Data analyse rapport

Machine learning casus

School: HAN

Course: ADS

Auteur: Timoy van Balkom & Daan van Vugt

Datum: 2024-03-29

Inhoudsopgave

[Inleiding 3](#_Toc161418770)

[Descriptieve statistische kentallen 4](#_Toc161418771)

[Continuous features 4](#_Toc161418772)

[Categoriale features 4](#_Toc161418773)

[Histogrammen 5](#_Toc161418774)

[Cl.thickness 5](#_Toc161418775)

[Cell.size 7](#_Toc161418776)

[Cell.shape 9](#_Toc161418777)

[Marg.adhesion 11](#_Toc161418778)

[Epith.c.size 13](#_Toc161418779)

[Bare.nuclei 15](#_Toc161418780)

[Bl.cromatin 17](#_Toc161418781)

[Normal.nucleoli 19](#_Toc161418782)

[Mitoses 21](#_Toc161418783)

[Class 23](#_Toc161418784)

[Boxplots 24](#_Toc161418785)

[CL.thickness 24](#_Toc161418786)

[Cell.size 25](#_Toc161418787)

[Cell.shape 26](#_Toc161418788)

[Marg.adhesion 27](#_Toc161418789)

[Epith.c.size 28](#_Toc161418790)

[Bare.nuclei 29](#_Toc161418791)

[Bl.cromatin 30](#_Toc161418792)

[Normal.nucleoli 31](#_Toc161418793)

[Mitoses 32](#_Toc161418794)

[Tests 33](#_Toc161418795)

[Proportie test 33](#_Toc161418796)

[Manova test 33](#_Toc161418797)

[Scatter plots 36](#_Toc161418798)

[Pairplot 36](#_Toc161418799)

[Correlatiecoëfficiënt tabel 37](#_Toc161418800)

[Cell.size en Cell.shape 39](#_Toc161418801)

[Bare.nuclei en cell.shape 40](#_Toc161418802)

[Bare.nuclei en cell.size 41](#_Toc161418803)

[Conclusie 42](#_Toc161418804)

[Figurenlijst 43](#_Toc161418805)

[Bibliografie 44](#_Toc161418806)

# Inleiding

In dit document staat de analyse van de breastcancer dataset, aangeleverd door de HAN. In dit document wordt gekeken of de data in de dataset geschikt is voor het ontwikkelen van een machine learning-model voor het diagnosticeren van borstkanker. Hiervoor wordt gekeken naar alle features in de dataset, met visualisaties en testen om te bepalen of de data betrouwbaar en bruikbaar is en waar mogelijk punten zitten die gecorrigeerd moeten worden, waar data weg gehaald moet worden of waar data getransformeerd moet worden.

# Descriptieve statistische kentallen

Om te beginnen is het zaak om met gebruik van descriptieve statistische kentallen een beeld te schetsen van de data die we voor ons hebben.

Hieronder staan twee tabellen, de eerste tabel bevat alle continuous features en de tweede de categoriale feature, in de aangeleverd breastcancer\_csv dataset is slechts 1 categoriale feature, alle andere features zijn continuous.

De waardes in onderstaande tabellen zijn berekend met behulp van matplotlib.pyplot library in Python.

## Continuous features

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | % Miss. | Card. | Min. | 1ste Qrt. | Mean | Median | 3dr Qrt. | Max. | Std. Dev. |
| Cl.thickness | 699 | 0.0 | 10 | 1.0 | 2.0 | 4.4177 | 4.0 | 6.0 | 10.0 | 2.8157 |
| Cell.size | 699 | 0.0 | 10 | 1.0 | 1.0 | 3.1345 | 1.0 | 5.0 | 10.0 | 3.0515 |
| Cell.shape | 699 | 0.0 | 10 | 1.0 | 1,0 | 3.2074 | 1.0 | 5.0 | 10.0 | 2.9719 |
| Marg.adhesion | 699 | 0.0 | 10 | 1.0 | 1.0 | 2.8069 | 1.0 | 4.0 | 10.0 | 2.8554 |
| Epith.c.size | 699 | 0.0 | 10 | 1.0 | 2.0 | 3.2160 | 2.0 | 4.0 | 10.0 | 2.2143 |
| Bare.nuclei | 699 | 2.2890 | 10 | 1.0 | 1.0 | 3.5447 | 1.0 | 6.0 | 10.0 | 3.6439 |
| Bl.cromatin | 699 | 0.0 | 10 | 1.0 | 2.0 | 3.4278 | 3.0 | 5.0 | 10.0 | 2.4384 |
| Normal.nucleoli | 699 | 0.0 | 10 | 1.0 | 1.0 | 2.8670 | 1.0 | 4.0 | 10.0 | 3.0536 |
| Mitoses | 699 | 0.0 | 9 | 1.0 | 1.0 | 1.5894 | 1.0 | 1.0 | 10.0 | 1.7151 |

Als we naar bovenstaande tabel kijken zien we dat de dataset over het algemeen erg compleet is, enkel Bare.nuclei heeft een klein percentage data dat ontbreekt, wat dit betekent en hoe dit opgelost kan worden zal verder onderzocht moeten worden.

Wat nog meer opvalt is dat de spreiding van Mitoses heel klein is, met een 1ste en 3dr kwartiel van beide 1.0, dit betekent dat het overgrote deel van de waardes in Mitoses 1.0 zijn met uitschieters naar de 10, ook dit zal later onderzocht moeten worden hoe dit kan en of hier nog aanpassingen aan moeten worden gedaan.

