

Závěrečná správa

Databáza:

Ako databázu sme použili súbor dát o hříboch. Obsahuje vzorky zodpovedajúce 23 druhom rebrových hříbov. V databáze máme 8124 hříbov s 23 príznakmi.

Popis příznakov

1. Jedlost': jedlé=e, jedovaté=p
2. Tvar klobúku: zvon=b, kuželový=c, konvexný=x, plochý=f, uzlový=k, klesnutý=s
3. Povrch klobúku: vláknitý=f, drážkový=g, šupinatý=y, hladký=s
4. Farba klobúku: hnedá=n, žlttohnedá=b, škoricová=c, g=šedá, zelená=r, ružová=p, fialová=u, červená=e, biela=w, žltá=y
5. podliatiny: ano=t, nie=f
6. pach: mandle=a, aníz=l, kreozot=c, rybí=y, hnusný=f, zatuchnutý=m, žiaden=n, ostrý=p, pikantné=s
7. Pripevnenie rebra: pripojený=a, klesajúce=d, voľné=f, vrúbkované=n
8. Umiestnenie repra: blízko=c, preplnené=w, vzdialené=d
9. Veľkosť rebra: široká = b, úzka = n
10. Farba rebra: čierna=k, hnedá=n, žlttohnedá=b, čokoláda=h, sivá=g, zelená=r, oranžová=o, ružová=p, fialová=u, červená=e, biela=w, žltá=y
11. Tvar stonky: zväčšujúci=e, zužujúci=t
12. Koreň stonky: baňatý=b, kyj=c, šálka=u, rovný=e, vláknitý=z, korene=r, chýba=?
13. Povrch stonky nad prstencom: vláknitý=f, šupinatý=y, hodvábný=k, hladký=s
14. Povrch stonky pod prstencom: vláknitý=f, šupinatý=y, hodvábný=k, hladký=s
15. Farba stonky nad prstencom: hnedá=n, žlttohnedá=b, škoricová=c, g=šedá, oranžová=o, ružová=p, červená=e, biela=w, žltá=y
16. Farba stonky pod prstencom: hnedá=n, žlttohnedá=b, škoricová=c, g=šedá, oranžová=o, ružová=p, červená=e, biela=w, žltá=y
17. Typ závoja: čiastočný=p, univerzálny=u
18. Farba závoja: hnedá=n, oranžová=o, biela=w, žltá=y
19. Počet prstencov: žiaden=n, jedna=o, dva=t
20. Typ prstencov: pavučinový=c, pomínuteľný=e, rozširujúci=f, veľký=l, nič=n, príveskový=p, oplášťovací=s, zóna=z
21. Farba stopy výtrusu: čierna=k, hnedá=n, žlttohnedá=b, čokoláda=h, zelená=r, oranžová=o, purpurová=u, w=biela, žltá=y
22. obsadenie: hojné=a, skupinové=c, početné=n, roztrúsené=s, niekoľko=v, osamelé=y
23. nálezisko: trávy=g, lístia=l, lúky=m, cesty=p, mestá=u, odpad=w, lesy=d

Ako je z popisu vidno niektoré huby nemajú zaradený koreň stonky. Keďže naša databáza obsahuje veľké množstvo húb sme tento nedostatok vyriešili nezaradením húb do vzorky, s ktorou sme následne pracovali a teda nám ostalo 5936 húb.

Zamerali sme sa najmä na zistenie jedlosti húb keďže v popise databázy bolo spomenuté, že sa jedlosť dá odvodiť s vysokou pravdepodobnosťou dokonca niekoľkými spôsobmi len na základe logických foriem typu farba stopy výtrusu=zelená, ktorá na 99% trafi jedovatý húb. Ďalší príznak, ktorý sme sa snažili odvodiť bolo nálezisko ktoré nám pripadalo, že sa bude dať tiež dobre odvodiť, keďže prostredie má vplyv na rastliny aj zvieratá

Redukcia príznakov

Na redukcii príznakov sme použili algoritmy PCA a ICA.

PCA

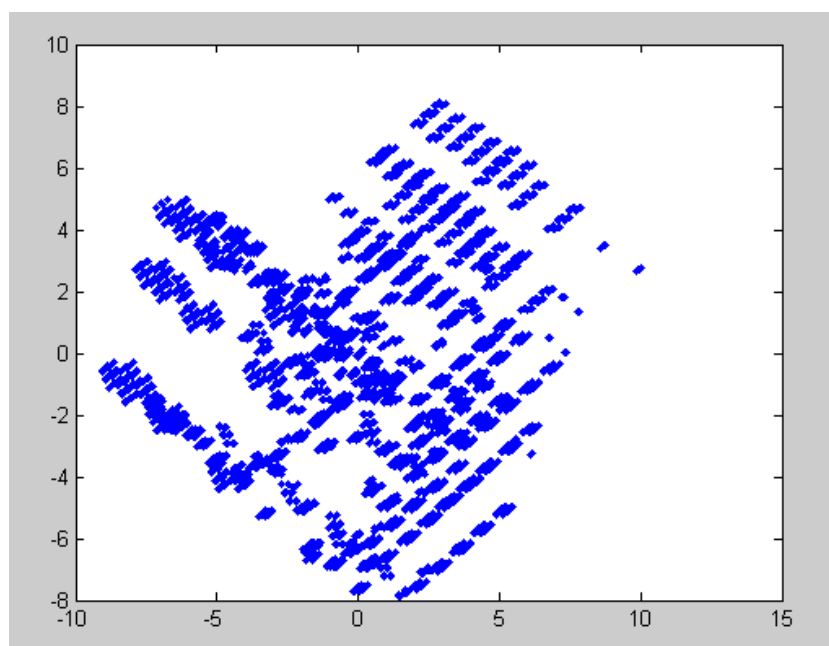
Tento algoritmus vyberá základné prvky na základe variance pričom sa snaží ponechať kolmost medzi prvkami. Toto však znamená, že najväčší vplyv budú mať príznaky s najväčším počtom hodnôt. V našom prípade teda budú mať najväčší vplyv príznaky farby rôznych častí húb. V prípadoch že zredukujeme dáta na málo dimenzii sa môže stať že nám ostanú hlavne farby, ktoré nemusia pomôcť pri predpovedi požadovaného príznaku.

