SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Fakulta informatiky a informačných technológií

# ZADANIE 4

## Úloha a) – klasifikácia

Dávid Kromka
Cvičenie: utorok 16:00
12.12.2021

Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00

# Obsah

Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00

# 1. Riešený problém

Cieľom zadania je vytvorenie algoritmu na klasifikáciu vygenerovaných bodov vložených do 2-rozmerného priestoru s veľkosťou od -5000 do 5000 bodov v smere osi X aj Y. Priestor je rozdelený do 4 rovnako veľkých štvorcových častí podľa súradníc, kde každá časť má pridelenú farbu: červenú, zelenú, modrú a fialovú. Na začiatku je v každom z týchto častí 5 bodov, ktorých pridelená farba je známa. Program postupne generuje náhodne nové body postupne v súradnicových intervaloch podľa jednotlivých farieb s 99% pravdepodobnosťou, zvyšné 1% sú body vygenerované náhodne v rámci celého priestoru. Pomocou algoritmu K-NN sa zistí farba novo vygenerovaného bodu a tento bod je vložený do poľa k ostatným bodom.

# 2. Opis riešenia

Prvým krokom riešenia je vytvorenie poľa súradníc, v ktorom sú uložené súradnice známych bodov a zároveň ich farby, ktoré sú dopredu známe podľa ich súradníc. Do tohto poľa sa budú ukladať aj vygenerované body a im pridelená farba.

V ďalšom kroku sa vo funkcii generate() generujú nové body, kde sa najprv určí, či bod bude vygenerovaný v rámci súradníc konkrétnej farby (pravdepodobnosť 99%) alebo v rámci celého prostredia (pravdepodobnosť 1%). Ak sa vygenerovaný bod má nachádzať v rámci súradníc konkrétnej farby, určí sa, ktorá farba to má byť. Zvyšok po delení iterácie číslom 4 zaručuje, že vygenerovaná farba bude postupne červená, zelená, modrá a fialová. Následne sa náhodne vygenerujú súradnice nového bodu v rámci súradníc zvolenej farby na ploche. Aby nevznikali duplicitné body, kontroluje sa ich výskyt na ploche a ak už vygenerovaný bod existuje, náhodne sa zvolia nové súradnice. Nakoniec sa bod uloží do poľa a generujú sa ďalšie body.

Ak sú všetky body generované, potrebujeme ich klasifikovať vo funkcii classify(array, k), kde v poli array sú súradnice x a y vygenerovaného bodu a k je počet susedov, ktorý sa využíva v algoritme KNN pri klasifikácii. V cykle sa prechádza všetkými súradnicami bodov v poli self.array a z nich sa ráta euklidovská vzdialenosť od súradníc x a y. Do poľa points sa uloží k bodov s najmenšou vzdialenosťou zoradené vzostupne. Najčastejšie sa vyskytujúca farba v týchto bodoch bude farba nového bodu. Ak je viacero najčastejšie sa vyskytujúcich farieb, berie sa do úvahy aj vzdialenosť bodov a vyberie sa tá, s ktorou body sú najbližšie k novému bodu. Funkcia vracia pridelenú farbu, ktorá sa vo funkcii generate() porovnáva s farbou, ktorá by mala byť bodu pridelená a ukladá sa počet správne pridelených farieb.

Vzorec na výpočet euklidovskej vzdialenosti 2 bodov:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

Pri ohodnocovaní bodov sa ohodnotený bod pridá do grafického znázornenia bodov a po ohodnotení všetkých bodov sa toto grafické znázornenie vykreslí. V programe je využitý scatter plot z knižnice matplotlib.pyplot. Znázornenia a výsledky správneho ohodnotenia sú v časti 4. Testovanie.

Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00

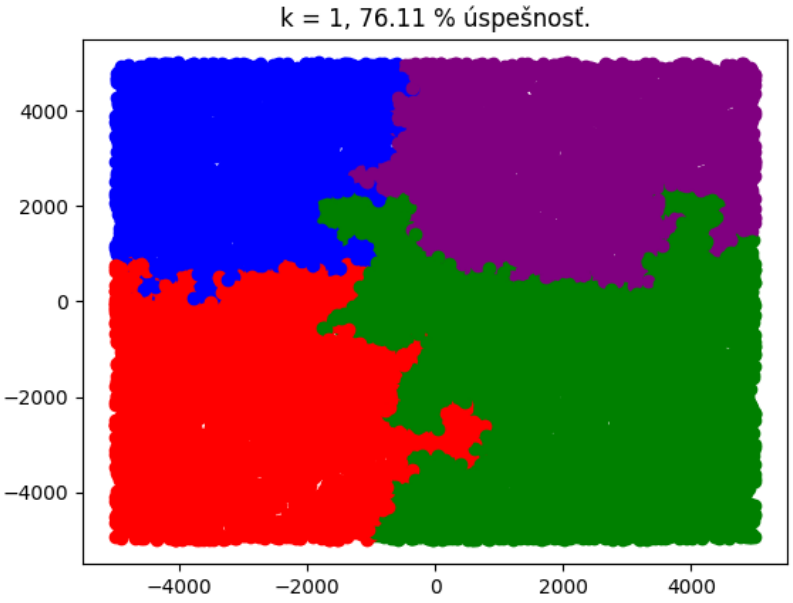
Ohodnocovanie sa vykonáva nad rovnakými bodmi pre rôznu veľkosť k, postupne pre k=1, k=3, k=7 a k=15.

# 3. Použitý algoritmus

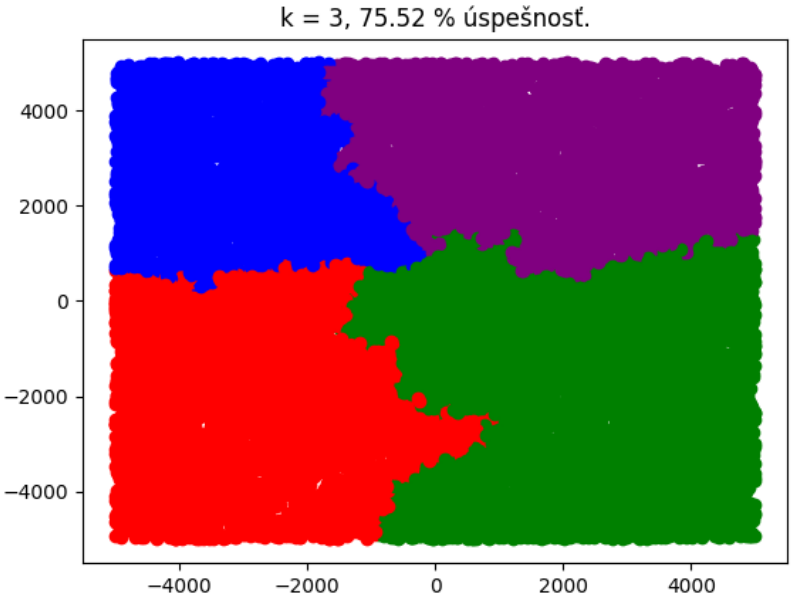
1. Generovanie nových bodov bez opakovanie s postupným striedaním rozmedzia súradníc podľa farby
2. Určenie čísla k – počet susedov
3. Pridelenie farby  bodu podľa algoritmu KNN
4. Uloženie nového bodu s farbou do poľa
5. Kým nie sú všetky vygenerované body klasifikované, znova bod 3
6. Vyhodnotenie úspešnosti a grafický výstup
7. Koniec alebo znova bod 2

