaško, AIS ID: 96835. Cvičiacci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00



Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00

#### Test 4.

Úspešnosť klasifikácie:

**Body:** 20 020

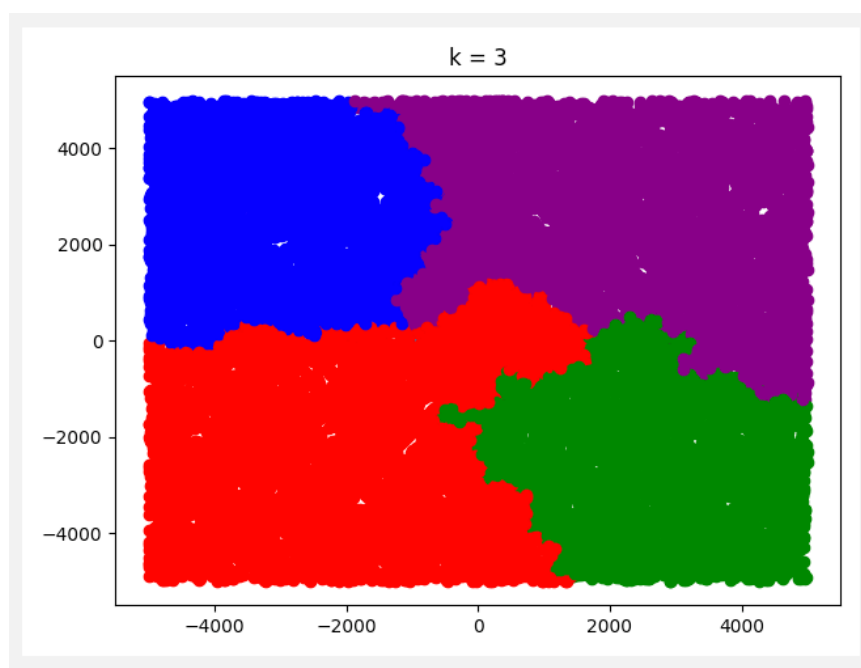
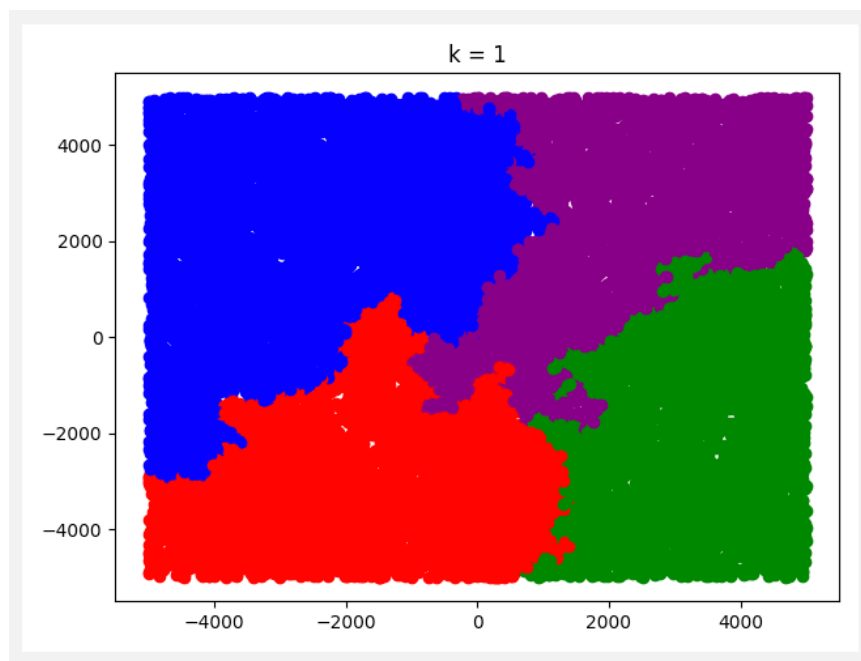
K = 1: 74.21500%

