Data:

Mushroom Database:
http://archive.ics.uci.edu/ml/datasets/Mushroom

Popis dat:

Tato datová sada obsahuje popis hypotetických příkladů hub ze skupiny Agaricus and Lepiota. Každá houba je klasifikována zda je jedovatá (poisonous), jedlá (edible) nebo nevíme a bylo žádoucí takovou houbu zařadit pro jistotu mezi jedovaté(poisonous).

Atributy:

Vstup – všechny atributy jsou nominální

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10. stalk-shape: enlarging=e,tapering=t
11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Výstup:

0 – class: edible=e, poisonous=p (třída je první písmenko v každém řádku)

Úkoly: Každému studentovi je přidělena celá datová sada

Předzpracování dat:

1) Vyberte si tři po sobě jdoucí atributy (například 1-3, nebo 20-22, apod.)
2) Zkonstruujte pro tyto atributy histogramy. Můžete na základě histogramů nalézt nějaké odlehlé hodnoty, nebo určit vztah atribut třída.
3) Atribut 11 stalk-root obsahuje chybějící hodnoty ?, navrhněte jak tyto hodnoty nahradit.

Strojové učení:

1) Vyberte dvě metody strojového učení a pokuste se klasifikovat houby na jedlé a jedovaté na základě vstupních hodnot.
2) Zkonstruujte křivku učení pro obě vybrané metody a jejich výsledky porovnejte.

3) Zkuste další metodu validace dat.
4) Zhodnoťte, jaká je pravděpodobnost, že se houbami otrávíte.

Asociační pravidla:

1) Existují v datech nějaká významná pravidla pro klasifikaci?
2) Jedno takové pravidlo slovně popište.