ICA

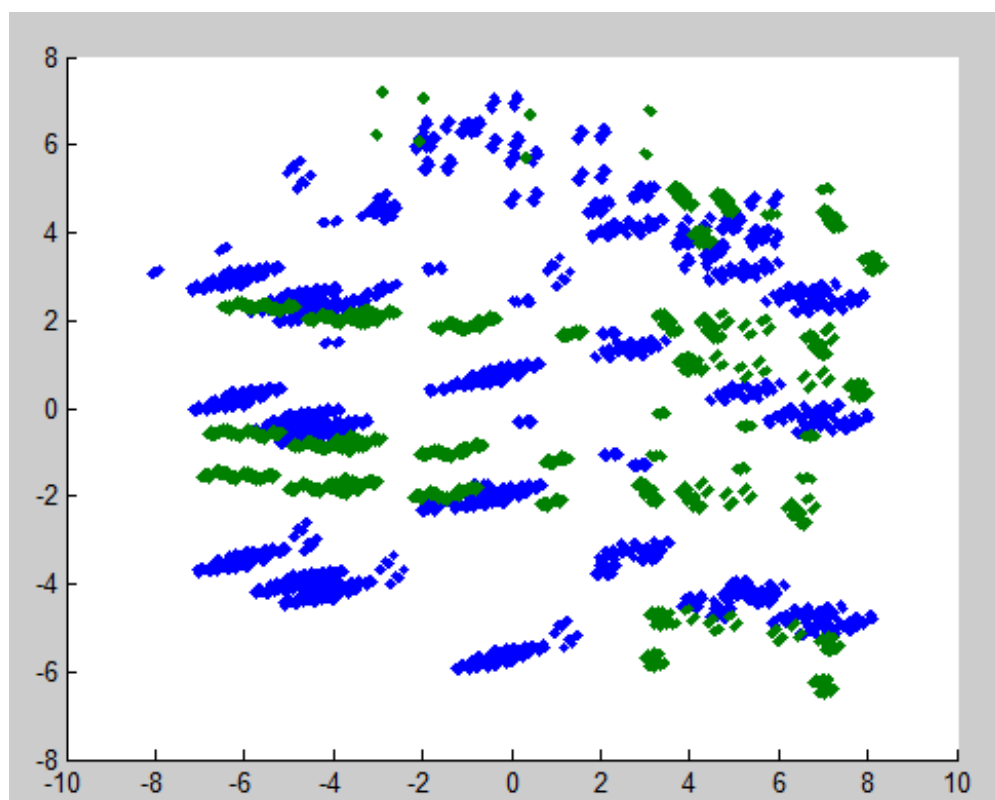
ICA predpokladá že príznaky dostane lineárnou kombináciou skrytých premenných. Na rozdiel od väčšiny algoritmov vrátane PCA tento algoritmus tiež predpokladá negausovské rozloženie dát. Tento algoritmus však nevie separovať perfektne gausovské dáta, avšak keďže nevieme ako sú naše huby rozložené je tento algoritmus dobrá druhá voľba.

Vizualizácia dát

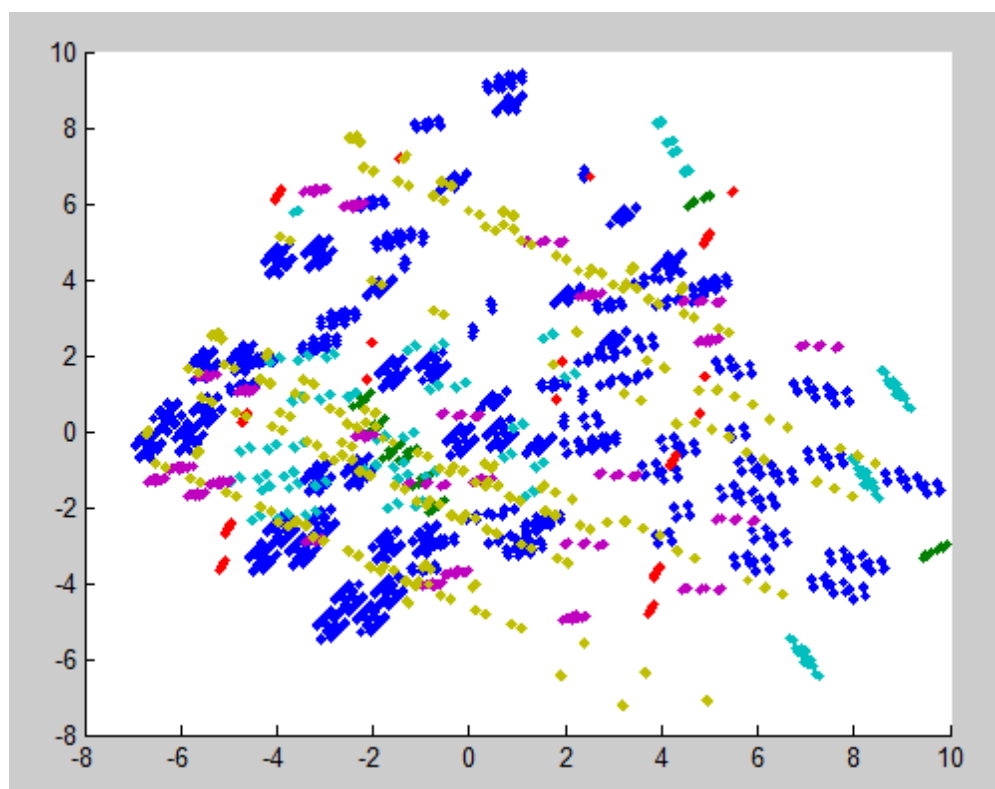
Nezredukované 23 rozmerné dáta:



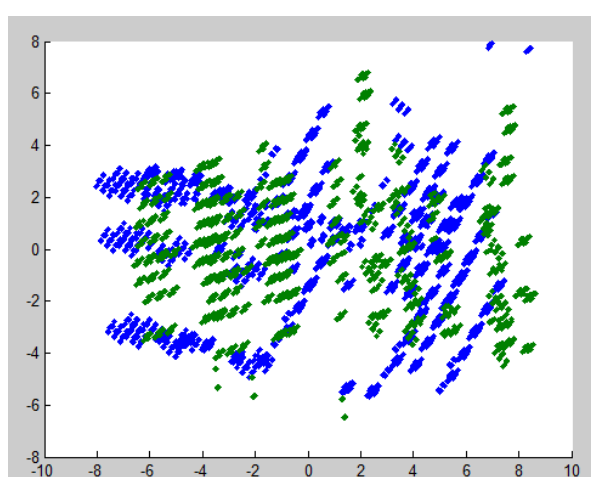
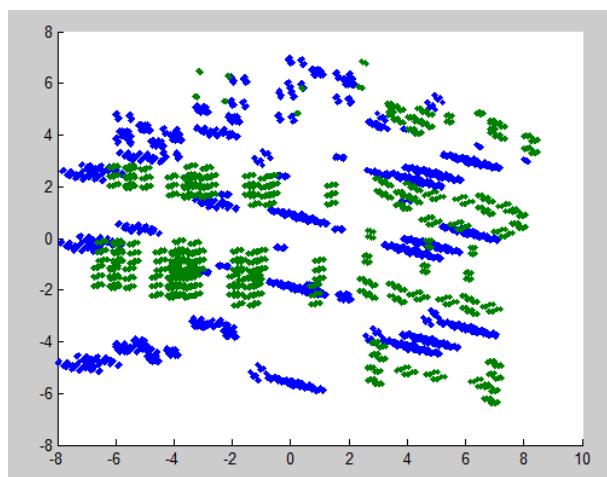
Dáta rozdelené podľa jedlosti:



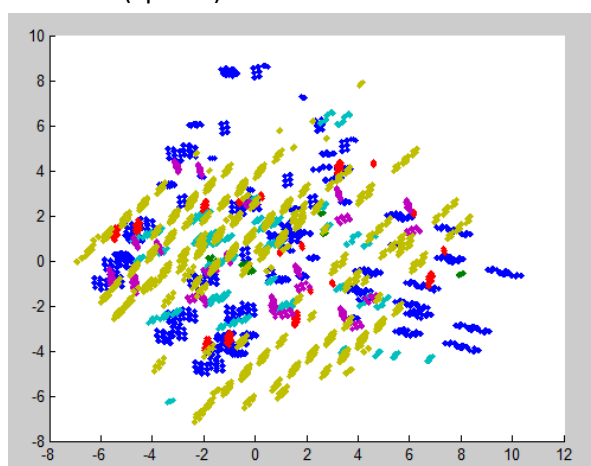
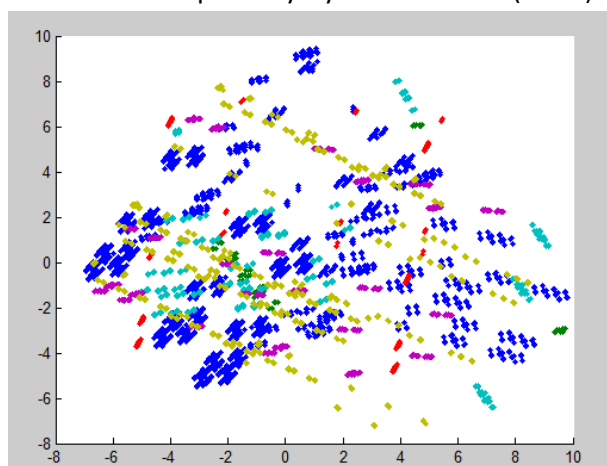
Dáta rozdelené podľa náleziska:



Dáta rozdelené podľa jedlosti zredukované na 5 dimenzii (vľavo) 2 dimenzie (vpravo)



Dáta rozdelené podľa výskytu 10 dimenzii (vľavo) 2 dimenzie (vpravo)



Klasifikácia

Dáta sme klasifikovali pomocou lineárneho klasifikátora, neurónových sietí, SOM a rozhodovacieho stromu. Všetky tieto klasifikátory sme najprv naučili sa rozhodovať na trénovacej množine a následne sme ich nechali rozhodovať na ostatku dát.

