Data Mining Lab 1

Niek de Visscher 10667474

Deel 1:

1) De regel om te kijken of er getennist kan worden: (outlook = overcast)

De accuracy van deze regel is 100%

2) accuracy voor de test = 100% accuracy voor de cross-validation = 64.3%

3.1) Bij de ZeroR is de accuracy voor beide gevallen: 64.3%

3.2)Ik vind de ZeroR een redelijke baseline, omdat er hierbij meteen wordt gekeken naar de meest voorkomende class. Dit zorgt ervoor dat er hieruit ook meteen gevonden kan worden welke attributen de meeste invloed hebben op deze class. Ik kan persoonlijk geen andere baseline verzinnen.

3.3)Met behulp van cross-validation kan er gezorgd worden dat er gegeneraliseerd kan worden voor andere datasets. Door cross-validation kan dus worden gekeken hoe accuraat het model is voor de gegeven training en test data.

4) Het is nodig om te kijken naar andere gegevens dan de accuracy omdat hieruit afgeleid kan worden wat nu het verschil is tussen de verschillende methodes. Zo kan gezien worden uit de confusion matrix dat er een verschil is tussen de ZeroR 10 fold en J48 10 fold terwijl deze wel beiden dezelfde accuracy geven.

Deel 2:

1.1) De accuracy van de geteste set is: 85%

1.2)Er wordt eerst een onderscheid gemaakt of de nieuwe instantie veren heeft, zo ja dan is het een vogel, anders wordt gekeken naar het feit of het melk drinkt/produceert(?) zo ja dan is het een zoogdier, anders wordt gekeken naar de tanden. Als het tanden heeft en vinnen is het vis, als het tanden heeft en geen vinnen is het een reptiel als het minder of gelijk aan 2 poten heeft, heeft het meer dan 2 poten is het een amfibie. Als het geen tanden heeft is het een ongewervelde als het niet kan vliegen en een insect als het wel kan vliegen. Op deze manier wordt voor elk instantie gekeken onder welke class het valt.

1.3)Er is 1 intantie die verkeerd is geclassificeerd, namelijk een insect dat is geclassificeerd als ongewervelde. Dit betreft de termiet, deze wordt verkeerd geclassificeerd, omdat het een insect is dat niet vliegt. Hierdoor wordt het onder de ongewervelden gezet.

1.4) Hiervoor is de accuracy: 92.08%

1.5)

2.1) Voor de originele: 100% voor de aangepaste: 94.4%

2.2) Mijn data:

@relation weather

@attribute outlook {sunny, overcast, rainy}

```
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute favoritesport {hockey, tennis, golf}

@data
sunny,85,85,FALSE,golf
sunny,80,90,TRUE,hockey
overcast,83,86,FALSE,golf
rainy,70,96,FALSE,tennis
rainy,68,80,FALSE,tennis
rainy,65,70,TRUE,hockey
overcast,64,65,TRUE,hockey
sunny,72,95,FALSE,golf
sunny,69,70,FALSE,golf
rainy,75,80,FALSE,hockey
sunny,75,70,TRUE,hockey
overcast,72,90,TRUE,hockey
overcast,81,75,FALSE,golf
rainy,75,91,TRUE,hockey
overcast,61,81,TRUE,tennis
sunny,80,65,FALSE,golf
sunny,81,74,TRUE,hockey
overcast,61,91,TRUE,tennis
```

-De decision tree voor de dataset:

windy = TRUE

| temperature <= 61: tennis (2.0)

| temperature > 61: hockey (7.0)

windy = FALSE

| outlook = sunny: golf (4.0)

| outlook = overcast: golf (2.0)

| outlook = rainy: tennis (3.0/1.0)

De gevonden decision tree houd hier als eerste rekening met het feit of het waait of niet in tegenstelling tot de originele waar eerst werd gekeken naar de outlook. Verder wordt er ook rekening gehouden met de temperatuur waar in de originele helemaal geen rekening mee werd gehouden. Deze verschillen tussen de originele en deze decision tree zorgen ervoor dat er een gehele andere manier is voor het classificeren van instanties.

2.3) Er is geen verschil tussen de decision trees met of zonder pruning. De reden dat er geen verschil opgemerkt kan worden is dat er geen generalisaties gemaakt kunnen worden voor attributen omdat er voor elk attribuut duidelijk gezegd kan worden tot welke class het leid.