P

afrash ramjit

2151957

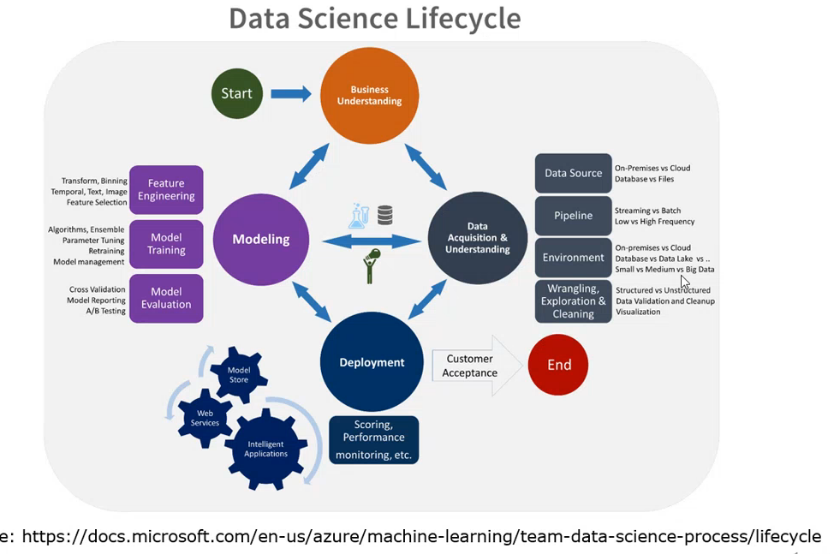
Portfolio assignment 102

Afrash Ramjit

2151957

# Inleiding

In dit document zie je hieronder op de afbeelding een aantal processen. Deze processen zal ik linken aan de assignments die ik heb uitgevoerd en dan zie je duidelijk wat de taken en processen zijn van een data scientist.



# Data acquisition & understanding

Data acquisition & understanding heb ik gedaan in opdracht 3 doordat ik hier een univeriate analyse heb gedaan op de penguin dataset.

Ik heb ook het process Data acquisition & understanding gedaan in opdracht 4 omdat ik hierin een dataset ging uitzoeken voor latere opdrachten. In opdracht 5 is dit proces gebruikt omdat ik hier een begin maakte aan de life expactency dataset. In opgave 6 is dit process toegepast op de dataset digimon en hierop is dan een univeriate analyses gedaan op 2 categorical kollommen en 2 numerieke kollommen. In opdracht 9 is een bivariate analyses gedaan op numerieke kollommen, hierbij is dan gekeken naar de correlaties. Bij opdracht 10 is een bivariate analyses gedaan op 2 kollommen op de digimon dataset. Bij opdracht 11 is een bivariate analyses gedaan op de penguin dataset. Dit is gedaan op de categorical data. In opdracht 12 is een bivariate analyses gedaan op 3 numerical kolommen. Bij opdracht 13 is een bivariate analyses gedaan op de penguin dataset en hier is dan ook een chi squared test gedaan om te kijken of er een kans is op correlatie is tussen 2 kollommen. In opdracht 14 is hetzelfde gedaan maar dan op de digimon dataset.

# Modeling

Dit proces is toegepast in opdracht 15. In opdracht 15 zijn er predictions gemaakt op de penguin dataset. Hier is een test set en een trainset gemaakt en is de accuracy berekent. Op het einde van deze opdracht is er een plot gemaakt om de decision tree visueel te weergeven.

Bij opdracht 16 is hetzelfde gedaan maar dan voor de digimon dataset, dus hier is ook het modelling proces toegepast, dit is gedaan op een categorical kolom.

Bij opdracht 17 is het proces modeling toegepast, dit is gedaan op numerical data van body\_mass\_g. hier is dan een decision tree gemaakt en word er geprobeerd om gegevens te voorspellen. De RMSE ( root mean square error, wat is de gemiddelde fout die we gaan zien.) wordt hier berekend voor de train data en voor de test data. De regression tree wordt hierna geplot om dit visueel te weergeven.

Bij opdracht 18 wordt een decision tree getrained op een numerieke kollom. Deze tree wordt geplot.

In opdracht 19 word een cluster model gemaakt op de penguin dataset, er wordt hier een algoritme toegepast om het cluster model te maken. En hier wordt dan een extra cluster kolom aangemaakt op deze dataset. Deze cluster wordt dan visueel weergeven en we vergelijken de cluster met de species kollom van de penguin.

# Business Understanding

Dit is gedaan bij opdracht 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 en 14. Hier hebben we meer kennis opgedaan over de penguin dataset, de digimon dataset en de life expactancy dataset. Bij het proces Data acquisition & understanding is hier verder op in gegaan en dit zie je dan ook terug onder dit kopje.

# Deployment

Is niet gedaan.