Verder zien we dat alle waardes een spreiding hebben tussen de 1.0 en 10.0 en de algehele dataset lijkt realistisch te zijn, dit is een goed teken echter zal de data nog verder onderzocht moeten worden om ook te kijken waar eventueel correcties moeten worden toegepast.

## Categoriale features

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | % Miss. | Card. | Mode | Mode Freq. | Mode % | 2nd Mode | 2nd Mode Freq. | 2nd Mode % |
| Class | 699 | 0.0 | 2 | Benign | 458 | 65.5222 | Malignant | 241 | 34.4778 |

De categoriale feature in de dataset heet Class en is een binary dependent variabele, het is daarom logisch dat we in de tabel terugzien dat deze feature slechts uit 2 waardes kan bestaan. We zien dat het overgrote deel (65,5%) Benign als waarde heeft en slechts 34,5 % Malignant, dit betekent dat in 65,5% van de gevallen het gaat om een goedaardige tumor en in 34,5% van de gevallen het gaat om kanker.

# Histogrammen

Hieronder is voor iedere feature in de dataset een histogram en ondersteunend density plot gemaakt, omdat de dataset relatief weinig features bevat is het voor de compleetheid goed haalbaar om iedere feature te visualiseren.

De histogrammen zijn gemaakt met behulp van de matplotlib.pyplot library en Python en de density plots zijn gemaakt met behulp van de Seaborn library in Python.

## Cl.thickness

Als we naar onderstaande staafdiagram en density plot kijken lijkt het op een bimodale verdeling, hoewel het moeilijk te zien is zijn er wel 2 pieken, op 1 en op 5 met nog een kleinere piek op 10 maar die is bij lange na niet zo significant als bij de andere twee punten.

### Histogram

Afbeelding met diagram, schermopname, lijn, Perceel

Automatisch gegenereerde beschrijving

Figuur 1 Histogram Cl.thickness

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 2 Density plot Cl.thickness

## Cell.size

Als we naar onderstaande histogram en density plot kijken zien we het overgrote deel van de data op 1 liggen en significant minder data op 2 en hoger met een kleine piek op 10.

### Histogram

Afbeelding met tekst, schermopname, Perceel, lijn

Automatisch gegenereerde beschrijving

Figuur 3 Histogram Cell.size

### Density plot

Afbeelding met tekst, diagram, Perceel, ontwerp

Automatisch gegenereerde beschrijving

Figuur 4 Density plot cell.size

## Cell.shape

Als we naar onderstaande histogram en density plot kijken zien we eigenlijk een vergelijkbaar beeld als bij Cell.shape, veel waardes op 1 en significant minder bij 2 en hoger, hoewel hier wel meer data verdeling is ten opzichte van cell.size is het nog steeds minimaal. Dit zou kunnen wijze op een hoge correlatie tussen de twee features maar dat moet later onderzocht worden.

### Histogram

Afbeelding met tekst, schermopname, Perceel, diagram

Automatisch gegenereerde beschrijving

Figuur 5 Histogram Cell.shape

### Density plot

Afbeelding met diagram, Perceel, tekst, schermopname

Automatisch gegenereerde beschrijving

Figuur 6 Density plot Cell.shape

## Marg.adhesion

We zien in de histogram en density plot hieronder, net als bij cell.size en cell.shape, veel waardes op 1 en weinig vanuit 2 en hoger, ook dit zou kunnen betekenen dat er een goede correlatie is tussen marg.adhesion en cell.size en/of cell.shape maar dat zal wederom later onderzocht moeten worden, het kan namelijk ook toeval zijn en totaal geen correlatie betekenen.

### Histogram

Afbeelding met tekst, schermopname, Perceel, lijn

Automatisch gegenereerde beschrijving

Figuur 7 Histogram Marg.adhesion

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 8 Density plot Marg.adhesion

## Epith.c.size

Als we naar onderstaande histogram en density plot kijken zien we een kleine verschuiven ten opzichte van bovenstaande histogrammen, namelijk een groot deel van de waardes zijn 2. Wat wel hetzelfde is dat is dat wederom het overgrote deel van de data een waarde is en er niet veel variatie/verdeling is onder de datapunten.

### Histogram

Afbeelding met tekst, schermopname, Perceel, lijn

Automatisch gegenereerde beschrijving

Figuur 9 Histogram Epith.c.size

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 10 Density plot Epith.c.size

## Bare.nuclei

We zien in de histogram en density plot hieronder, net als bij cell.size en cell.shape, veel waardes op 1 en weinig vanuit 2 en hoger, met een piek op 10, dit zouden mogelijk outliers kunnen zijn die bij de boxplot naar voren kunnen komen.

### Histogram

Afbeelding met tekst, schermopname, lijn, Perceel

Automatisch gegenereerde beschrijving

Figuur 11 Histogram Bare.nuclei

### Density plot

Afbeelding met diagram, tekst, Perceel, ontwerp

Automatisch gegenereerde beschrijving

Figuur 12 Density plot Bare.nuclei

## Bl.cromatin

Als we kijken naar onderstaande histogram en density plot zien we veel data op de 1, 2 en 3 met een kleine piek op 7. Het overgrote deel van de data bevindt zich echter wel op de 1 t/m 3 waardes, potentieel dat hierdoor hogere waardes naar voren gaan komen als outliers maar dit moet bij de boxplots verder onderzocht worden.