Lineárny klasifikátor

Rozdeľuje vstupné dáta nadrovinou, tak aby mal jednu triedu na jednej strane a druhú triedu na druhej strane nadroviny.

Predpoklad je, že lineárny klasifikátor nebude veľmi úspešný, lebo dáta by museli byť lineárne separabilné vo veľkých dimenziách, alebo zredukovateľné do veľmi malej dimenzie.

Neurónové siete

Sieť ktorá si na trénovacej množine upraví váhy vstupných dát a prechodov medzi neurónmi šírením chyby, tak aby sme dostávali čo najlepšie výsledky. Pri vhodnom učení dosiahne globálne minimum chyby a preto ak zvolíme vhodnú trénovaciu množinu je pravdepodobné že bude pri klasifikácii správne klasifikovať naše huby.

Predpoklad je, že neurónové siete by mohli byť veľmi úspešné. Náhodné trénovanie ukázalo, že sú schopné trénovaciu množinu pri jedlosti rozdeliť so 100% úspešnosťou a nálezisko s 80% úspešnosťou.

SOM

Samo organizujúce mapy sú podobne ako neurónové siete sústava pospájaných neurónov avšak sú pospájané do mriežky a vstupné dáta sú spojené s každým neurónom. Výsledkom SOM je mapa kde podobné huby budú mapované blízko seba a rozdielne huby oddelene.

Predpoklad je, že SOM by mohli byť pomerne úspešné, keďže huby s rovnakými vlastnosťami by mohli byť podobné aj v iných aspektoch.

Rozhodovací strom

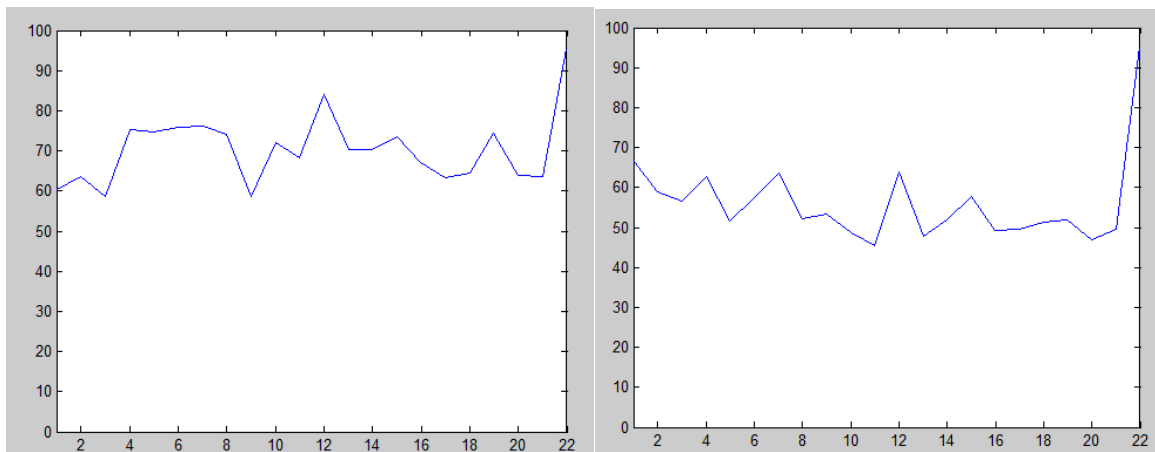
Vytvára strom tak že pre každý uzol vezme príznak s najlepším ohodnotením a rozvetví ho na základe hodnoty tohto príznaku. Dáta rozdeľuje na základe vetvenia stromu na podmnožiny až kým nedostane jednoznačné rozdelenie požadovaného príznaku. Teda každý list obsahuje len hľadaný príznak s jednou hodnotou.

Rozhodovací strom by mal byť tiež pomerne úspešný. Vybrali sme ho preto, aby sme overili pravidlá „starých mám“, ktoré hovoria o jedlosti huby podľa niektorých jej znakov tak ako rozhodovací strom.

