Gegevenswetenschap opdracht

Dataset: https://archive.ics.uci.edu/dataset/915/differentiated+schildklier+kanker+recidief

Aanleveren:

De volgende lijst helpt je bij het maken van een notitieblok met stappen die je neemt in datascience. Zoals vermeld in de les is je opdracht echter niet beperkt tot deze stappen of vragen, dus wees creatief. In je mondelinge presentatie moet je elke beslissing die je hebt genomen verantwoorden. Het is ook de gelegenheid voor de docent om je relevante vragen te stellen over je implementatie.

1. Laad de dataset.

2. Maak een beschrijvende analyse.

3. Bekijk de correlatiematrix.

4. Selecteer de variabelen die je gaat proberen te bestuderen.
Gebruik een correlatiematrix om te zien hoe de variabelen aan elkaar gerelateerd zijn en geef je mening (intuïtie) over waarom deze variabelen gecorreleerd zijn of niet.

5. Selecteer welke van de variabelen de afhankelijke en welke de onafhankelijke variabele wordt. Leg uit waarom.
Bijvoorbeeld: als je het risico op kanker bij patiënten wilt voorspellen op basis van de rest van de variabelen, dan zou je onafhankelijke variabele het risico zijn.
Opmerking: Alle studenten zullen mogelijk een andere kijk hebben op wat de voorspellende variabele zal zijn, aangezien het aan de student is om er zelf een te kiezen.

6. Maak voor visualisatiedoeleinden een scatterplot voor deze variabelen om te zien of hun patroon lineair is. Zo niet, ga dan terug naar stap 3.

7. Voer classificatie en regressie uit en probeer te begrijpen wat het verschil daartussen is.

- Voer feature engineering uit, d.w.z. construeer 3 extra features en maak een voorspellend model, neem kruisvalidatie op in je implementatie. Kies minimaal 3 variabelen om 1 binaire variabele te classificeren. Pas een model toe, gebruik kruisvalidatie en bereken de juiste metriek (gebruik ten minstens nauwkeurigheid, precisie en ROC-plot).
- Kies 4 variabelen om een regressieanalyse uit te voeren op een variabele. Zorg ervoor dat de verklarende variabelen in je gegevens een lineaire correlatie van minder dan 0,9 heeft.

8. Test of het model goed presteert of niet. Geef resultaten van metrieken, zoals nauwkeurigheid, precisie en f1-score weer.

9. Rapporteer. Is je model een goede voorspeller? Geeft je model extra inzicht in de afhankelijke variabelen die je eerder niet had? Als het antwoord op beide vragen nee is, ga dan terug naar stap 3.

10. Wat is, volgens de verkregen data in stap 1, de waarschijnlijkheid/voorspelling-van een recidief bij patiënten bij wie goed-gedifferentieerde schildklierkanker is vast-gesteld?

11. Op dit punt zou je deze vraag moeten kunnen beantwoorden: Hoe kan uw voor-spellend model(len) in het algemeen helpen bij deze data?

12. Kun je clusters van informatie maken?
    Hoeveel clusters heb je nodig?
    Waarom deze clusters?

13. Heeft het zin om een op dichtheid gebaseerde clustering uit te voeren? Waarom? Waarom niet? Als het zinvol is, moet je het implementeren en de resultaten vergelijken met de vorige Machine Learning-technieken.

Veel succes!