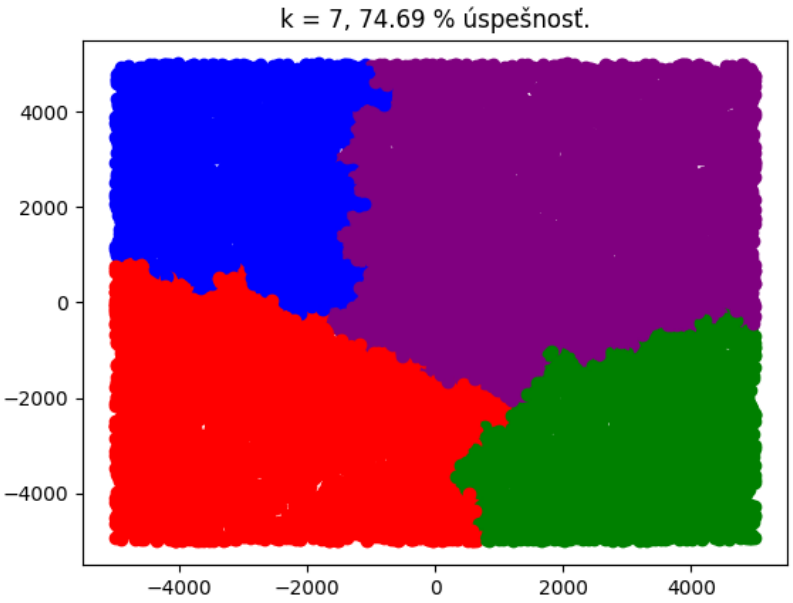
# 4. Testovanie

## 4.1. Test 1

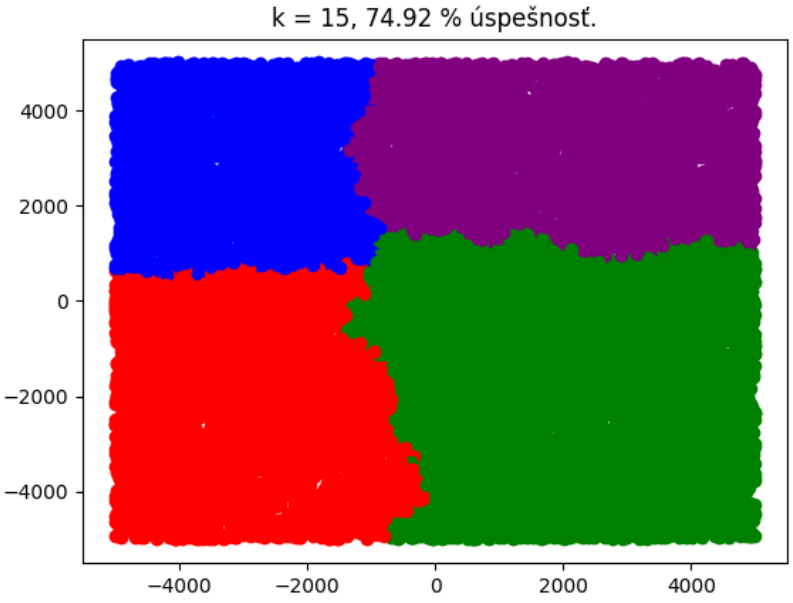


Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00



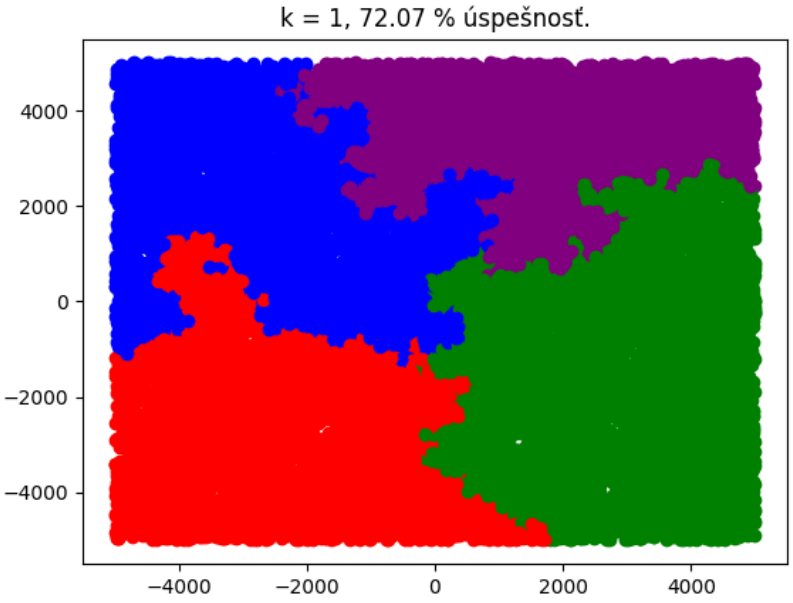
Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00

