

Groei Rapport BD04

Data Science



**Groepsleden**:

Szonja Hollós

Roy Wendries

**Opdrachtgever(s):**

Data Science

**Documentinformatie**

**Versiebeheer**

|  |  |  |  |
| --- | --- | --- | --- |
| Versie | Gewijzigd | Auteur | Datum |
| V0.1 | Document opgezet | Roy Wendries | 26/04/2022 |
| V0.2 | Document aangevuld met Fase 1 | Szonja Hollós | 28/05/2022 |
| V0.3 | Document aangevuld met Fase 2 | Szonja Hollós | 13/06/2022 |

**Groepsleden**

|  |  |
| --- | --- |
| Naam | E-Mailadres |
| Szonja Hollós | 1934104hollos@zuyd.nl |
| Roy Wendries | 1747290wendries@zuyd.nl |

Inhoud

[1. Fase 1: Multiple Linear Regression & Logistic Regression 3](#_Toc106011812)

[1.1 Multiple Linear Regression (MLR) 3](#_Toc106011813)

[1.2 Logistic Regression 4](#_Toc106011814)

[1.3 Vergelijking 4](#_Toc106011815)

[2. Fase 2: Decision Trees & Random forests 5](#_Toc106011816)

[2.1 Decision Trees 5](#_Toc106011817)

[2.2 Random forests 7](#_Toc106011818)

[3. Fase 3: Neurale Netwerken & Support Vector Machines 9](#_Toc106011819)

[3.1 Neurale Netwerken 9](#_Toc106011820)

[3.2 Support Vector Machines 9](#_Toc106011821)

[3.3 Vergelijking 9](#_Toc106011822)

[4. Eindevaluatie 9](#_Toc106011823)

# Fase 1: Multiple Linear Regression & Logistic Regression

Er is gekozen voor de Toyota Corolla dataset. Dit heeft verschillende redenen:

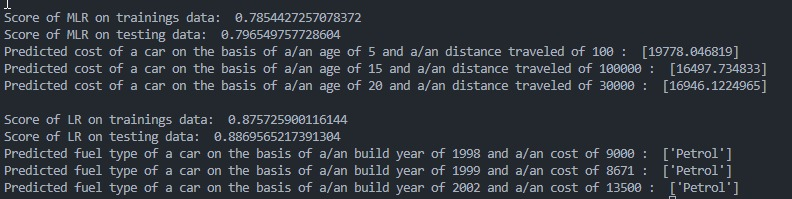
* Zowel voor classification als regression zijn er op het eerste oog voldoende mogelijkheden te vinden binnen de dataset.
* Er is van tevoren gekeken naar de verschillende variabelen waarmee vergelijkingen kunnen worden gemaakt. Op basis hiervan is o.a. de inschatting gemaakt dat de leeftijd van een auto en de kilometertelling aan elkaar gerelateerd zijn en samen eventueel invloed kunnen hebben op de prijs van een Toyota Corolla. Deze vergelijking zal dan ook meegenomen worden in het uitwerken van MLR.
* Hiernaast is er een vermoeden dat er een relatie zit tussen het bouwjaar en de prijs van een Toyota Corolla en het type brandstof dat gebruikt wordt met als opties ‘Diesel’ of ‘Petrol’.
* De dataset bevat een ruime hoeveelheid records om het model te trainen.

## Multiple Linear Regression (MLR)

Bij Multiple Linear Regression worden in ieder geval twee variabelen meegenomen om te kijken of deze invloed hebben op een andere (kwantitatieve) variabele. Hiernaast is er een test gedaan met drie variabelen.