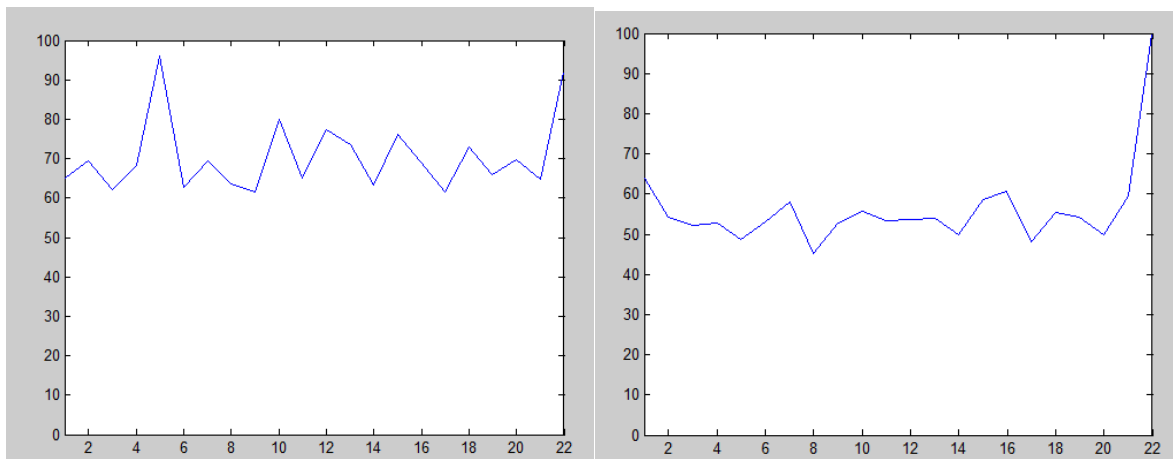
ROC krivky

Na validáciu sme použili vrstvenú k-násobnú krížovú validáciu. Na X-ovej osi ROC kriviek máme počet parametrov, na Y-ovej osi percentuálnu úspešnosť. Najprv ROC krivky zobrazujúce úspešnosť zisťovania jedlosti a potom krivky pre zisťovanie náleziska.

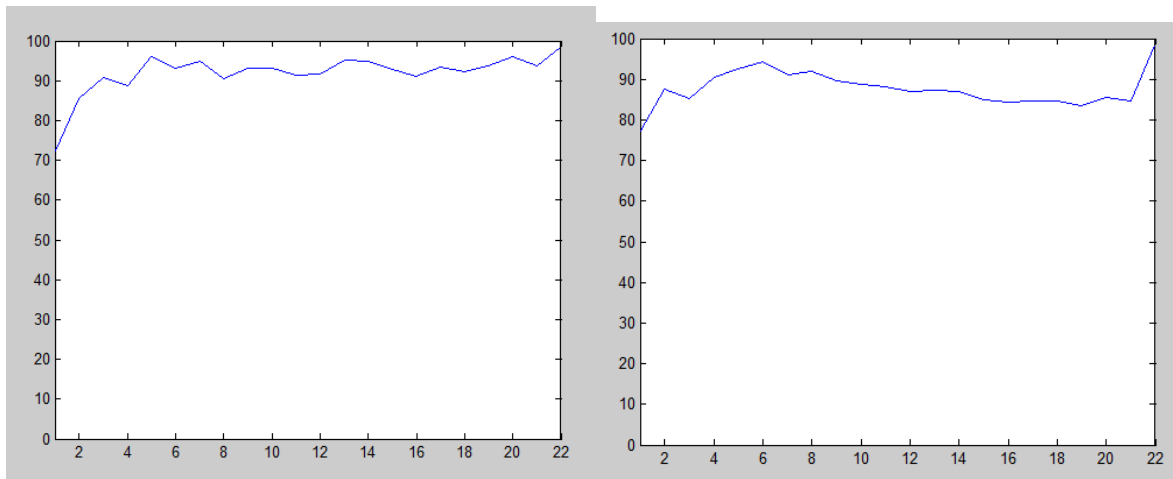
ROC krivka na zisťovanie **jedlosti** pre **lineárny klasifikátor** naľavo s PCA na pravo ICA



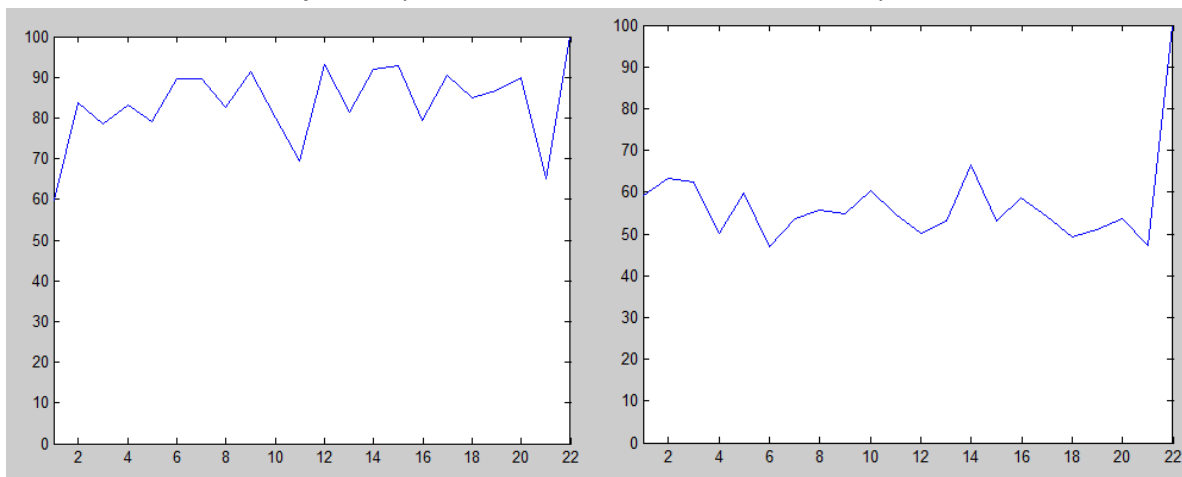
ROC krivka na zisťovanie **jedlosti** pre **neurónové siete** naľavo s PCA na pravo ICA