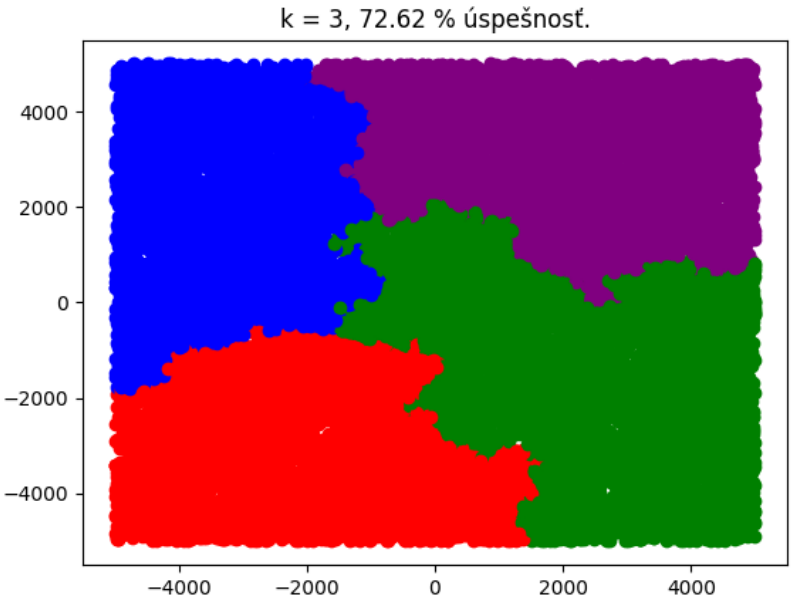



```
15222
Čas vykonania pre k = 1: 680.9478664398193 s
15104
Čas vykonania pre k = 3: 682.9336981773376 s
14937
Čas vykonania pre k = 7: 682.5983061790466 s
14985
Čas vykonania pre k = 15: 684.8177642822266 s
```

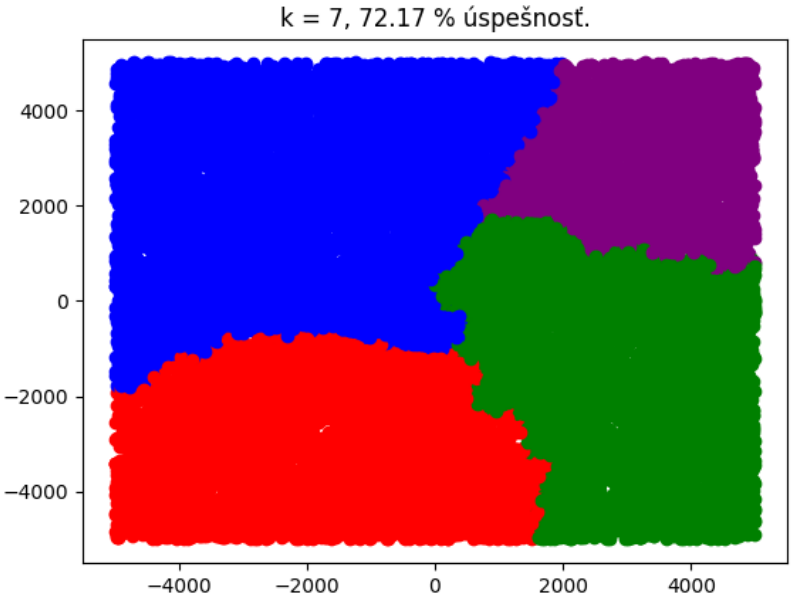
Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00

