Istraživanje podataka - primer 2. testa

1. Dati skup podataka predstavlja broj pojavljivanja pojmova X, Y, Z u različitim dokumentima. Algoritmom K-sredina identifikovati 3 klastera u tim podacima. Pri tom, koristiti kosinusno rastojanje kao meru sličnosti.

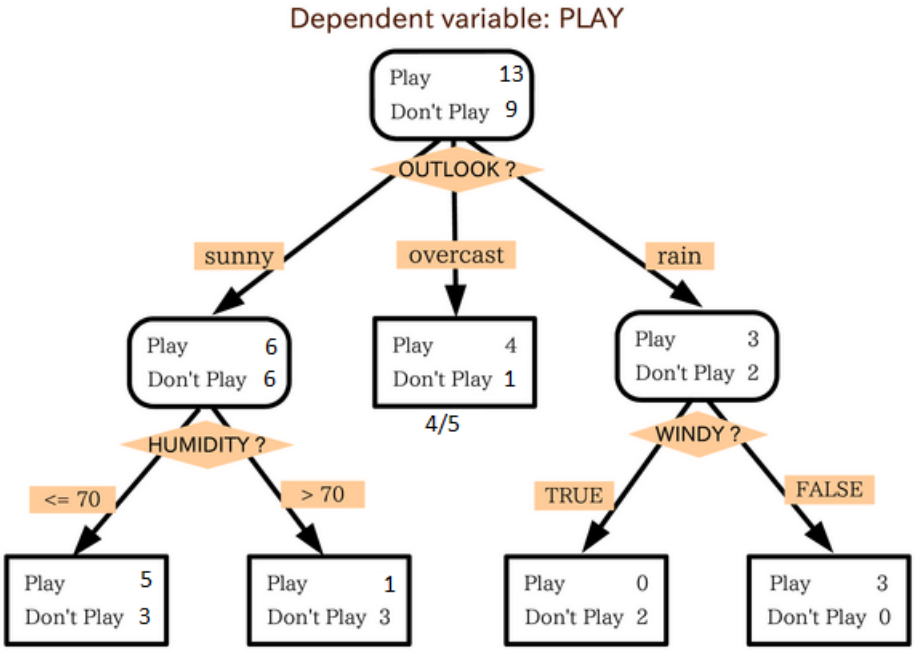
X	Y	Z
2	2	10
5	5	5
6	5	7
2	3	12
20	13	2
25	4	1

2. Date su niske ACU, ACC, GCU, GCC i UAU. Koristeći Hamingovo rastojanje za niske napraviti matricu različitosti datih niski i zatim izvršiti hijerarhijsko klasterovanje korišćenjem *max* veze. Rezulatat prikazati dendogramom.

3. Na osnovu datih podataka o osobama proceniti da li osoba sa osobinama (*Fizička aktivnost=Da, Umerena ishrana=Ne, Stres=Da, Visok pritisak=Ne*) ima bolest srca korišćenjem stabla odlučivanja. Kao meru nečistoće koristiti Ginijev indeks.

Fizička aktivnost	Umerena ishrana	Stres	Visok pritisak	Obolenje srca
Da	Da	Ne	Da	Ne
Da	Ne	Da	Ne	Ne
Da	Da	Da	Da	Da
Ne	Ne	Da	Da	Da
Ne	Da	Ne	Ne	Ne
Ne	Da	Da	Ne	Da

4. Na slici je prikazano drvo odlučivanja za klasifikaciju.



Izdvojiti pravila za klasifikaciju i odrediti najbolje pravilo za klasifikaciju prema pouzdanosti. Napraviti matricu konfuzije za trening skup i odrediti preciznost.