### Histogram

Afbeelding met schermopname, diagram, lijn, Perceel

Automatisch gegenereerde beschrijving

Figuur 13 Histogram Bl.cromatin

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 14 Density plot Bl.cromatin

## Normal.nucleoli

We zien in de histogram en density plot hieronder, net als bij cell.size en cell.shape, veel waardes op 1 en weinig vanuit 2 en hoger, ook dit zou kunnen betekenen dat er een goede correlatie is tussen normal.nucleoli en cell.size en/of cell.shape maar dat zal wederom later onderzocht moeten worden, het kan namelijk ook toeval zijn en totaal geen correlatie betekenen.

### Histogram

Afbeelding met tekst, schermopname, Perceel, diagram

Automatisch gegenereerde beschrijving

Figuur 15 Histogram Normal.nucleoli

### Density plot

Afbeelding met tekst, diagram, Perceel, lijn

Automatisch gegenereerde beschrijving

Figuur 16 Density plot Normal.nucleoli

## Mitoses

We zien in de histogram en density plot hieronder, net als bij cell.size en cell.shape, veel waardes op 1 en weinig vanuit 2 en hoger, ook dit zou kunnen betekenen dat er een goede correlatie is tussen mitoses en cell.size en/of cell.shape maar dat zal wederom later onderzocht moeten worden, het kan namelijk ook toeval zijn en totaal geen correlatie betekenen.

### Histogram

Afbeelding met schermopname, lijn, tekst, Perceel

Automatisch gegenereerde beschrijving

Figuur 17 Histogram Mitoses

### Density plot

Afbeelding met tekst, diagram, Perceel, schermopname

Automatisch gegenereerde beschrijving

Figuur 18 Density plot Mitoses

## Class

In onderstaande histogram zien we net als in de kentallen tabel dat de class features uit 2 waardes kan bestaan, benign en malignant, waar benign het vaakst voorkomt, wederom dit betekent dat het om een goedaardige tumor gaat, waar malignant een kanker tumor betekent.

Voor deze histogram is geen ondersteunende density plot gemaakt omdat die niks zou toevoegen, we weten de verdeling omdat er slechts 2 waardes zijn.

### Histogram

Afbeelding met schermopname, Rechthoek, lijn, plein

Automatisch gegenereerde beschrijving

Figuur 19 Histogram Class

# Boxplots

Hieronder is voor iedere continuous feature in de dataset een boxplot gemaakt, hier is voor gekozen omdat het relatief weinig features zijn en daarom voor de compleetheid en volledigheid is het beter om alle features te analyseren in een boxplot zodat we niks over het hoofd zien.

De boxplots zijn gemaakt met behulp van de Seaborn library in Python.

## CL.thickness

Als we kijken naar de Cl.thickness zien we geen outliers, het gemiddelde ligt wel meer naar links van de meetstaaf wat betekent dat er een aantal waardes redelijk boven het gemiddelde liggen maar niet genoeg om outliers te zijn. Aan de hand van deze boxplot hebben we geen rede om aan te nemen dat we data moeten aanpassen/verwijderen voor deze feature.

Afbeelding met schermopname, diagram, Rechthoek, lijn

Automatisch gegenereerde beschrijving

Figuur 20 Boxplot Cl.thickness

## Cell.size

Als we kijken naar de Cell.size zien we geen outliers, het gemiddelde ligt echt aan de linker kant van de maatstaaf wat betekent dat er een aantal waardes redelijk boven het gemiddelde liggen maar niet genoeg om outliers te zijn. Opvallende dat de minimum, mediaan en 1ste kwartiel elkaar overlappen wat suggereert dat het merendeel van de waardes erg laag ligt rond de 1. Aan de hand van deze boxplot hebben we geen rede om aan te nemen dat we data moeten aanpassen/verwijderen voor deze feature.

Afbeelding met schermopname, Rechthoek, plein, diagram

Automatisch gegenereerde beschrijving

Figuur 21 Boxplot Cell.size

## Cell.shape

Als we kijken naar de Cell.shape zien we geen outliers, het gemiddelde ligt echt aan de linker kant van de maatstaaf wat betekent dat er een aantal waardes redelijk boven het gemiddelde liggen maar niet genoeg om outliers te zijn. Opvallende dat de minimum, mediaan en 1ste kwartiel elkaar overlappen wat suggereert dat het merendeel van de waardes erg laag ligt rond de 1. Aan de hand van deze boxplot hebben we geen rede om aan te nemen dat we data moeten aanpassen/verwijderen voor deze feature.

Afbeelding met Rechthoek, diagram, plein, schermopname

Automatisch gegenereerde beschrijving

Figuur 22 Boxplot Cell.shape

## Marg.adhesion

Als we kijken naar de boxplot van marg.adhesion zien we 2 outliers, op 9 en 10 respectievelijk. Dit komt neer op 60 records in de dataset waar de marg.adhesion 9 of 10 als waarde heeft. Het merendeel van deze regels (58) heeft als resultaat malignant, wat betekent dat het hier gaat om gediagnostiseerde kanker. Om deze rede vinden wij het verstandig om deze outliers mee te nemen in het model omdat het hier overduidelijk vaak om kanker gaat zou het onverstandig zijn om de outliers te verwijderen en daarmee een hoop positieve uitslagen weg te gooien.