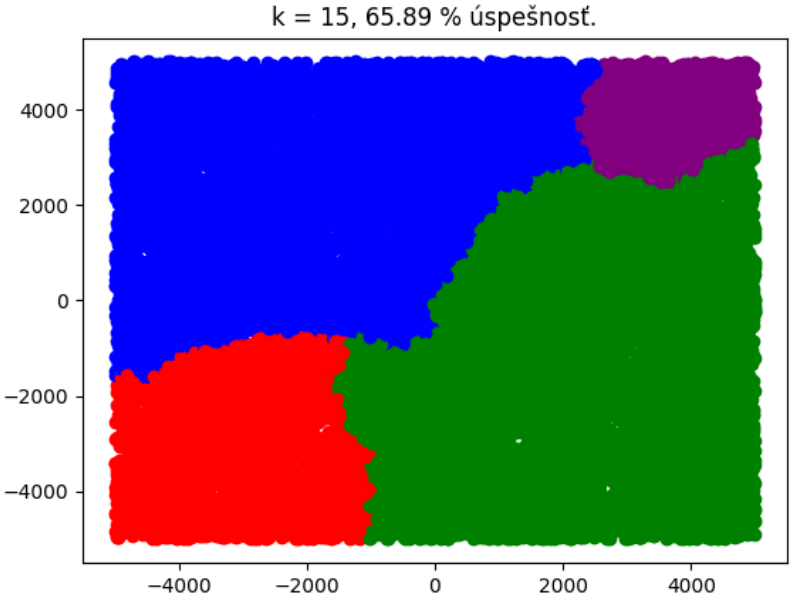
## 4.2. Test 2



Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00




```
k = 1, 72.07 % úspešnosť.
Čas vykonania pre k = 1: 690.7666802406311 s
k = 3, 72.62 % úspešnosť.
Čas vykonania pre k = 3: 669.8025033473969 s
k = 7, 72.17 % úspešnosť.
Čas vykonania pre k = 7: 721.5314273834229 s
k = 15, 65.89 % úspešnosť.
Čas vykonania pre k = 15: 693.1343395709991 s
```

Meno: Dávid Kromka
Ič: 110834
Cvičenie: Slíž, utorok 16:00

### 4.3. Test 3

Test 3 je vykonaný bez grafického znázornenia, čo malo výrazný vplyv na čas vykonávania a môžeme usúdiť, že grafické znázornenie bodov zaberá väčšinu času behu programu.

```
k = 1, 73.22 % úspešnosť.
Čas vykonania pre k = 1: 129.97339606285095 s
k = 3, 75.46 % úspešnosť.
Čas vykonania pre k = 3: 131.3465700149536 s
k = 7, 74.43 % úspešnosť.
Čas vykonania pre k = 7: 141.7304129600525 s
k = 15, 58.04 % úspešnosť.
Čas vykonania pre k = 15: 136.4710431098938 s
```

## 5. Zhodnotenie

Z výsledkov testovania vyplýva, že najvyššiu úspešnosť má klasifikácia pri hodnote k=1, k=3 a dobré výsledky sú aj pri k=7. Pri hodnote k=15 boli výsledky v každom z testov výrazne horšie od ostatných.

Čas vykonania klasifikácie pre k s grafickým výstupom trvá v priemere 600 až 700 sekúnd, bez grafického znázornenia je to výrazne menej, niečo vyše 2 minút. Zložitosť by bolo možné znížiť tak, že by sa pri klasifikácii neprechádzalo celým poľom, ale len určitou časťou plochy, v ktorej sa bod nachádza a v nej sa vyhľadávali susedia. Body by sa nepridávali do jedného veľkého poľa, ale do samostatných menších polí podľa súradníc.

Program je implementovaný v jazyku Python 3.9.7 s využitím knižníc random, math, matplotlib a time.