K = 3: 79.28500%

K = 7: 75.16500%

K = 15: 47.88500%

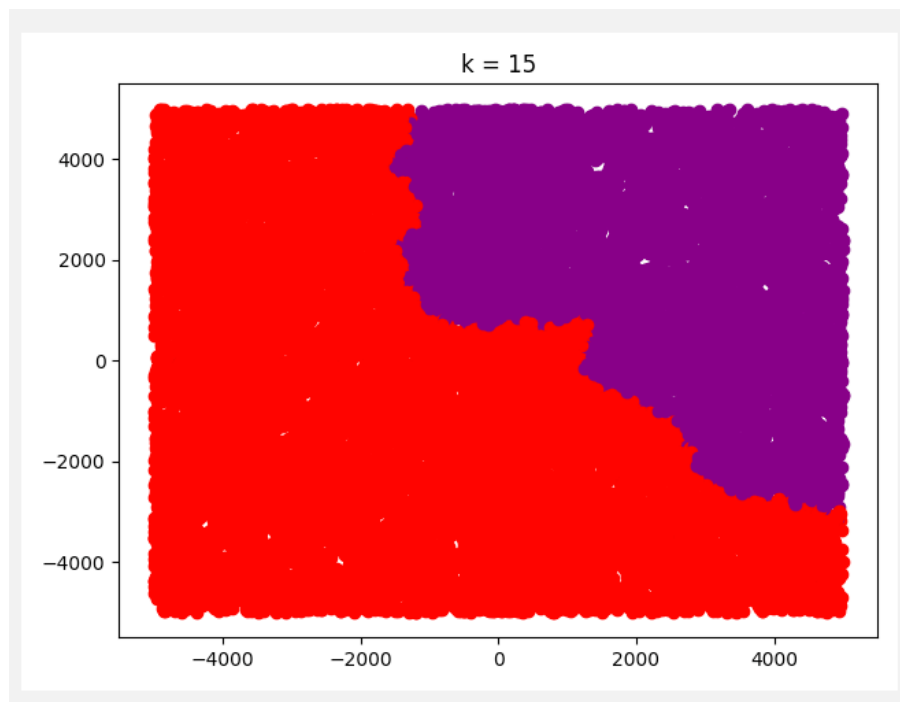
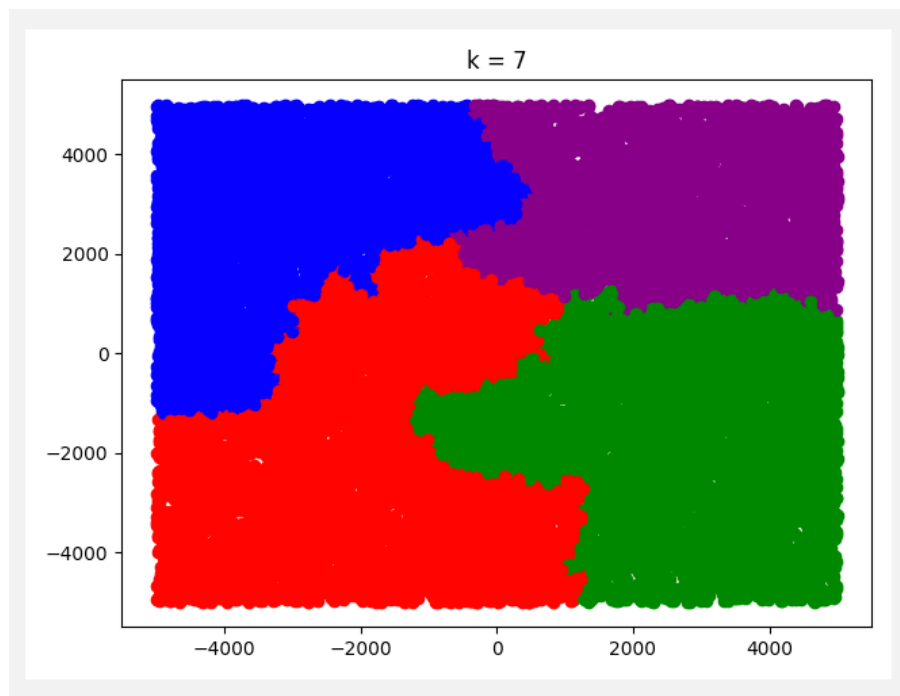
**Vizualizácia:**



Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00



Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00

### Test 5.

Úspešnosť klasifikácie:

**Body:** 40 020

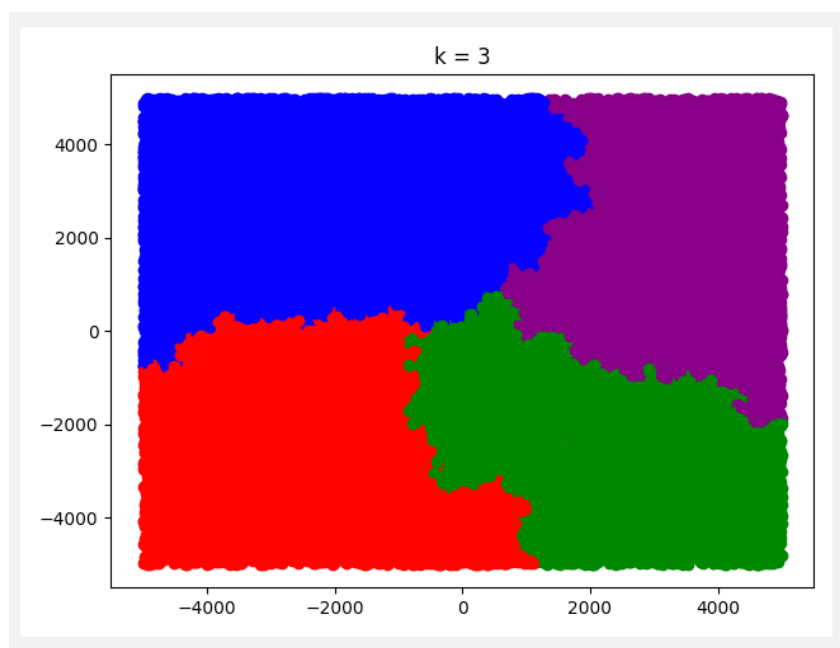
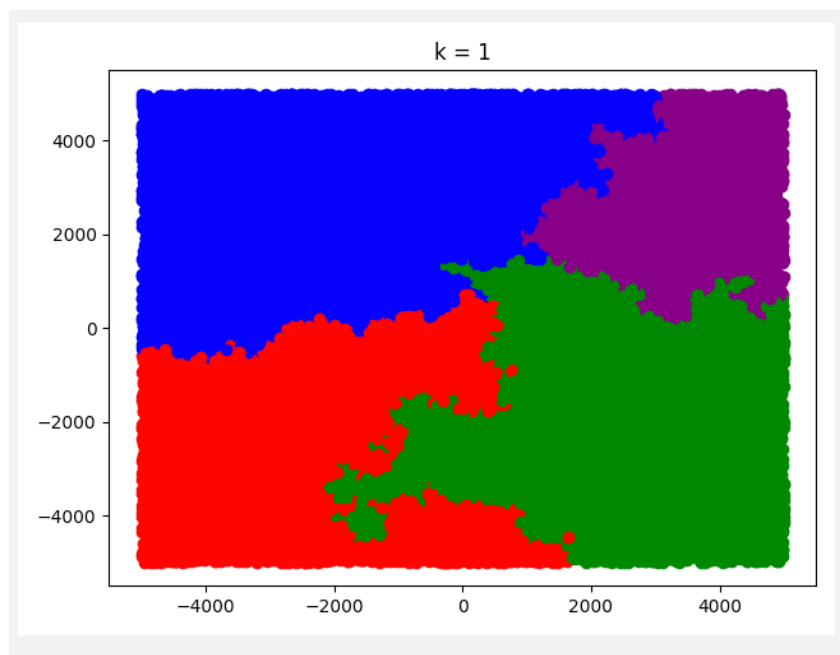
K = 1: 75.54250%

