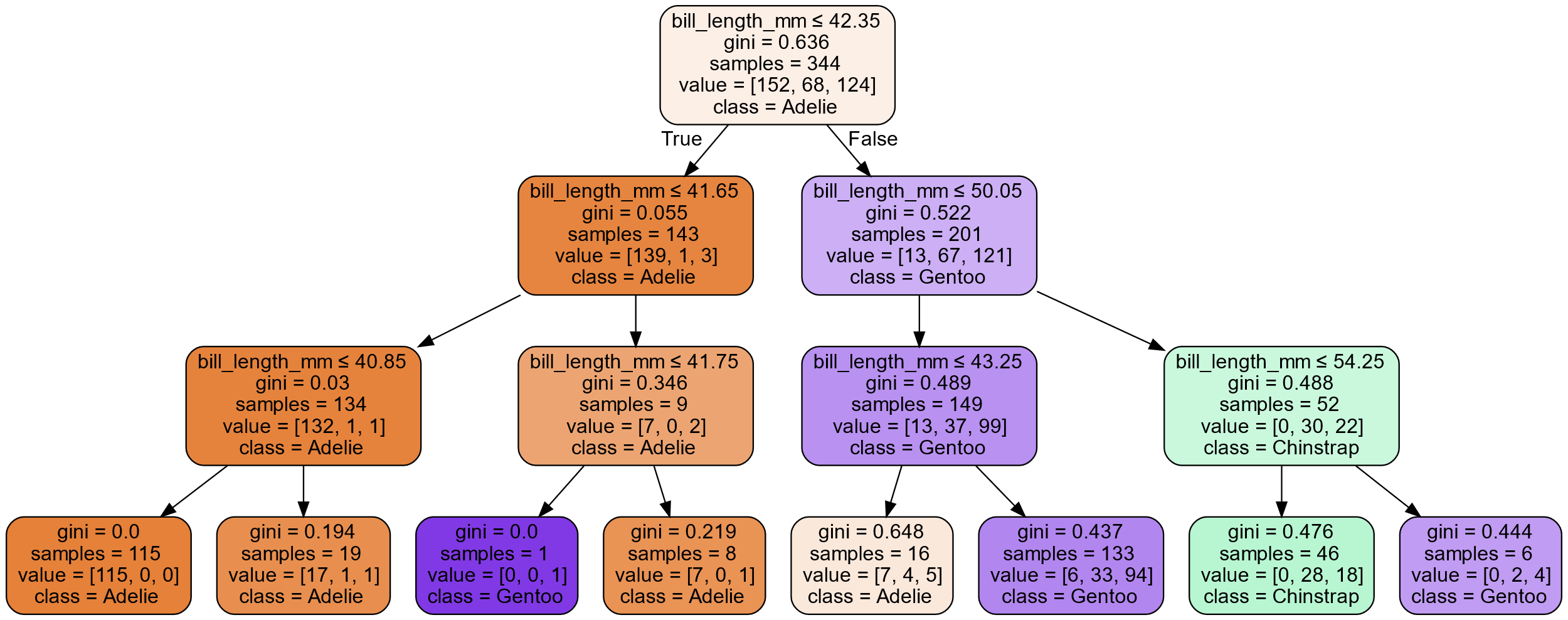
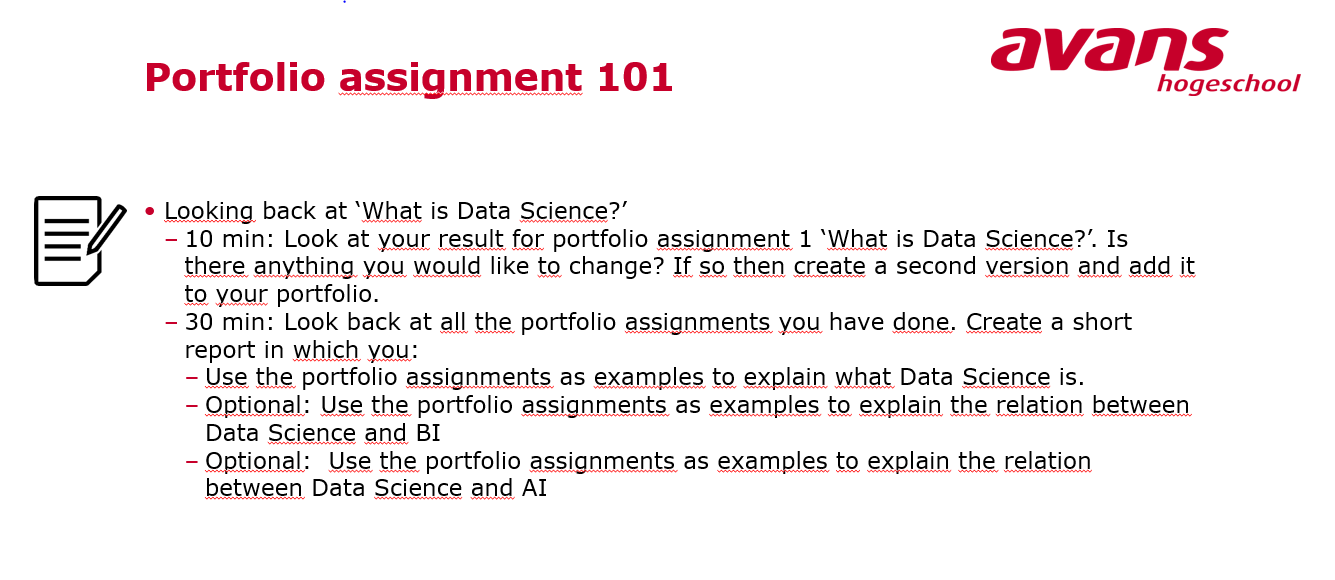
Bij Data Science is het belangrijk om op het begin je data goed weer te geven zodat je een indruk hebt waar het over gaat en niet blind met de data aan de slag gaat. Hieronder een voorbeeld:

Hier zie je de eerste 5 rijen uit de desbetreffende dataset penguins. Zo kun je snel zien welke kolommen er zijn en wat voor inhoud ze hebben. Doordat je dit weet kun je als je wilt meer gaan bekijken of al meteen analyses uitvoeren. Zoals je ook kunt zien is er in een diagram weergegeven hoeveel er van elke sekse aanwezig zijn in de dataset.

Nadat je een duidelijk beeld hebt wat er in je dataset zit kun je gaan kijken wat je wilt weten. Zo kun je gaan kijken of er correlations zijn tussen de verschillende kolommen of juist totaal niet. Dit kun je mooi weergeven in een tabel met kleurschema of in een scatterplot waar je in een oogopslag ziet tussen twee kolommen of ze veel met elkaar te maken hebben of niet.

Zoals je in deze afbeelding kunt zien kun je op verschillende manieren de resultaten weergeven van een snelle analyse op de data. Hier is gekeken of er mogelijke verbanden zijn tussen verschillende kolommen. In de tweede tabel kun je in een enkele oogopslag zien of ze wel of niet een correlation hebben door hoe verder ze van 0 afzitten en dus een donkerdere kleur krijgen. Rood voor een positief verband en blauw voor een negatief verband. Uit die tabel kan je ook de zien wat bijvoorbeeld de grootst positieve combinatie is van twee kolommen en die zou je dus ook in een scatterplot tegenover elkaar kunnen zetten om te kijken of er mogelijk nog uitschieters tussen zitten. Maar onder Data Science valt ook het feit dat je met bestaande data dingen kunt gaan voorspellen. Zo kun je met behulp van een algoritme je data laten analyseren zodat ervan geleerd kan worden. Wanneer dit is gedaan kan je met bijvoorbeeld een decisiontree duidelijk weergeven waar verwacht wordt wat de data in een bepaalde kolom wordt op basis van de date in de andere kolom(men) die je wilt gebruiken bij deze analyse. Bij de afbeelding hieronder wordt er gekeken naar welke Class er wordt verwacht dat je bent wanneer de bill\_length\_mm aan een bepaalde waarde voldoet.