Afbeelding met schermopname, diagram, Rechthoek, tekst

Automatisch gegenereerde beschrijving

Figuur 23 Boxplot Marg.adhesion

## Epith.c.size

Als we kijken naar onderstaande scatter plot zien we 3 outliers, op 8, 9 en 10 respectievelijk. Dit komt neer op 54 regels in de dataset waar de epith.c.size een waarde heeft die als outlier wordt beschouwd. We zien als we naar de data kijken wederom veel malignant resultaten (51), wat ook opvalt is dat, hoewel een aantal van de outliers van marg.adhesion aanwezig zijn in de records, er ook veel records zijn waar de waarde van marg.adhesion geen outlier is, dit duidt erop dat we met veel andere records in de dataset te maken hebben ten opzichte van de boxplot analyse van marg.adhesion. Om deze redenen hebben wij ervoor gekozen om ook deze data in de data set te laten zien, om dezelfde reden als bij de marg.adhesion outliers.

Afbeelding met diagram, tekst, lijn, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 24 Boxplot Epith.c.size

## Bare.nuclei

Als we kijken naar de Bare.nuclei zien we geen outliers, het gemiddelde ligt echt aan de linker kant van de maatstaaf wat betekent dat er een aantal waardes redelijk boven het gemiddelde liggen maar niet genoeg om outliers te zijn. Opvallende dat de minimum, mediaan en 1ste kwartiel elkaar overlappen wat suggereert dat het merendeel van de waardes erg laag ligt rond de 1. Aan de hand van deze boxplot hebben we geen rede om aan te nemen dat we data moeten aanpassen/verwijderen voor deze feature.

Afbeelding met schermopname, Rechthoek, plein, tekst

Automatisch gegenereerde beschrijving

Figuur 25 Boxplot Bare.nuclei

## Bl.cromatin

Als we kijken naar Bl.cromatin zien we 1 outlier, op 10. Dit komt neer op 20 records in de dataset waar Bl.cromatin een waarde van 10 heeft. Van al deze 20 records is het resultaat malignant, opvallend is ook dat het overgrote deel van de marg.adhesion waardes in deze records outliers zijn al kan niet hetzelfde gezegd worden voor de epith.c.size, waar ongeveer 40% outlier waardes zijn. Dit betekent dat er wel degelijk enige overlap is met de outliers van marg.adhesion in ieder geval. Desalniettemin hebben wij er toch voor gekozen om deze outliers ook in de dataset te houden, om dezelfde reden als eerder, dat het hier gaat om alleen malignant resultaten.

Afbeelding met diagram, lijn, Rechthoek, plein

Automatisch gegenereerde beschrijving

Figuur 26 Boxplot Bl.cromatin

## Normal.nucleoli

Als we kijken naar de boxplot van normal.nucleoli zien we 2 outliers, op 9 en 10 respectievelijk. Dit komt neer op 77 records in de dataset waar normal.nucleoli een waarde heeft van 9 of 10. Ook hier zien we weer dat het gaat om enkel malignant resultaten, op 1 na, en opvallend is ook dat we bij de andere features waar we outliers zagen niet extreem veel outliers zien in deze records, dit duidt erop dat het om veel andere records gaat dan bij de andere features. We hebben er wederom voor gekozen om geen outlier data te verwijderen om dezelfde rede als eerder, dat het hier wederom gaat om alleen malignant resultaten en mogelijk andere records dan bij de andere features.

Afbeelding met schermopname, Rechthoek, diagram, plein

Automatisch gegenereerde beschrijving

Figuur 27 Boxplot Normal.nucleoli

## Mitoses

De mitoses boxplot laat verreweg de meeste outliers zien, met 8 outliers op 2, 3, 4, 5, 6, 7, 8 en 10 respectievelijk. Dit komt neer op 120 records in de dataset waar mitoses een waarde heeft van 2 of hoger. Hoewel we hier iets meer benign resultaten zien is het overgrote deel (107) nog steeds malignant. Opvallend is wederom dat we geen overduidelijke aanwezigheid zien van outlier data bij de andere features, dit duidt erop dat er veel andere data tussen zit die niet bij de andere outliers naar boven kwam. Om deze reden hebben wij ervoor gekozen om wederom geen data te verwijderen, om wederom dezelfde reden als hierboven ook al een aantal keer vermeldt.

Afbeelding met tekst, diagram, lijn, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 28 Boxplot Mitoses

# Tests

## Proportie test

Voor de Proportie test hebben we gekeken naar de Class feature en wat de kans is op een positief (malignant) of negatief (benign) resultaat. Hiervoor is een alpha van 0.05 gebruikt en vanwege de omvang van de dataset is gekozen voor de normale verdeling.

* Benign: 0.6200 – 0.6905
* Malignant: 0.3095 – 0.3800

Dit betekent dat we met 95% zekerheid kunnen zeggen dat de kans op een benign resultaat tussen de 62% en 69.05% ligt en de kans op een malignant resultaat tussen de 30.95% en 38% ligt. We kunnen dus met hoge mate van zekerheid zeggen waar de werkelijke proportie waarschijnlijk ligt.

