Hoofdstuk 1

Features

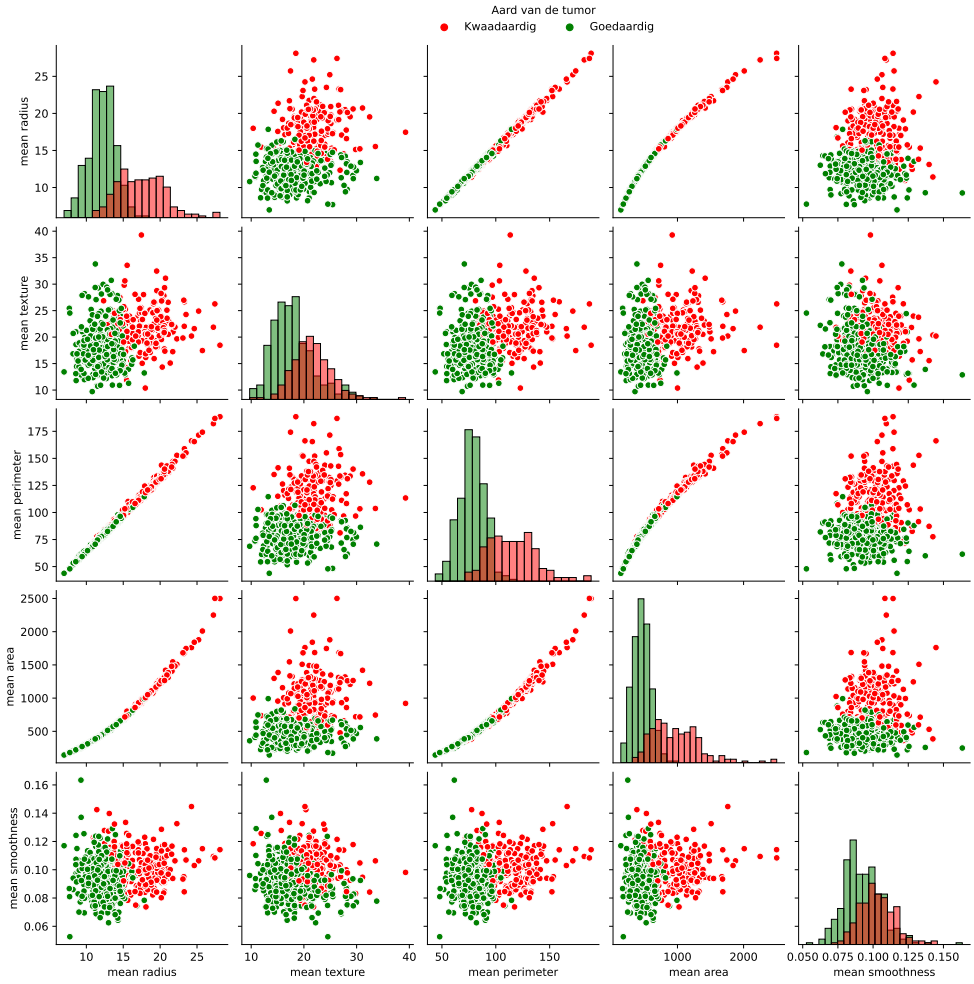
1.1 Overzicht

Zoals in de inleiding reeds vermeld werd, bestaat onze dataset uit een 500-tal datapunten met elk 30 features. Hieronder lijsten we de eerste features op om de lezer een idee te geven van de data die voor handen is:

- Straal (gemiddelde van afstanden van het middelpunt tot punten op de omtrek)
- Textuur (standaardafwijking van grijswaarden)
- Omtrek
- Oppervlakte
- Gladheid (lokale variatie in straallengtes)
- Compactheid ($omtrek^2/oppervlakte - 1.0$)
- Concaafheid (ernst van concave delen van de contour)
- Concave punten (aantal concave delen van de contour)
- Symmetrie
- ...

1.2 Limitaties

Zoals we verder zullen ontdekken in hoofdstuk **??**, botsen we tegen een limitatie wat betreft het aantal features die we kunnen gebruiken voor ons model. Indien we het aantal features opdrijven, worden de wiskundige berekeningen namelijk een pak complexer en zouden we dus meer tijd nodig hebben om dit helemaal uit te werken. Daarnaast wordt het visueel voorstellen van de dataset ook des te moeilijker bij een toenemend aantal features. Wegens deze twee limitaties, zullen we ons dus beperken tot slechts twee features per tumor.



Figuur 1.1: Een pairplot van de eerste 5 features in de tumordataset.

1.3 Selectie

Hoe selecteren we nu de twee interessante features? We kunnen daarvoor gebruik maken van een *pairplot*, zoals er een te zien is in figuur 1.1. Een *pairplot* is een visuele weergave van de relatie tussen paren van features in een dataset. In machine learning worden dit soort voorstellingen vaak gebruikt om patronen en verbanden tussen de verschillende features van een dataset te identificeren. Op basis van deze pairplot bepaalden wij dat *mean texture* (textuur) en *mean radius* (straal) twee interessante features waren om te bestuderen.