# Database 4 **Eindverslag**

Michelle van Setten 23-08-2020

# Inhoudsopgave

Inhoudsopgave	2
Inleiding	3
Theoretisch kader	4
Wat is data mining?	4
Supervised data mining	4
Unsupervised data mining	4
Classificatie	4
Regressie	4
Onregelmatigheids detectie	4
Clustering	5
Associatie leren	5
Methode	6
vooronderzoek	6
Software keuze	9
Resultaten	10
Eerste ronde	10
Tweede ronde	11
Conclusie en aanbeveling	13
Bronnen	14

## **Inleiding**

Vanwege veel klantrecensies dat er geen realistische prijzen zijn voor de auto's van Used Cars, is er de opdracht gegeven om onderzoek te doen naar wat een realistische verkoopprijs is voor een auto die Used Cars inkoopt. Alhoewel de meeste mensen die een auto kopen bij Used Cars positief terugkijken op de verkoopprijs van de auto, wilt Used Cars natuurlijk meer auto's verkopen en daartoe is de onderzoeksvraag geformuleerd:

"Hoe kan bedrijf x de realistische verkoopprijs van een auto voorspellen door middel van historische gegevens?"

Used Cars beschikt over een grote dataset van tweedehands auto die te koop staat. Deze dataset zal geanalyseerd worden om te kijken of er verbanden zijn tussen de diverse eigenschappen van een auto. Dit analyseren zal gebeuren door middel van datamining.

In dit verslag wordt er eerst beschreven wat datamining is en welke relevante technieken er bestaan. Vervolgens wordt gekeken naar de inhoud van de dataset en de eigenschappen die we gaan gebruiken. Daarna komen de resultaten van het minen aan bod en kan je zien wat je met de resultaten kan doen. Er wordt afgesloten met een conclusie en aanbevelingen.

### Theoretisch kader

### Wat is data mining?

Data mining is het zoeken naar patronen in grote hoeveelheden data. Dit kan op verschillende manieren gedaan worden. Er zijn twee methodes waar dit mee kan worden gedaan, supervised en unsupervised. (Houthoofd, 2015) (McNulty, 2015) (Oracle, 2017)

### Supervised data mining

Bij supervised datamining wordt data gebruikt om een model te maken waarmee je nieuwe data kan voorspellen. Dit kan gedaan worden door een model te trainen. Het model krijgt dan een set met complete data waarin wordt aangegeven wat hij moet voorspellen. Dit kan het vinden van een categorie zijn (classificatie) of het bepalen van een getal (regressie). (McNulty, 2015)

### **Unsupervised data mining**

Unsupervised datamining doe je waarmee je gaat zoeken naar nieuw patronen in de data. Dit model krijgt geen training en moet zelf gaan uitzoeken wat hij moet doen met de data. Dit zou je willen doen als je opzoek bent naar vreemde waardes (onregelmatigheid detectie), je groepen wil maken maar niet weet waar je op wil groeperen (clustering), of als je opzoek ben naar relaties tussen eigenschappen van een waarde (associatie leren). (McNulty, 2015)

### Classificatie

Classificatie wordt gebruikt om binnen data te bepalen bij welke vooraf vastgestelde groep een waarde hoort. Dit wordt gedaan door de waarde op te breken in kleine blokjes en die te vergelijken met de kleine blokjes van waarde die al in de groep zitten. Elk klein blokje krijgt ook een waarde mee om te bepalen hoeveel invloed hij heeft over de groep van de waarde. (Han, Kamber, & Pei, 2012) (McNulty, 2015) (Chin, 2018) (Oracle, 2017)

### Regressie

Bij regressie wordt bijna hetzelfde gedaan als bij classificatie alleen gaat het hier om het bepalen getal. Dat getal kan bijvoorbeeld een prijs zijn of een score. Daarnaast kunnen alleen getallen meedoen aan deze techniek. Bijvoorbeeld een naam kan niet meedoen in een regressie omdat het geen getal is. (Chin, 2018) (McNulty, 2015) (Oracle, 2017) (Foltz, 2018)

### Onregelmatigheids detectie

Om in een dataset te zoeken naar bijvoorbeeld fraude kan onregelmatigheid detectie worden ingezet. Dit model gaat de hele dataset langs om te kijken wat het patroon is in de data en welke waardes daar sterk van afwijken. (Han, Kamber, & Pei, 2012) (Oracle, 2017) (Chin, 2018)

### Clustering

Als er data is waar groepen van gemaakt moeten worden, maar het is niet helemaal duidelijk wat de groepen moeten zijn, dan is het handig om clustering in te zetten. Met clustering kan gaat het model langs de data en probeert deze te groeperen. Het model probeert dan zo veel mogelijk waardes bij elkaar te zetten die op elkaar lijken en dat groepen zo min mogelijk op elkaar lijken. (Han, Kamber, & Pei, 2012) (Oracle, 2017) (Chin, 2018)