## Manova test

Om te kijken of er statistische verschillen zijn tussen de verschillende features aan de hand van de waarde van de Class feature, hebben wij gekozen om een Manova test uit te voeren, zodat we alle continuous features kunnen vergelijken met de Class feature.

Hieronder staan de resultaten van de Manova test, allereerst zien we de intercept resultaat tabel, dit is de baseline waar alles is gemeten aan de hand van de waarde 0 of Benign in deze test. De tweede tabel bevat de resultaten van het vergelijken van beide waardes, 0 en 1 of Benign en Malignant in het geval van deze test.

Aan de hand van de extreem lage P waarde kunnen we concluderen dat er een significant verschil is tussen de groepen. Deze verschillen kunnen mogelijk bruikbaar zijn voor het onderscheiden van maligne en benign tumoren, maar verdere analyse zal nodig zijn om de voorspellende waarde van deze verschillen te evalueren. Dit wordt hieronder gedaan aan de hand van een post-hoc test.

### Nulhypothese

* H0 = μ1 = μ2
* HA = μ1 ≠ μ2

### Intercept resultaat

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | Value | Num DF | Den DF | F Value | Pr > F |
| Wilks’ lambda | 0.2257 | 9.0000 | 673.000 | 256.4890 | 0.0000 |
| Pillai’s trace | 0.7743 | 9.0000 | 673.000 | 256.4890 | 0.0000 |
| Hotelling-Lawley trace | 3.4300 | 9.0000 | 673.000 | 256.4890 | 0.0000 |
| Roy’s greatest root | 3.4300 | 9.0000 | 673.000 | 256.4890 | 0.0000 |

### Class resultaat

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class | Value | Num DF | Den DF | F Value | Pr > F |
| Wilks’ lambda | 0.1567 | 9.0000 | 673.000 | 402.4991 | 0.0000 |
| Pillai’s trace | 0.8433 | 9.0000 | 673.000 | 402.4991 | 0.0000 |
| Hotelling-Lawley trace | 5.3826 | 9.0000 | 673.000 | 402.4991 | 0.0000 |
| Roy’s greatest root | 5.3826 | 9.0000 | 673.000 | 402.4991 | 0.0000 |

### Post-hoc test

Voor de post-hoc test is gekozen om een Tukey’s HSD post-hoc test uit te voeren waar iedere continuous feature wordt vergeleken met de Class feature.

Als we dan naar onderstaande resultaten kijken zien we een hoop waardes maar waar we vooral naar opzoek zijn is of we de nulhypothese kunnen verwerpen, dit wordt aangeduid met de kolom ‘Reject’, als deze True is kunnen we de nulhypothese verwerpen en als deze False is niet. We zien dat bij alle features op 1 na we de nulhypothese kunnen verwerpen, dit betekent dat we voor al deze features kunnen concluderen dat er een significant verschil is afhankelijk van de waarde van Class (benign of malignant) wat suggereert dat deze feature potentieel waardevolle voorspellers zijn voor het diagnosticeren van borstkanker.

Als we dan kijken naar de enige feature waar we de nulhypothese niet kunnen verwerpen, Bare.nuclei, dan zien we hier ook dat er geen waardes zijn voor de meandiff, p-adj, lower en upper waarde en we zien ook dat residual een lagere Df waarde heeft dan bij de andere features. Dit klopt ook omdat we eerder bij de kentallen al zagen dat er een klein procent ontbrekende waardes was voor deze feature en het lijkt erop dat deze ontbrekende waardes de test inaccuraat maken voor deze feature.

Als we de records uit de dataset selecteren waar bare.nuclei NaN als waarde heeft zien we voornamelijk benign als resultaat en veel lage waardes voor de overige features met een paar uitschieters richting de 10 maar het overgrote deel lijkt onder de 4 te zitten qua waarde. Het gaat in totaal om 16 records, gezien de waarde van de meeste records relatief laag is en het resultaat voor 14/16 benign is zou de overweging ook gemaakt kunnen worden om de records waar bare.nuclei geen waarde heeft uit de dataset te verwijderen, het lijkt erop dat dit geen significante negatieve invloed zou hebben op de kwaliteit van het model waar zelf data toevoegen dat mogelijk wel zou hebben wanneer dit niet goed gedaan wordt.

#### Cl.thickness

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 2837.0594 | 2837.0594 | 733.2070 | 6.8356 |
| Residual | 697.0 | 2696.9607 | 3.8694 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 4.2387 | 0.0 | 3.9313 | 4.546 | True |

#### Cell.size

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 4347.8536 | 4347.8536 | 1408.5272 | 1.7479 |
| Residual | 697.0 | 2151.5055 | 3.0868 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 5.2473 | 0.0 | 4.9728 | 5.5218 | True |

#### Cell.shape

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 4134.5197 | 4134.5197 | 1419.3055 | 2.9456 |
| Residual | 697.0 | 2030.4016 | 2.9131 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 5.1169 | 0.0 | 4.8503 | 5.3836 | True |

#### Marg.adhesion

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 2763.1188 | 2763.1188 | 657.7937 | 1.1098 |
| Residual | 697.0 | 2927.8802 | 4.2006 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 4.1831 | 0.0 | 3.8629 | 4.5033 | True |

#### Epith.c.size

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 1595.4957 | 1595.4957 | 608.7196 | 4.3462 |
| Residual | 697.0 | 1826.8848 | 2.6211 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 3.1787 | 0.0 | 2.9257 | 3.4316 | True |