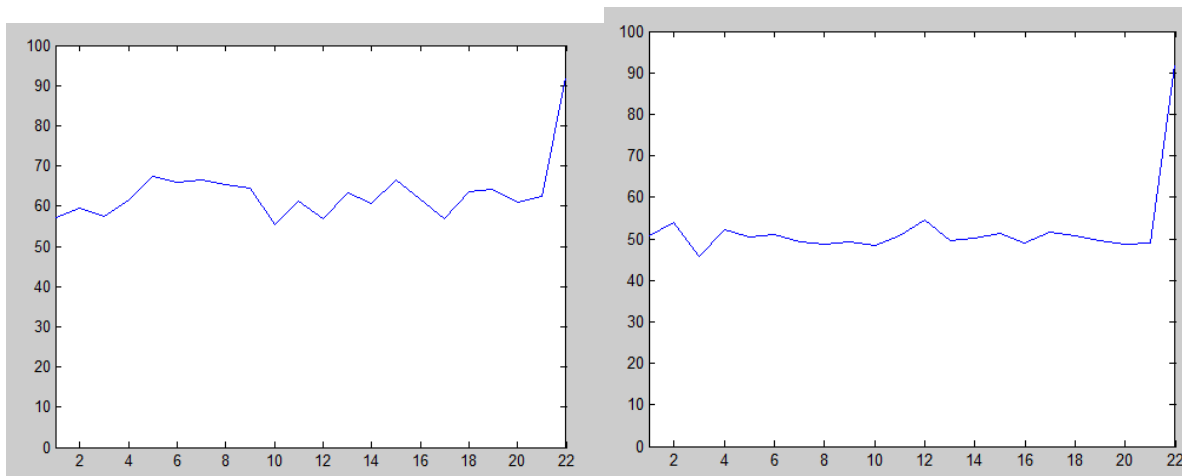
ROC krivka na zisťovanie **jedlosti** pre **SOM** naľavo s PCA na pravo ICA



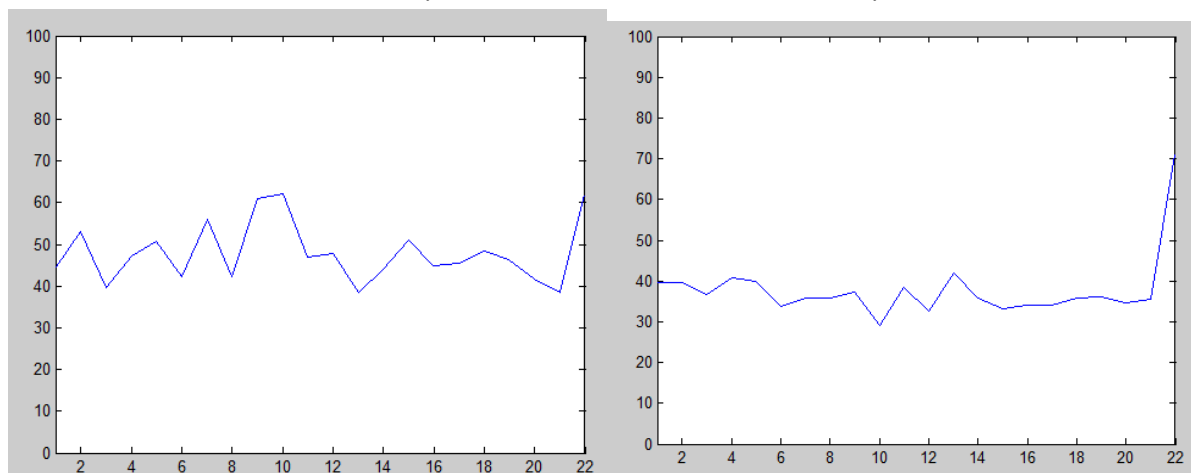
ROC krivka na zisťovanie **jedlosti** pre **rozhodovací strom** naľavo s PCA na pravo ICA



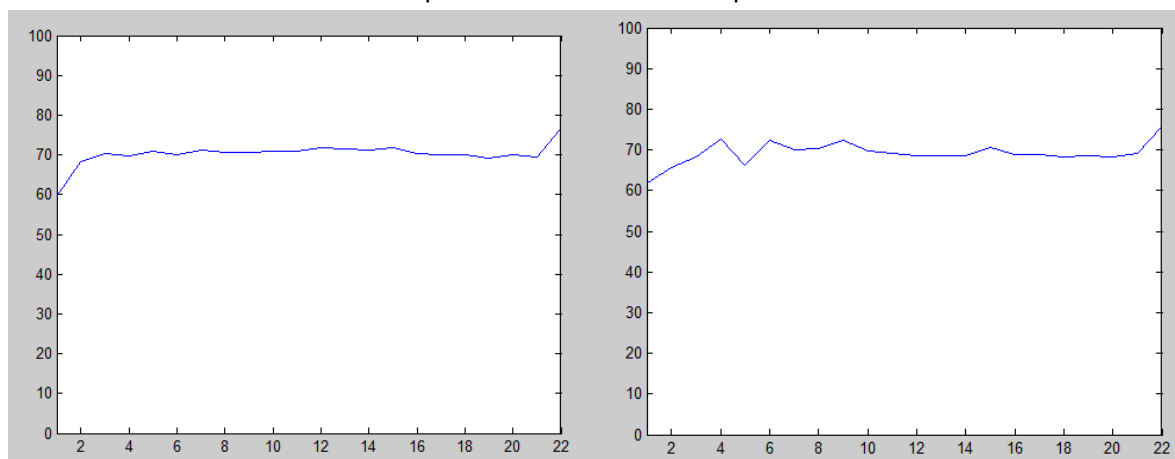
ROC krivka na zisťovanie **náleziska** pre **lineárny klasifikátor** naľavo s PCA na pravo ICA