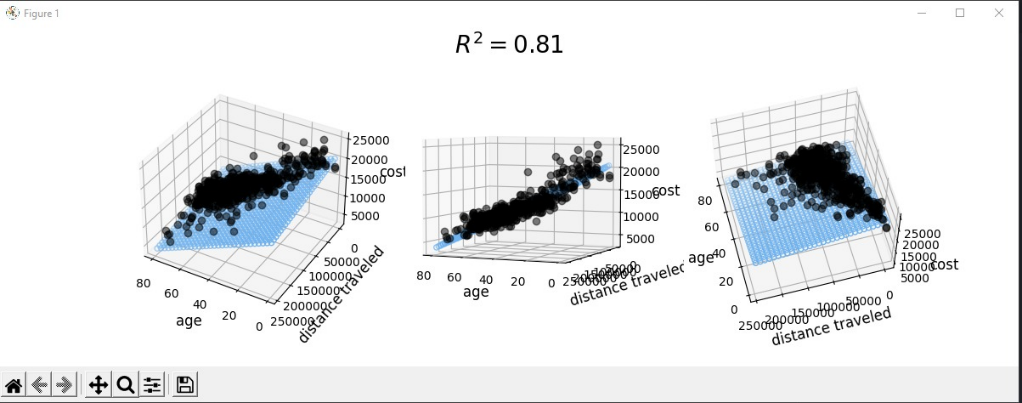
De vraag die met dit model beantwoordt zal worden luidt: Heeft de leeftijd en kilometertelling invloed op de prijs van een Toyota Corolla? Daarnaast wordt er nog gekeken of de ‘Quarterly Tax’ hier een relatie mee heeft als zijnde een derde feature.

Er zijn drie verschillende scenario’s beschreven met twee features als voorbeeld met als resultaat een Accuracy van (afgerond op 3 decimalen) 0,785 op de trainingsdata en 0,797 op de testdata.



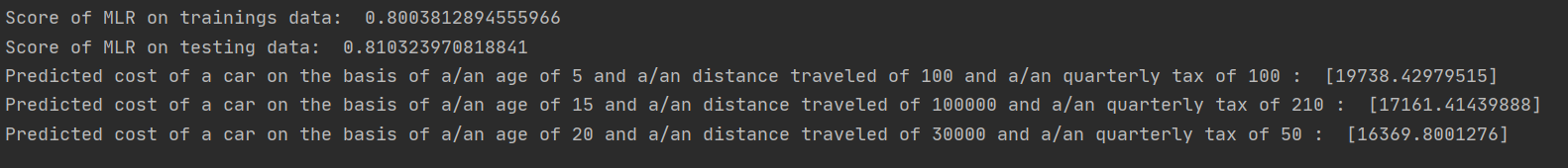
In wezen houdt dit in dat de voorspelling (kijkende naar de testdata) in 79,7% van de gevallen juist is, maar daarmee ook in 20,3% van de gevallen onjuist.

Hieronder een visualisatie van de resultaten in één van de runs waarbij enkel twee features zijn gebruikt:



Hierboven is vanaf verschillende aanzichten te zien hoe de data verdeeld is. De verschillende datapunten lijken redelijk in één lijn verdeeld te zijn zoals de bedoeling is bij MLR.

Bij het gebruiken van drie features zijn de resultaten niet veel beter of slechter dan bij twee features, althans wanneer men naar de score kijkt zoals hieronder weergegeven.

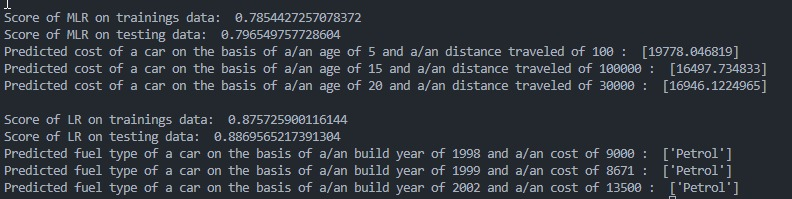


## Logistic Regression

Bij Logistic Regression worden in ieder geval twee variabelen meegenomen om te kijken welke invloed deze hebben op een discrete variabele.

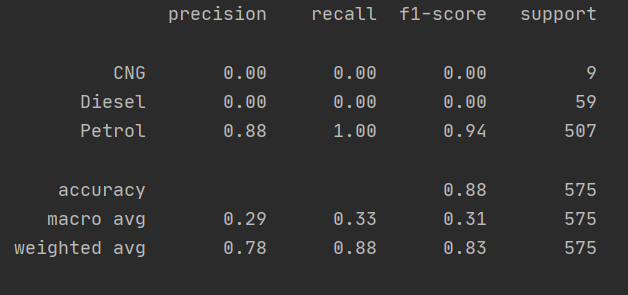
De vraag die met dit model beantwoord zal worden luidt: Heeft het bouwjaar en de prijs van een Toyota Corolla een relatie met het type brandstof dat deze gebruikt? Hiernaast wordt er nog een derde variabele meegenomen om te zien wat voor effect dit heeft op de accuracy, namelijk Quarterly Tax.