K = 3: 78.42500%

K = 7: 79.69250%

K = 15: 65.57250%

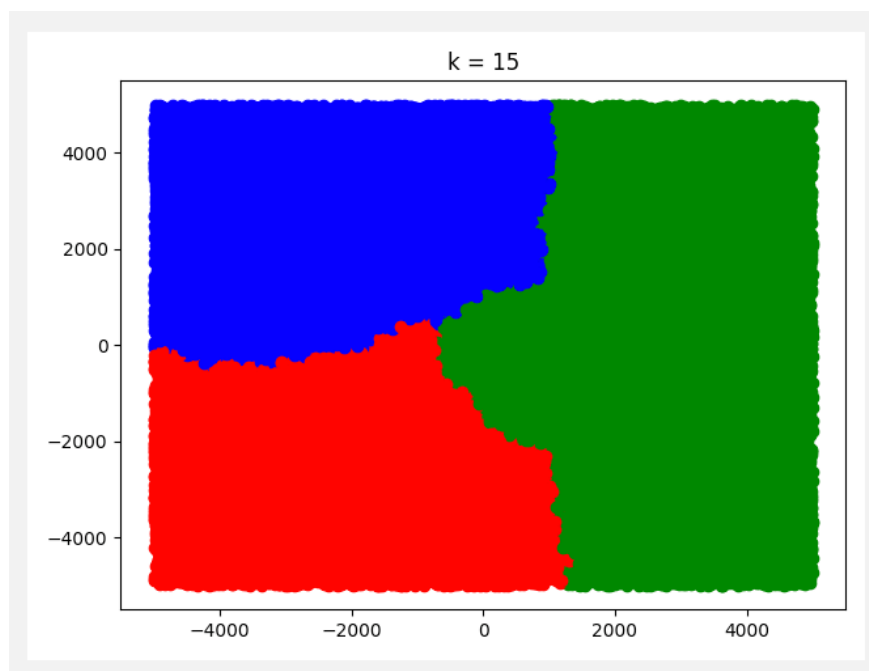
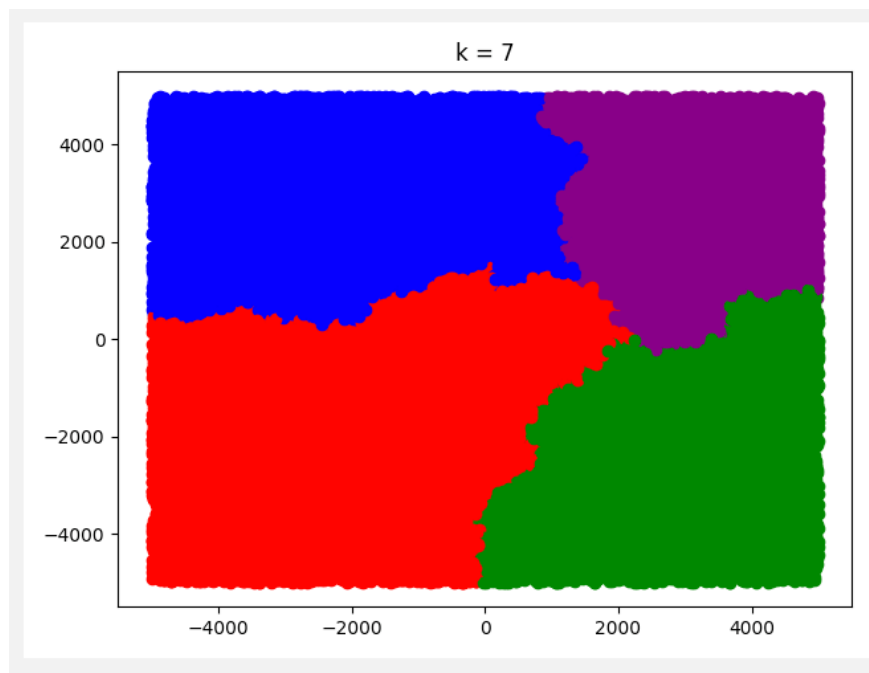
**Vizualizácia:**



Štefan Dudaško, AIS ID: 96835. Cvičiaci: Ing. Boris Slíž

UI: zadanie 4, úloha A - klasifikácia

Cvičenie štvrtok: 16:00



## Zhodnotenie

Najprv zhodnotíme obsah dát, ktoré sme spracovali. Vo väčšine prípadov mal náš klasifikátor úspešnosť väčšiu ako 70 percent. Vidíme, že pri vyšších  $k$  sa zvyšuje aj faktor náhodnosti, a je vyššia šanca, že vygenerované prvky budú mať určenú klasifikáciu inú, ako by sme čakali, a táto „chyba“ sa ďalej kumuluje a prevláda jedna skupina nad druhou. Skupina s menej prvkami je teda ďalej utláčaná, a ak by sme generovali ešte viac bodov, pravdepodobne by boli rozdiely skupín ešte väčšie.

Ďalej môžeme vidieť, že niektoré body sú úplne mimo skupiny, je to len v raritných prípadoch, ktoré nastávajú zväčša kvôli tomu, že generujeme pri šanci 1% bod, ktorý môže vytvoriť ostrovčeky. Body sú na seba ale dosť natlačené, tak to nie je veľmi pekne vidieť. Tieto ostrovčeky však pekne diverzifikujú skupinu a vnášajú nám do dát trochu náhodného faktoru.

Zaujímavé bolo tiež pozorovať úpravu trénovacích dát, tie nie sú ukázané vo vizualizáciách, ale je zaujímavé, že ak odoberieme jednej skupine na začiatku nejaký trénovací bod, je extrémne oslabená, a výsledky experimentu sa líšia ešte viac, ako by sme možno čakali, dokonca viac, ako keď pridáme väčší náhodný faktor ako 1 percento.

Projekt ako taký hodnotím za veľmi zaujímavý a rád by som aj v budúcnosti pracoval na viac projektoch, kde si môžem pekne vizualizovať výsledky pokusov.