Data analyse rapport

Machine learning casus

School: HAN

Course: ADS

Auteur: Timoy van Balkom & Daan van Vugt

Datum: 2024-03-29

Docent:

Inhoudsopgave

[Inleiding 2](#_Toc161315137)

[Descriptieve statistische kentallen 3](#_Toc161315138)

[Continuous features 3](#_Toc161315139)

[Categoriale features 3](#_Toc161315140)

[Histogrammen 4](#_Toc161315141)

[Cl.thickness 4](#_Toc161315142)

[Cell.size 6](#_Toc161315143)

[Cell.shape 7](#_Toc161315144)

[Marg.adhesion 8](#_Toc161315145)

[Epith.c.size 9](#_Toc161315146)

[Bare.nuclei 10](#_Toc161315147)

[Bl.cromatin 11](#_Toc161315148)

[Normal.nucleoli 12](#_Toc161315149)

[Mitoses 13](#_Toc161315150)

[Class 14](#_Toc161315151)

[Boxplots 15](#_Toc161315152)

[CL.thickness 15](#_Toc161315153)

[Cell.size 16](#_Toc161315154)

[Cell.shape 17](#_Toc161315155)

[Marg.adhesion 18](#_Toc161315156)

[Epith.c.size 19](#_Toc161315157)

[Bare.nuclei 20](#_Toc161315158)

[Bl.cromatin 21](#_Toc161315159)

[Normal.nucleoli 22](#_Toc161315160)

[Mitoses 23](#_Toc161315161)

[Tests 24](#_Toc161315162)

[Proportie test 24](#_Toc161315163)

[Manova test 24](#_Toc161315164)

[Scatter plots 27](#_Toc161315165)

[Pairplot 27](#_Toc161315166)

[Correlatie coëfficiënt tabel 28](#_Toc161315167)

[Conclusie 30](#_Toc161315168)

[Figurenlijst 31](#_Toc161315169)

[Bibliografie 32](#_Toc161315170)

# 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, middels 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 terug zien 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

### 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

### 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

### 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

### 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

### 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

### 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

### 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

### 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

### 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

Afbeelding met schermopname, diagram, Rechthoek, lijn

Automatisch gegenereerde beschrijving

Figuur 20 Boxplot Cl.thickness

## Cell.size

Afbeelding met schermopname, Rechthoek, plein, diagram

Automatisch gegenereerde beschrijving

Figuur 21 Boxplot Cell.size

## Cell.shape

Afbeelding met Rechthoek, diagram, plein, schermopname

Automatisch gegenereerde beschrijving

Figuur 22 Boxplot Cell.shape

## Marg.adhesion

Afbeelding met schermopname, diagram, Rechthoek, tekst

Automatisch gegenereerde beschrijving

Figuur 23 Boxplot Marg.adhesion

## Epith.c.size

Afbeelding met diagram, tekst, lijn, Rechthoek

Automatisch gegenereerde beschrijving

Figuur 24 Boxplot Epith.c.size

## Bare.nuclei

Afbeelding met schermopname, Rechthoek, plein, tekst

Automatisch gegenereerde beschrijving

Figuur 25 Boxplot Bare.nuclei

## Bl.cromatin

Afbeelding met diagram, lijn, Rechthoek, plein

Automatisch gegenereerde beschrijving

Figuur 26 Boxplot Bl.cromatin

## Normal.nucleoli

Afbeelding met schermopname, Rechthoek, diagram, plein

Automatisch gegenereerde beschrijving

Figuur 27 Boxplot Normal.nucleoli

## Mitoses

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 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. Het advies is daarom om aan de hand van deze test voor deze feature bij het voorbereiden van de data dit gat op te vullen met betrouwbare waardes of deze feature niet te gebruiken in de uiteindelijke dataset omdat deze momenteel niet betrouwbaar gebruikt kan worden.

#### 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.

## Pairplot

Afbeelding met tekst, diagram, schermopname, Parallel

Automatisch gegenereerde beschrijving

Figuur 29 Pairplot

## Correlatie coëfficiënt 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.

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

# Conclusie

# Figurenlijst

[Figuur 1 Histogram Cl.thickness 4](#_Toc161260755)

[Figuur 2 Density plot Cl.thickness 5](#_Toc161260756)

[Figuur 3 Histogram Cell.size 6](#_Toc161260757)

[Figuur 4 Density plot cell.size 6](#_Toc161260758)

[Figuur 5 Histogram Cell.shape 7](#_Toc161260759)

[Figuur 6 Density plot Cell.shape 7](#_Toc161260760)

[Figuur 7 Histogram Marg.adhesion 8](#_Toc161260761)

[Figuur 8 Density plot Marg.adhesion 8](#_Toc161260762)

[Figuur 9 Histogram Epith.c.size 9](#_Toc161260763)

[Figuur 10 Density plot Epith.c.size 9](#_Toc161260764)

[Figuur 11 Histogram Bare.nuclei 10](#_Toc161260765)

[Figuur 12 Density plot Bare.nuclei 10](#_Toc161260766)

[Figuur 13 Histogram Bl.cromatin 11](#_Toc161260767)

[Figuur 14 Density plot Bl.cromatin 11](#_Toc161260768)

[Figuur 15 Histogram Normal.nucleoli 12](#_Toc161260769)

[Figuur 16 Density plot Normal.nucleoli 12](#_Toc161260770)

[Figuur 17 Histogram Mitoses 13](#_Toc161260771)

[Figuur 18 Density plot Mitoses 13](#_Toc161260772)

[Figuur 19 Histogram Class 14](#_Toc161260773)

[Figuur 20 Boxplot Cl.thickness 15](#_Toc161260774)

[Figuur 21 Boxplot Cell.size 16](#_Toc161260775)

[Figuur 22 Boxplot Cell.shape 17](#_Toc161260776)

[Figuur 23 Boxplot Marg.adhesion 18](#_Toc161260777)

[Figuur 24 Boxplot Epith.c.size 19](#_Toc161260778)

[Figuur 25 Boxplot Bare.nuclei 20](#_Toc161260779)

[Figuur 26 Boxplot Bl.cromatin 21](#_Toc161260780)

[Figuur 27 Boxplot Normal.nucleoli 22](#_Toc161260781)

[Figuur 28 Boxplot Mitoses 23](#_Toc161260782)

[Figuur 29 Pairplot 25](#_Toc161260783)

# Bibliografie

**Het huidige document heeft geen bronnen.**