Er zijn drie verschillende scenario’s beschreven voor de twee features als voorbeeld met als resultaat een Accuracy van (afgerond op 3 decimalen) 0,876 op de trainingsdata en 0,887 op de testdata.

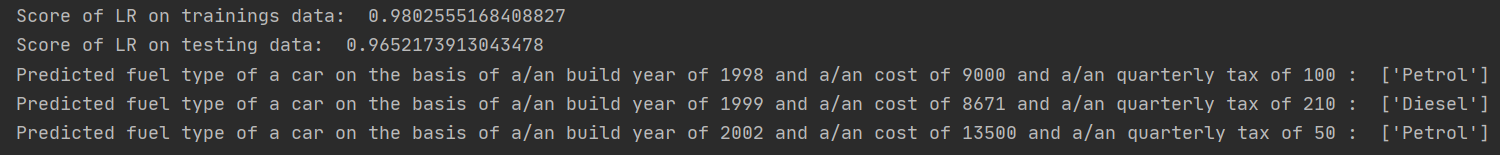


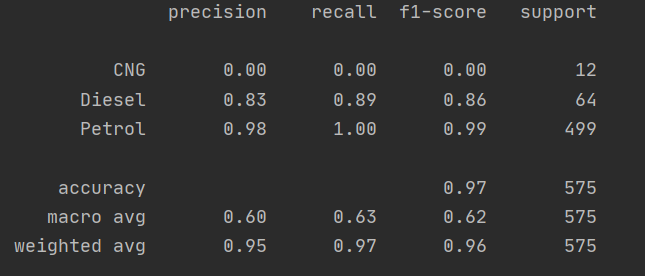
Dit houdt in dat de voorspelling (kijkende naar de testdata) in 88,7% van de gevallen juist is, maar daarmee ook in 11,3% van de gevallen onjuist.

Er zijn nog meer manieren om aan te tonen hoe goed dit model presteert, dit is bijvoorbeeld te presenteren in een tabel zoals in onderstaande afbeelding.



Bij drie features komt er het volgende resultaat uit:





Er kan op basis van bovenstaande scores gesteld worden dat het model beter presteert met 3 features dan met 2 features.

## Vergelijking

De twee bovenstaande methodes zijn niet direct vergelijkbaar gezien zij allebei een ander resultaat opleveren en daarbij niet dezelfde zaken in verband brengen met elkaar. Echter wijst de testdata erop dat er een sterker verband te vinden is tussen het bouwjaar, prijs en type brandstof dan de leeftijd van de auto, kilometertelling en de prijs. Wanneer ook de quarterly tax wordt meegenomen heeft dit bij MLR in dit geval weinig impact op de score, maar bij LR juist wel.

Dit betekent niet dat deze uitkomst als waarheid moet worden gezien, omdat dit afhankelijk kan zijn van hoe het model zelf werkt en de dataset zelf. Het is enkel een indicatie van hoe deze dataset geïnterpreteerd wordt op dit moment door het gebruikte model.

# Fase 2: Decision Trees & Random forests

Na Multiple Linear Regression en Linear Regression wordt nu gekeken naar Decision Trees en Random forests.

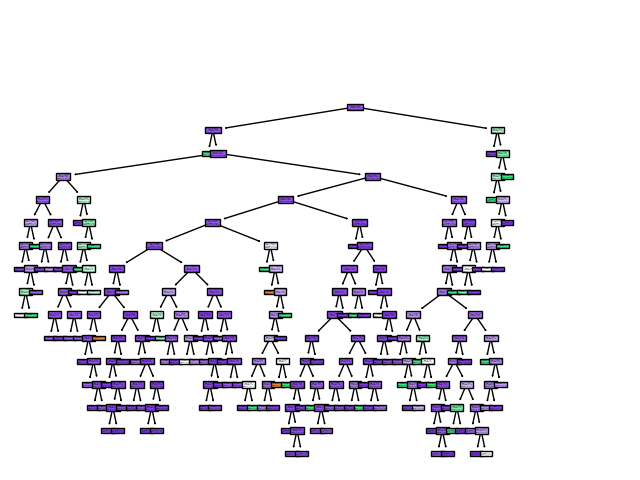
## Decision Trees

Voor Decision Trees wordt naar dezelfde features gekeken als voor MLR en Logistic Regression en dus naar zowel 2 als 3 features. Er wordt gebruik gemaakt van Regression én Classification en deze kunnen dan ook elk meegenomen worden in de uiteindelijke vergelijking.