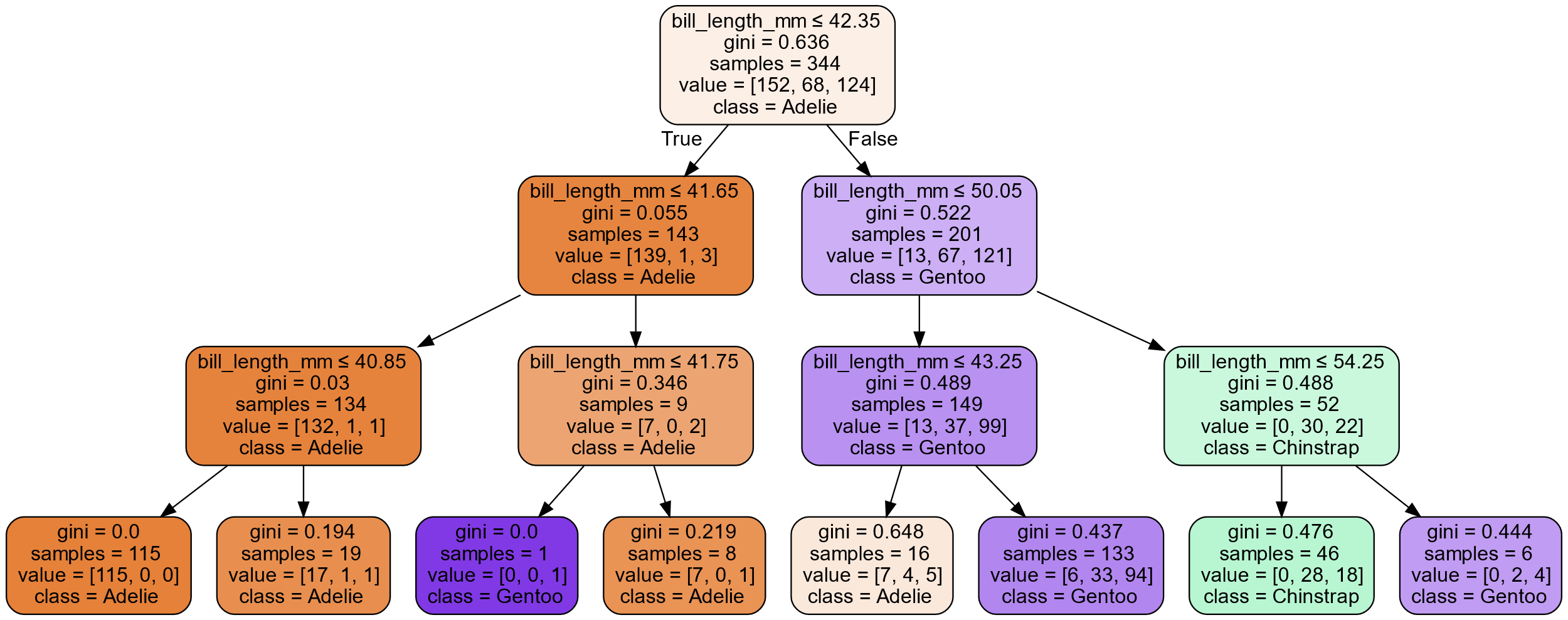
- Ik zou aan mijn eerste assignment niet echt iets aanpassen aangezien alles dat we hebben gedaan in deze introductie op Jupyter Notebook onder die kopjes valt.

Bij Data Science is het belangrijk om vragen te stellen zodat je weet wat je wilt gaan onderzoeken of analyseren. Zodra je dit weet kun je aan de slag om bijpassende data te gaan verzamelen van verschillende bronnen zoals webservers, logs, databases, API’s en online repositories.

Om ervoor te zorgen dat je fatsoenlijk met deze data aan de slag kunt moet deze data opgeschoond worden en mogelijk vervormd. Hierbij verwijder je de inconsistente datatypes, fout gespelde attributen, missende en dubbele waardes. Als dit is gebeurd kun je het met informatica tools verbouwen zodat je het beter kunt begrijpen.  Bij deze afbeelding zie je zulke data weergegeven in een tabel. Hierbij zou je er nog voor kunnen kiezen om de data die bestaat uit NaN eruit te halen omdat dit geen toegevoegde waarde heeft aan bepaalde analyses.

Wanneer de data zijn opgeruimd en mogelijk getransformeerd kun je de keuze gaan maken welke features je van de data wel in je model wilt en welke niet. Zo kun je voorkomen dat je uiteindelijke model niet klopt.

Wanneer je hebt gekozen voor een bepaald model kun je die gaan trainen. Als voorbeeld gaan we kijken naar decision trees. Dit model kan getraind worden met behulp van python. Hierbij gebruik je een test dataset.

Hierbij bepaal je op welke features je het model gaat trainen om te kijken wat de waarde van een specifiek attribuut is, in dit voorbeeld species. De feature die hierbij is gekozen is de bill\_length\_mm. Wanneer het getraind is kun je de decision tree ook weergeven waarbij je kan zien op welke keuze de uitkomst gebaseerd is.

De gegevens die je op deze manieren uit je dataset haalt kun je uiteindelijk gebruiken om bijvoorbeeld een dashboard te maken waarin je op een duidelijke en overzichtelijke manier deze gegevens weer te geven aan bijvoorbeeld de manager van een bedrijf die je de opdracht had gegeven. Hiermee moet je ze uiteindelijk overtuigen dat wat je gedaan hebt nuttig is voor hen en ze het daadwerkelijk kunnen gebruiken.

Wanneer het allemaal is goedgekeurd moet je ervoor zorgen dat alles getest is en goed werkt. Als dit is gebeurd en alles klopt kan het worden gedeployed zodat er met behulp van repo’s en dashboards real-time analyses kunnen worden uitgevoerd. Hierna hoef je alleen nog de deployment environment te onderhouden.