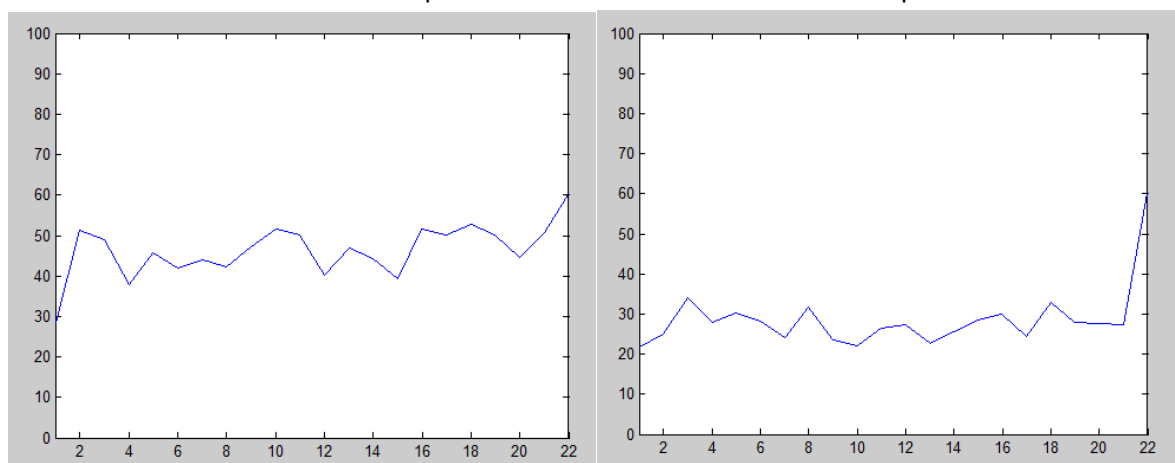
ROC krivka na zisťovanie **náleziska** pre **neurónové siete** naľavo s PCA na pravo ICA



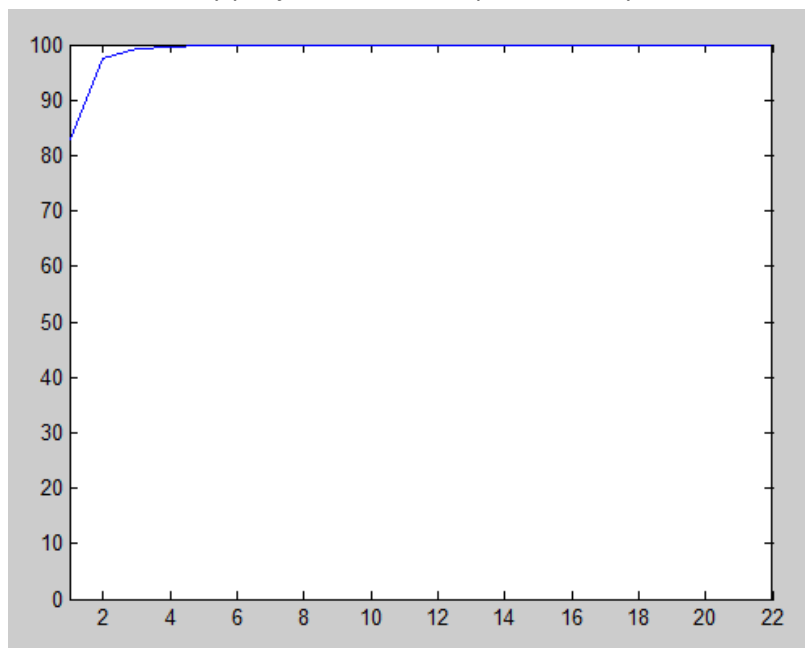
ROC krivka na zisťovanie **náleziska** pre **SOM** naľavo s PCA na pravo ICA



