

Hoofdstuk 1

Classificatie

1.1 Wat is classificatie

Classificatie is een machine learning techniek die binnen het gebied van supervised learning valt. Bij deze methode worden gegevens in een dataset ingedeeld in verschillende klassen, het kan worden uitgevoerd op zowel gestructureerde als ongestructureerde gegevens. Het proces begint met het voorspellen van de klasse van bepaalde datapunten.. De klassen worden vaak aangeduid als target, label of categorie. Het ultieme doel van classificatie is om te bepalen in welke categorie de nieuwe gegevens zullen vallen.

Om het concept een beetje beter te kunnen begrijpen bekijken we een eenvoudige voorstelling van de toepassing in verband met tumoren. Hierbij willen we aan de hand van classificatie onderzoeken of een tumor goedaardig of kwaadaardig is. Dit is een voorbeeld van een binaire classificatie aangezien er slechts twee klassen kunnen zijn, namelijk goedaardig of kwaadaardig.

1.2 Terminologie

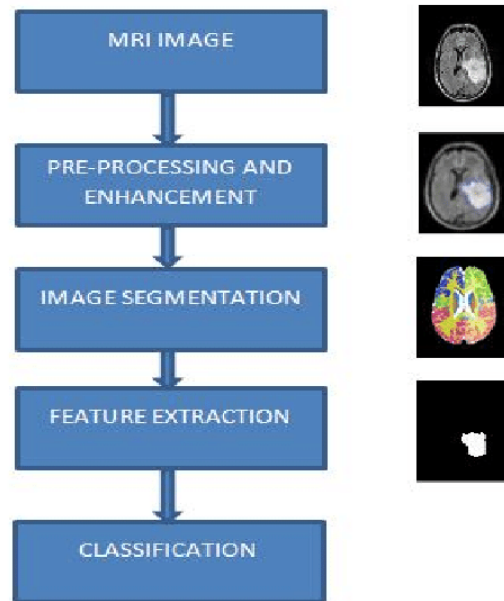
De **classifier** is een algoritme dat gebruikt wordt om de invoergegevens toe te wijzen aan een specifieke categorie. In ons geval heeft de classifier trainingsgegevens nodig om te begrijpen wat het verband is tussen de gegeven invoervariabelen en de klasse. Als de classifier nauwkeurig getraind is, kan deze worden gebruikt om na te gaan of de tumor goedaardig of kwaadaardig is.

Het **classificatiemodel** voorspelt de klasse of categorie van de gegevens aan de hand van de invoergegevens die tijdens de training zijn verstrekt.

Het **feature** of kenmerk is een meetbare eigenschap van iets dat je waarneemt. Bij tumor-classificatie kunnen bijvoorbeeld grootte, vorm, compositie en locatie van het beeld dienen als kenmerken.

Er bestaan ook classificaties met **meerdere klassen** of **meerdere labels**. Bij meerdere klassen wordt elk voorbeeld toegewezen aan één label, zoals bijvoorbeeld het identificeren van verschillende soorten dieren. Bij meerdere labels wordt elk voorbeeld toegewezen aan een reeks labels, zo kan een afbeelding meerdere objecten bevatten.

Overfitting treedt op wanneer een model te specifiek is voor de trainingsgegevens en slecht presteert op nieuwe, ongeziene gegevens. **Onderfitting** daarentegen treedt op wanneer het model te eenvoudig is en niet in staat is de complexiteit van de gegevens vast te leggen.



Figuur 1.1: Classificatie van een tumor. De eerste stap is de voorbewerking en verbetering van de MRI afbeelding. Fijnere details worden verbeterd en de ruis wordt uit het beeld gehaald. Bij image segmentation wordt het beeld opgedeeld in verschillende delen of segmenten om het beter te kunnen analyseren. Voor de features wordt rekening gehouden met bepaalde parameters, zoals grootte, vorm, compositie en locatie van het beeld. Volgens de resultaten die zijn verkregen uit de feature-extractie, wordt de classificatie van de tumor uitgevoerd.

1.3 Enkele toepassingen

- **Beeldherkenning:** Classificatie wordt vaak gebruikt in beeldherkenningstoepassingen, zoals het identificeren van objecten in foto's of video's.
- **Tekstclassificatie:** Hier wordt classificatie gebruikt om tekst te categoriseren, bijvoorbeeld het identificeren van spam-e-mails of het toewijzen van artikelen aan specifieke onderwerpen.
- **Medische diagnose:** Naast het eerder genoemde voorbeeld van tumorklassificatie, wordt classificatie ook toegepast op het diagnosticeren van verschillende medische aandoeningen op basis van verzamelde gegevens.