#### Bare.nuclei

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 6128.9447 | 6128.9447 | 1426.2403 | 3.4011 |
| Residual | 681.0 | 2926.4433 | 4.2973 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | NaN | NaN | NaN | NaN | False |

#### Bl.cromatin

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 2375.7667 | 2375.7667 | 933.2873 | 9.8491 |
| Residual | 697.0 | 1774.2762 | 2.5456 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 3.8788 | 0.0 | 3.6295 | 4.1281 | True |

#### Normal.nucleoli

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 3301.7676 | 3301.7676 | 717.6280 | 3.1242 |
| Residual | 697.0 | 3206.8591 | 4.6009 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 4.5727 | 0.0 | 4.2375 | 4.9078 | True |

#### Mitoses

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabele | Df | Sum\_sq | Mean\_sq | F | PR(>F) |
| Class | 1.0 | 367.6660 | 367.6660 | 152.0402 | 9.6821 |
| Residual | 697.0 | 1685.4957 | 2.4182 | NaN | NaN |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Groep 1 | Groep 2 | Meandiff | P-adj | Lower | Upper | Reject |
| 0 | 1 | 1.5259 | 0.0 | 1.2829 | 1.7689 | True |

# Scatter plots

Hieronder is allereerst een pairplot gemaakt om alle feature-to-feature scatter plots te visualiseren, wij hebben hier vervolgens een aantal interessante relaties uitgepakt waar we op inzoomen om verder te kijken naar de correlatie tussen de beide features, deze features zijn uitgekozen aan de hand van hun correlatiecoëfficiënt met de Class feature.

## Pairplot

Afbeelding met tekst, diagram, schermopname, Parallel

Automatisch gegenereerde beschrijving

Figuur 29 Pairplot

## Correlatiecoëfficiënt tabel

In de onderstaande tabel staan de correlatiecoëfficiënten van het bovenstaande pair plot. Alle rood gemarkeerde kolommen zijn kolommen waar een feature wordt vergeleken met zichzelf en hebben daarbij altijd een correlatiecoëfficiënt van 1. Als we kijken naar de gehele correlatiecoëfficiënten tabel zien we veel sterke en een aantal zeer sterke positieve correlaties tussen de features, dit laat zien dat veel features elkaar matig tot sterk beïnvloeden.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cl.thickness | Cell.size | Cell.shape | Marg.adhesion | Eptith.c.size | Bare.nuclei | Bl.cromatin | Normal.nucleoli | Mitoses | Class |
| Cl.thickness | 1 | 0.6449 | 0.6546 | 0.4864 | 0.5218 | 0.5931 | 0.5584 | 0.5358 | 0.3500 | 0.7160 |
| Cell.size | 0.6449 | 1 | 0.9069 | 0.7056 | 0.7518 | 0.6917 | 0.7557 | 0.7229 | 0.4587 | 0.8179 |
| Cell.shape | 0.6546 | 0.9069 | 1 | 0.6831 | 0.7197 | 0.7139 | 0.7359 | 0.7194 | 0.4389 | 0.8189 |
| Marg.adhesion | 0.4864 | 0.7056 | 0.6831 | 1 | 0.5996 | 0.6706 | 0.6667 | 0.6034 | 0.4176 | 0.6962 |
| Epith.c.size | 0.5218 | 0.7518 | 0.7197 | 0.5996 | 1 | 0.5857 | 0.6161 | 0.6289 | 0.4791 | 0.6828 |
| Bare.nuclei | 0.5931 | 0.6917 | 0.7139 | 0.6706 | 0.5857 | 1 | 0.6806 | 0.5483 | 0.3392 | 0.8227 |
| Bl.cromatin | 0.5584 | 0.7557 | 0.7359 | 0.6667 | 0.6161 | 0.6807 | 1 | 0.6659 | 0.3442 | 0.7566 |
| Normal.nucleoli | 0.5358 | 0.7229 | 0.7194 | 0.6034 | 0.6289 | 0.5843 | 0.6659 | 1 | 0.4283 | 0.7122 |
| Mitoses | 0.3500 | 0.4587 | 0.4389 | 0.4176 | 0.4791 | 0.3392 | 0.3442 | 0.4283 | 1 | 0.4232 |
| Class | 0.7160 | 0.8179 | 0.8189 | 0.6962 | 0.6828 | 0.8227 | 0.7566 | 0.7122 | 0.4232 | 1 |

### Class en alle features

Allereerst kijken we naar alle features vergeleken met Class, hiervoor is benign omgezet tot 0 en malignant tot 1. Als we dan kijken naar de verschillende correlatiecoëfficiënt hierboven zien we voor alle, op mitoses na, features dat er een positieve correlatie is van meer dan 0.5 en ook voor mitoses zit de waarde niet ver onder de 0.5. Dit betekent dat alle features een sterke tot zeer sterke positieve correlatie hebben met Class en dat wanneer de waarde van een feature stijgt dit ook lijkt te zorgen voor een vergrootte kans om een malignant resultaat.

Het is belangrijk te onthouden dat correlatie geen oorzaak-gevolg relatie impliceert, het geeft alleen aan dat er een relatie is tussen de variabele maar het vertelt ons niet het volledige verhaal. Het kan namelijk zijn dat een of beide variabele door een andere factor wordt beïnvloed, of dat er helemaal geen causale relatie is. Maar zo op het eerste oogopslag lijkt het er wel op dat er een sterke correlatie is tussen de features en het resultaat in ieder geval.