Eerst zullen de resultaten van Classification worden toegelicht.

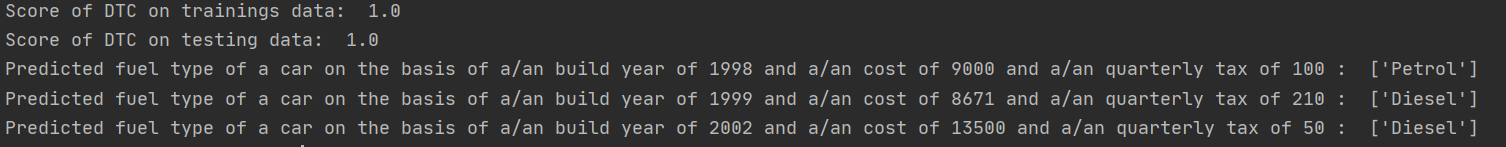
Afbeelding met tekst

Automatisch gegenereerde beschrijving



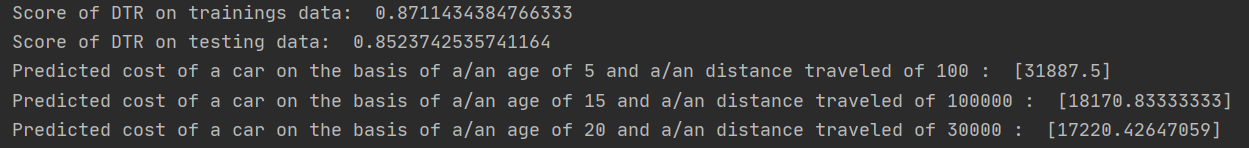
Hier is het resultaat te zien voor twee features. Kijkende naar de score is deze zeer accuraat. Echter is de visualisatie van de resultaten door de hoeveelheid nodes niet al te duidelijk en zijn de resultaten teveel uitgezoomd om de weg naar het resultaat toe te kunnen bekijken.

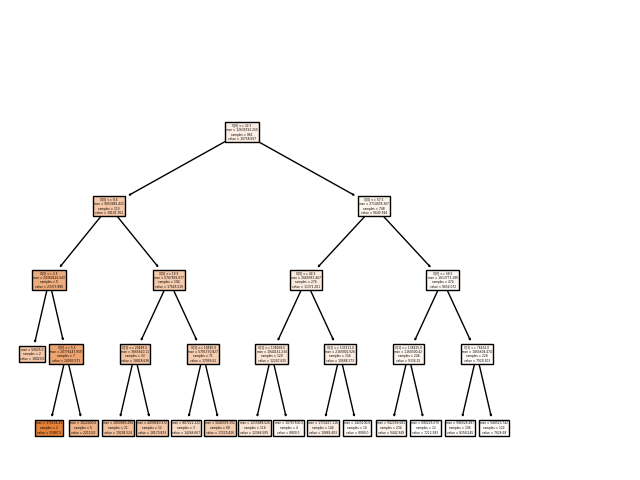
Wanneer er drie features worden meegenomen is de score van de trainingsdata én de testdata 1,0. Dit houdt in dat de trainingsdata 100% overeenkomt en daardoor sterk overfitted is. In dit geval heeft de testdata ook een score van 100% waardoor dit niet direct een probleem vormt, echter kan dit bij andere datasets wel tot problemen leiden.