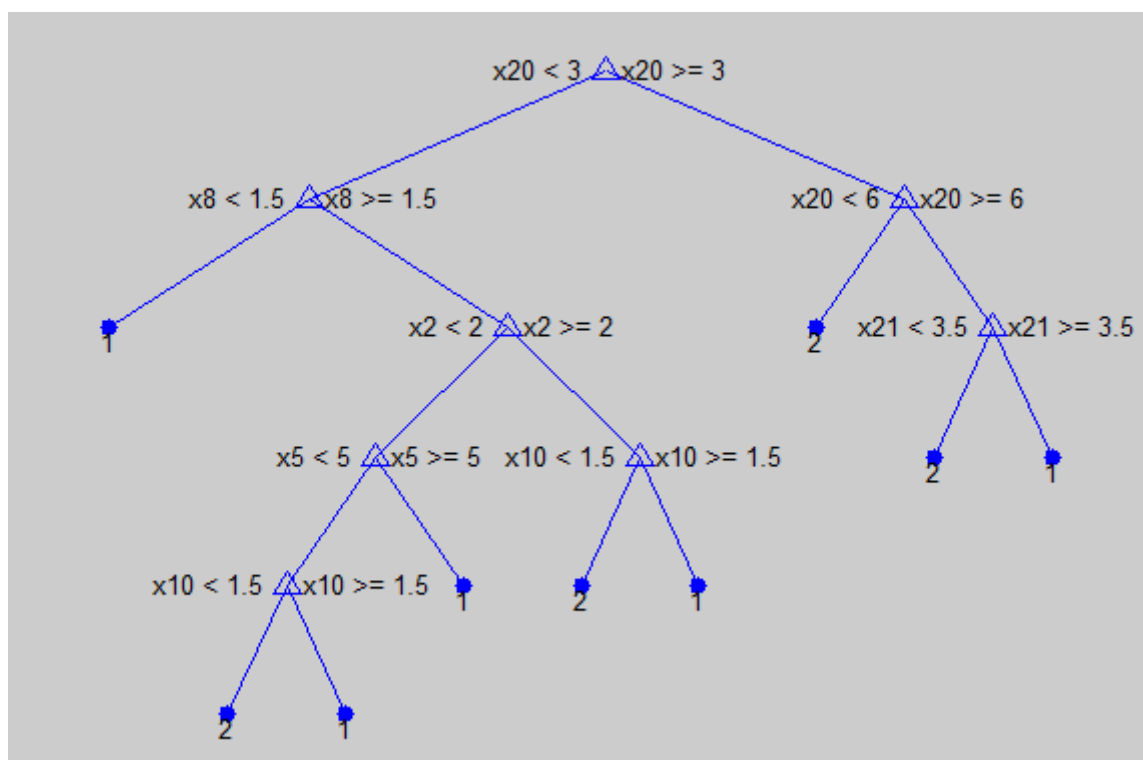
ROC krivka na zisťovanie **náleziska** pre **rozhodovací strom** naľavo s PCA na pravo ICA



Príklad ROC krivky pre jedlosť s PCA na pretrénovaných stromoch



Vizualizácia rozhodovacieho stromu pre jedlosť:



Confusion matice

Uvádame len niektoré confusion matice lebo ostatné boli veľmi podobné.

Confusion matica pre jedlosť na neurónových sieťach s PCA

	Jedlé	Jedovaté
Jedlé	741	2
Jedovaté	12	432

Confusion matica pre jedlosť na SOM s PCA

	Jedlé	Jedovaté
Jedlé	735	18
Jedovaté	18	415

Confusion matica pre náleziská na neurónových sieťach s PCA

	Tráva	Lístie	Lúky	Cesty	Mestá	Lesy
Tráva	334	0	39	31	69	31
Lístie	0	18	0	0	3	0
Lúky	12	0	19	0	0	0
Cesty	24	0	0	22	0	6
Mestá	0	0	0	0	0	0
Lesy	54	0	0	61	1	463

Confusion matica pre náleziská na SOM s PCA

	Tráva	Lístie	Lúky	Cesty	Mestá	Lesy
Tráva	339	2	41	36	38	43
Lístie	0	16	0	0	0	0
Lúky	5	0	11	0	0	0
Cesty	44	0	0	57	0	29
Mestá	8	0	0	0	35	0
Lesy	27	0	7	20	0	429

Vyhodnotenie

Z ROC kriviek je ľahko vidno že PCA algoritmus dosahoval všeobecne lepšie výsledky ako ICA algoritmus pričom najväčší rozdiel bol pri rozhodovacích stromoch.

Pri klasifikácii jedlosti húb sa klasifikátori pretrénovali. Separovali trénovaciu množinu vždy na 100% ale mali problém správne určiť huby v testovacej množine. Ako najúspešnejší klasifikátor dopadli SOM, ktoré mali najlepšie a najkonzistentnejšie výsledky. Z ROC kriviek je vidno, že nárast od ostatných klasifikátorov SOM neboli náhodné naprečenie aj keď im zvyšovanie počtu dimenzií nepomáhalo výrazne zvýšiť úspešnosť klasifikácie.

Rozhodovacie stromy boli tiež pomerne úspešné hlavne pri nezredukovanej dimenzii, čo sme očakávali. Jedlosť húb s celou databázou dokázali rozhodnúť so skoro 100% úspešnosťou, čo potvrdzuje predpoklad, že by mohli byť dobré kôli pravidlám medzi ľuďmi, ktoré na základe niektorých znakov huby rozlišujú jej jedlosť/nejedlosť.

Neurónové siete mali problém s pretrénovaním. Trénovaciu množinu vždy klasifikovali na 100% v prípade jedlosti a 80% v prípade náleziska. Avšak tieto rozdelenia nesedeli s testovacou množinou húb. Zaujímavé je zisťovanie jedlosti kde pri 5 dimenziách dosiahli neurónové siete cez 95% úspešnosť, lebo neboli ovplyvnené preučeními.

Lineárny klasifikátor fungoval až nečakane dobre. S plnou databázou dosahoval pre jedlosť húb skoro 100% úspešnosť a teda boli dáta lineárne separabilné. Nálezisko bolo určované lineárnym klasifikátorom po častiach. Teda pri plnej databáze mal v priemere 90% úspešnosť na otázku, či huba rastie/nerastie v jednom prostredí.

Rozhodovací strom pre nálezisko hore plný dole orezaný