We kunnen hier in elk geval wel uit oordelen dat de data aan de hand van de correlatiecoëfficiënt tabel geschikt is om borstkanker te voorspellen maar er zijn andere meet manieren nodig om hier een definitief oordeel over te kunnen vellen.

Hieronder zoomen we nog verder in op de 3 features die de hoogste correlatiecoëfficiënt hebben met Class om te kijken hoe deze onderling met elkaar correleren en zo hopelijk een beter beeld te krijgen van de algehele correlatie.

## Cell.size en Cell.shape

Als we naar de scatter plot kijken van cell.size en cell.shape dan zien we een duidelijke correlatie tussen de twee features, een groot deel van de data is gecentreerd rond de diagonale rij met weinig data punten die hiervan afwijken. Als we ook kijken naar de correlatiecoëfficiënt zien we een waarde van 0.9069, deze waarde ligt erg dicht tegen de 1 en laat een duidelijk sterke positieve correlatie zien tussen cell.size en cell.shape.

Dit betekent dat het effect van de cell.size groot is op de cell.shape en visa versa, het lijkt erop dat de cell.shape meegroeit met de cell.size, althans aan de hand van deze scatter plot en correlatiecoëfficiënt.

Afbeelding met schermopname, tekst, lijn, diagram

Automatisch gegenereerde beschrijving

Figuur 30 Scatter plot cell.size en cell.shape

## Bare.nuclei en cell.shape

Als we kijken naar onderstaande scatter plot zien we een spreiding van de datapunten maar we zien ook een hoop datapunten op of rond de diagonale lijn, met een aantal datapunten die daarbuiten vallen. Als we kijken naar de correlatiecoëfficiënt dan zien we een waarde van 0.7359, dit duidt op een redelijke sterke positieve correlatie tussen de beide features.

Dit betekent dat het effect van de bare.nuclei groot is op de cell.shape en visa versa, het lijkt erop dat de cell.shape meegroeit met de bare.nuclei, althans aan de hand van deze scatter plot en correlatiecoëfficiënt.

Afbeelding met schermopname, tekst, diagram, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 31 Scatter plot bare.nuclei en cell.shape

## Bare.nuclei en cell.size

Als we kijken naar onderstaande scatter plot zien we eigenlijk een soort gelijk beeld als bij de feature-to-feature vergelijking van bare.nuclei en cell.shape hierboven, namelijk een spreiding van datapunten maar ook veel rond de diagonale lijn. Ook de correlatiecoëfficiënt van 0.7557 laat zien dat er een redelijk sterke positieve correlatie is tussen de twee features.

Dit betekent dat het effect van de cell.size groot is op de bare.nuclei en visa versa, het lijkt erop dat de bare.nuclei meegroeit met de cell.size, althans aan de hand van deze scatter plot en correlatiecoëfficiënt.

Afbeelding met schermopname, tekst, diagram, lijn

Automatisch gegenereerde beschrijving

Figuur 32 Scatter plot bare.nuclei en cell.size

# Conclusie

Een belangrijk conclusie die we kunnen trekken aan de hand van de outliers in de boxplots is dat de outliers niet per se slecht zijn, sterker nog de outliers lijken juist een vergrootte kans op malignant aan te geven omdat het overgrote deel van de outliers malignant als resultaat heeft. Dit zou kunnen betekenen dat het wel degelijk outliers zijn maar het juist outliers zijn omdat er sprake is van kanker en deze waardes anders niet zouden voorkomen. Om die rede is het advies om geen outliers te verwijderen omdat het verwijderen ervan de kwaliteit van het uiteindelijke model kan comprimeren en niks wijst erop dat het behouden van de outliers hetzelfde zou doen, in tegendeel het behouden van de outliers lijkt een noodzakelijkheid om een betrouwbaar model te creëren.

Een andere belangrijke conclusie die we kunnen trekken is dat voor de records waar bare.nuclei geen waarde heeft het verstandig is om deze records uit de dataset te verwijderen en niet te gebruiken in het model. Zoals uitgebreid toegelicht in het hoofdstuk [Post-hoc test](#_Post-hoc_test) zal het verwijderen van de records geen significante negatieve impact hebben op de kwaliteit van het model maar het behouden van de records en daarmee het zelf invullen van de waardes voor bare.nuclei kan dit wel hebben, zeker als het niet goed gedaan wordt. Hierdoor geldt het advies om het risico niet te lopen en de records uit het model te laten.

Verder zien we dat er een correlatie lijkt te zijn tussen een hogere waarde van de features en een vergrote kans op een malignant resultaat, dit is vooral te zien in de scatter plots en correlatiecoëfficiënt maar komt ook terug bij de boxplot analyse wanneer we naar de data van de outliers gaan kijken, deze twee verschillende visualisaties samen maken het argument sterker dat hogere waardes voor de features de kans op malignant resultaat vergroot.

Om dan terug te komen op de informatievraag of de data in de dataset geschikt is voor het ontwikkelen van een machine learning-model voor het diagnosticeren van borstkanker. Wij concluderen aan de hand van deze data-analyse dat deze dataset geschikt is om te gebruikt te worden voor deze doeleinde. Daarbij hebben wij wel als duidelijke vereiste dat de records waar bare.nuclei geen waarde heeft uit de dataset verwijderd worden en niet in het model gebruikt zullen worden, de overige data kan zo gelaten worden en gebruikt worden in het model.