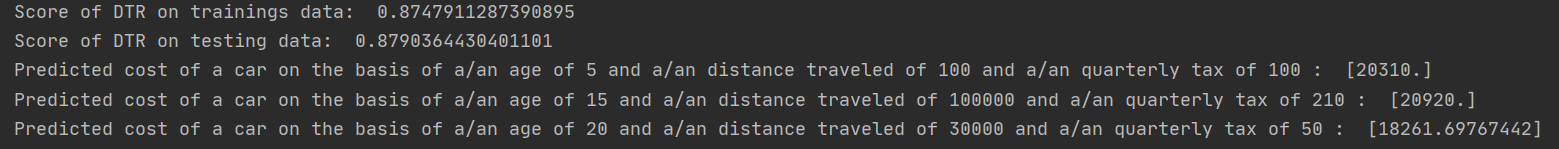


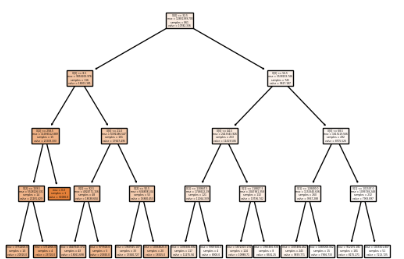


De resultaten van regression zijn in wezen niet te vergelijken met die van classification. In de uiteindelijke vergelijking aan het einde zullen deze resultaten dan ook worden vergeleken met MLR.

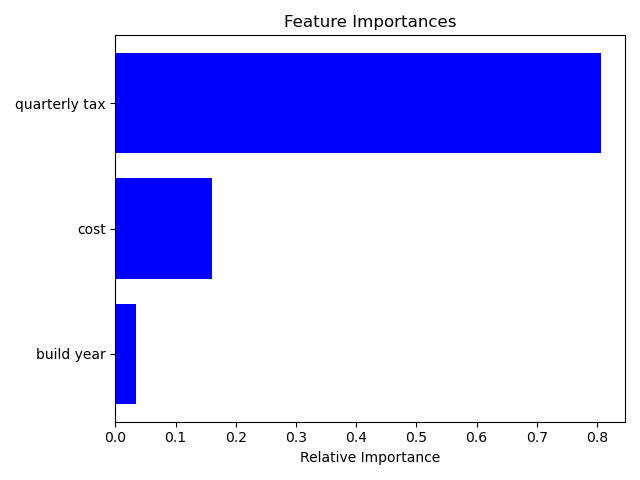
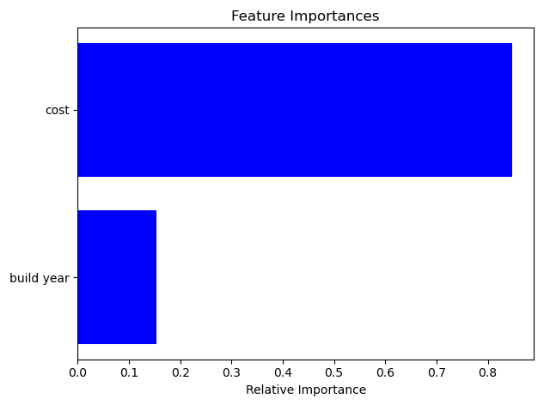
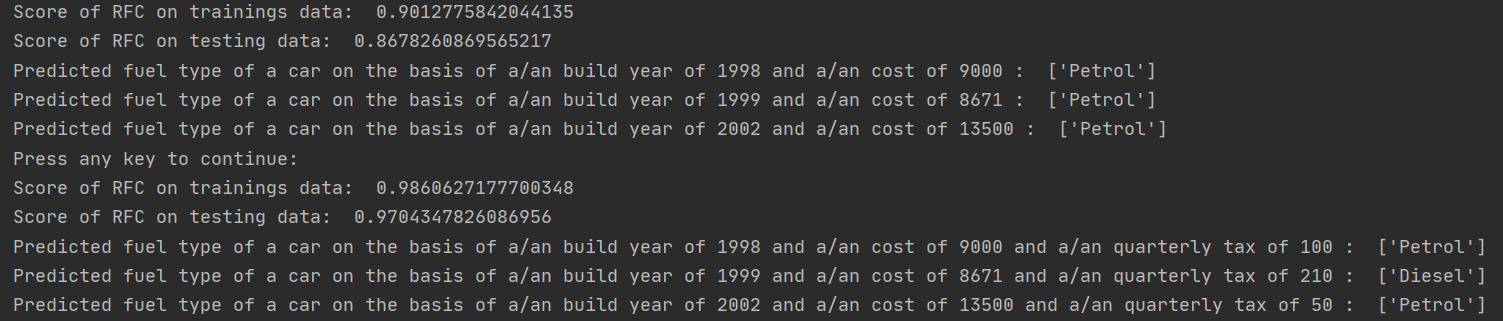




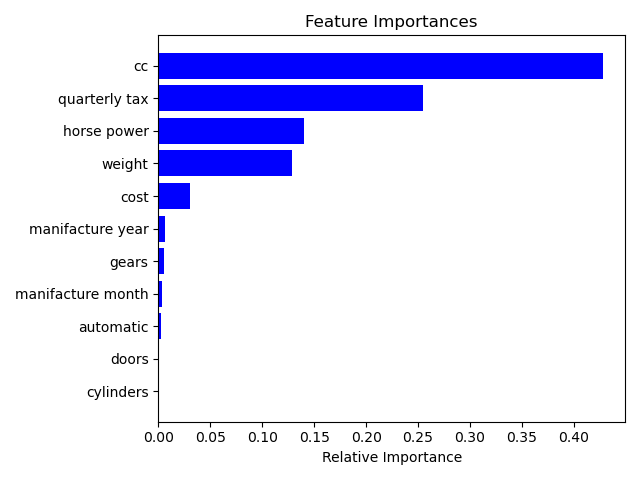
Hierboven zijn de resultaten te zien voor twee features. Kijkende naar de scores zijn deze accurater dan die van MLR en wel met ongeveer 5%. Bij het gebruik van drie features blijft dit verschil aanwezig. Met deze features werkt tot nu toe Decision Trees dus het beste.



## Random forests

Bij random forests zijn er een meer tests uitgevoerd dan bij de drie voorgaande technieken. Zo worden hier dezelfde features getest als bij Decision Trees, maar ook wordt er gekeken wat er gebeurd als er 11 features worden meegenomen. Hieronder staan de resultaten voor zowel 2 als 3 features wanneer er gebruik gemaakt wordt van classification. Hier is te zien dat 3 features ook in dit geval een accurater resultaat opleveren. Echter is het minder accuraat dan wanneer dezelfde features worden meegenomen bij Decision Trees.

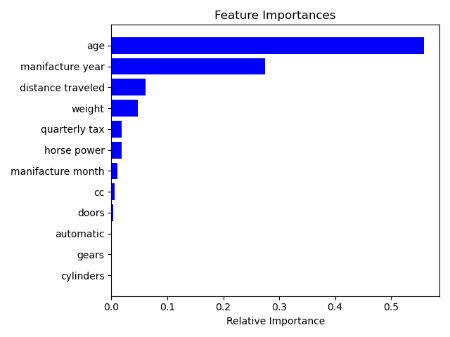
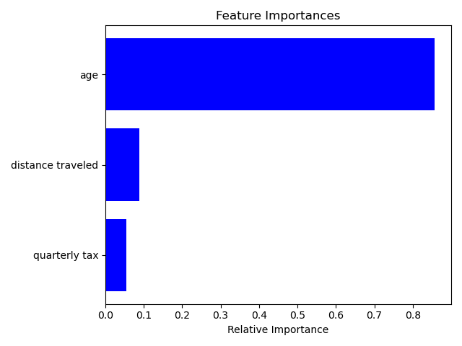
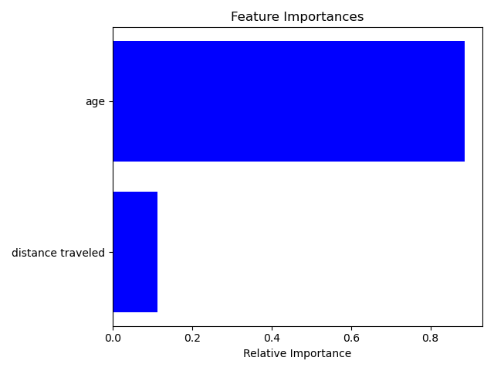
Het beste resultaat van random forests classification komt in dit geval wanneer er 11 features worden meegenomen. Bij het gebruiken van dit model zal dat dan ook de beste optie zijn voor deze dataset.

.

Bij regression zijn de resultaten als volgt voor alle tests zowel met 2, 3 als 11 features:

Afbeelding met tekst, buiten, schermafbeelding

Automatisch gegenereerde beschrijving



Hiernaar kijkende is het model overfitted, omdat de score wel bijna 100% is bij de trainingsdata, maar bij de testdata blijft de score tot wel 10% of zelfs meer achterwege. Deze trend zet door bij alle features en daarmee presteert deze ook niet beter dan Decion Trees.

# Fase 3: Neurale Netwerken & Support Vector Machines

## Neurale Netwerken

## Support Vector Machines

## Vergelijking

# Eindevaluatie