# Figurenlijst

[Figuur 1 Histogram Cl.thickness 5](#_Toc161418807)

[Figuur 2 Density plot Cl.thickness 6](#_Toc161418808)

[Figuur 3 Histogram Cell.size 7](#_Toc161418809)

[Figuur 4 Density plot cell.size 8](#_Toc161418810)

[Figuur 5 Histogram Cell.shape 9](#_Toc161418811)

[Figuur 6 Density plot Cell.shape 10](#_Toc161418812)

[Figuur 7 Histogram Marg.adhesion 11](#_Toc161418813)

[Figuur 8 Density plot Marg.adhesion 12](#_Toc161418814)

[Figuur 9 Histogram Epith.c.size 13](#_Toc161418815)

[Figuur 10 Density plot Epith.c.size 14](#_Toc161418816)

[Figuur 11 Histogram Bare.nuclei 15](#_Toc161418817)

[Figuur 12 Density plot Bare.nuclei 16](#_Toc161418818)

[Figuur 13 Histogram Bl.cromatin 17](#_Toc161418819)

[Figuur 14 Density plot Bl.cromatin 18](#_Toc161418820)

[Figuur 15 Histogram Normal.nucleoli 19](#_Toc161418821)

[Figuur 16 Density plot Normal.nucleoli 20](#_Toc161418822)

[Figuur 17 Histogram Mitoses 21](#_Toc161418823)

[Figuur 18 Density plot Mitoses 22](#_Toc161418824)

[Figuur 19 Histogram Class 23](#_Toc161418825)

[Figuur 20 Boxplot Cl.thickness 24](#_Toc161418826)

[Figuur 21 Boxplot Cell.size 25](#_Toc161418827)

[Figuur 22 Boxplot Cell.shape 26](#_Toc161418828)

[Figuur 23 Boxplot Marg.adhesion 27](#_Toc161418829)

[Figuur 24 Boxplot Epith.c.size 28](#_Toc161418830)

[Figuur 25 Boxplot Bare.nuclei 29](#_Toc161418831)

[Figuur 26 Boxplot Bl.cromatin 30](#_Toc161418832)

[Figuur 27 Boxplot Normal.nucleoli 31](#_Toc161418833)

[Figuur 28 Boxplot Mitoses 32](#_Toc161418834)

[Figuur 29 Pairplot 36](#_Toc161418835)

[Figuur 30 Scatter plot cell.size en cell.shape 39](#_Toc161418836)

[Figuur 31 Scatter plot bare.nuclei en cell.shape 40](#_Toc161418837)

[Figuur 32 Scatter plot bare.nuclei en cell.size 41](#_Toc161418838)

# Bibliografie

HAN. (sd). *ADS ML Nakijkmodel 2023-2024.xlsx.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/wqOJZQbD/eyJpdiI6Ik1URFMzSnRVbkt1dHd4R0dWb0hiMWc9PSIsInZhbHVlIjoiV0J2S2t1YWRVbXpPUnBrZmQyMFZEKzhiV25xa3RHSElrN1JmaDRwZkwrWVI0MWFCNmIwdSswcktuTVJzcVpkQiIsIm1hYyI6IjFjNGYwMzZlMjhhMDJhNWM5ZTU5MDk4NDI0MjdmNT

HAN. (sd). *Analyse van biopten data versie vj 2024.docx.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/wqOJZQbD/eyJpdiI6Ik9pcFEvNXFwTnpSQ3V6Ukt4UlN4R3c9PSIsInZhbHVlIjoiNjlhd3JUVzdDTzdtTWZPYXd4R0g2eTR1enhmMGpQS1BIZEM0QkNGekt2cEg4NC96eDhhOGprWW02YUpDSXc1WDY5SnZiY0picDd3QitpaDc0UVJQT2c9PSIsIm1hYyI6ImNmY2JiYT

HAN. (sd). *breastcancer\_csv.* Opgehaald van Onderwijsonline: https://han.onderwijsonline.nl/elearning/lessonfile/wqOJZQbD/eyJpdiI6ImNVQm9vQUI4eWNnM0RoMnEwOVZ0OXc9PSIsInZhbHVlIjoiTGxxSFhGM0l6UUxKZ0N4UzloZkQySEV0cGVGbzZOSTVrYUdNeWpaaDdHOD0iLCJtYWMiOiIyNjMzYzBiOThhYzYwOGY4N2VkYmE4YzUwODU5M2ZmNDZhZTE3OTRlMTQxNDU5ODBkOD

John D. Kelleher, B. M. (2015). *Fundementals of machine learning for predictive data analytics.* The MIT Press.

Statsmodels. (sd). *statsmodels.multivariate.manova.MANOVA.from\_formula*. Opgehaald van Statsmodels: https://www.statsmodels.org/stable/generated/statsmodels.multivariate.manova.MANOVA.from\_formula.html

Statsmodels. (sd). *statsmodels.stats.multicomp.pairwise\_tukeyhsd*. Opgehaald van Statsmodels: https://www.statsmodels.org/stable/generated/statsmodels.stats.multicomp.pairwise\_tukeyhsd.html