### **Associatie Ieren**

Een data mining techniek die we steeds meer terug zien is associatie leren. Bij deze techniek wordt er in de data gekeken of er waardes zijn die vaak samen komen. Bijvoorbeeld als je op YouTube video A bekijkt en daarna video B is er een grote kans dat je ook video C wilt kijken. Hiermee zou je ook persoonlijke suggesties aan klanten kunnen geven op webwinkels. (Oracle, 2017) (Chin, 2018)

### Methode

### vooronderzoek

De dataset die wij voor dit onderzoek gebruiken is afkomstig van de website kaggle.com en telt vijfentwintig kolommen en 509.569 rijen aan gegevens. Doordat het bestand slecht was te in te laden hebben wij er van te voren er voor gekozen om de rijen id, url, region\_url, image\_url en description weg te halen. Van deze rijen wisten wij dat ze geen invloed zouden hebben op ons onderzoek. De rijen die we wel hebben meegenomen zijn;

- Region
- Price
- Year
- Manufacturer
- Model
- Condition
- Cylinders
- Fuel
- Odometer
- Title\_status
- Transmission
- Vin
- Drive
- Size
- Type
- Paint\_color
- Country
- State
- Lat
- Long

Omdat we met ons onderzoek op zoek zijn naar prijzen hebben wij er voor gekozen regressie te gebruiken. Doordat we voor deze methode hebben gekozen vallen bijna alle kolommen af. Bij regressie kan er namelijk geen gebruik gemaakt worden van kolommen die geen getal als waarde hebben. De waarde die dan overblijven zijn price, year, odometer, lat, en long.

Naast dit feit, hebben wij ook een klein vooronderzoek uitgevoerd om te bepalen welke factoren er invloed hebben op de prijs van een tweedehands auto. De resultaten hiervan zijn te zien in tabel 1.

waardes x bron	1	2	3	4	5	6	7	8	9	10	aantal "ja"
bouwjaar	ja	ja	ja	nee	ja	ja	ja	nee	nee	ja	7
kilometerafstand	ja	ja	ja	nee	ja	ja	nee	ja	ja	ja	8
plaats	ja	nee	nee	ja	nee	ja	ja	nee	nee	nee	4
populariteit	ja	nee	nee	ja	nee	ja	nee	nee	nee	nee	3
model/fabrikant	ja	ja	nee	ja	ja	ja	ja	nee	ja	ja	8
aantal deuren	nee	ja	ja	nee	2						
kleur	nee	nee	ja	ja	nee	ja	ja	nee	ja	nee	5
grote van de auto	ja	nee	ja	ja	nee	nee	nee	nee	nee	nee	3
condidtie	nee	ja	ja	nee	ja	ja	ja	ja	ja	ja	8

Tabel 1 resultaten vooronderzoek, bronnen op volgorde; (autogeschiedenis.nl, sd) (Clercq, 2017) (Veldman, sd) (Adesa, 2016) (Truebil, sd) (Threewitt, 2019) (wewantanycar, sd) (Bassett, sd) (webuyanycar.com, 2013) (Vig, 2016)

Hieruit blijkt dat er vier factoren zijn die het meest voorkomen;

- Bouwjaar (Year)
- Kilometerafstand (odometer)
- Model/fabrikant (model/manufacturer)
- Conditie (Condition)

Tijdens het voorbereiden van de dataset zijn wij prijzen tegengekomen die ons niet realistisch leken. Om foute data uit de dataset te filteren zijn wij online gaan zoeken naar prijzen van tweedehands auto's. de resultaten van dat onderzoek is te zien in tabel 2.

min prijs (euro)	max prijs (euro)	website
109	1.280.000	gaspendaal.nl
75	1.999.999	marktplaats.nl
251	14.950.000	autotrack.nl
150	2.000.000	autoscout24.nl
100	2.200.000	autotrader.nl
137	4.485.999	gemiddelde
min prijs (dollar)	max prijs (dollar)	
155	5.063.055	

Tabel 2 prijzen tweedehands auto's

Vervolgens zijn we gaan kijken naar de compleetheid van de kolommen, zie tabel 3.

region	100%
price	100%
year	100%
manufacturer	96%
model	99%
condition	51%
cylinders	61%
fuel	99%
odometer	82%
status	99%
transmission	99%
vin	59%
drive	7 <b>2</b> %
size	33%
type	7 <b>2</b> %
color	68%
country	0%
state	100%
lat	98%
long	98%

Tabel 3 compleetheid per kolom in %

Hieruit blijkt dat conditie minder dan 70% compleet is en daardoor onbruikbaar is. Daarnaast zijn model en fabrikant geen numerieke waarde waardoor ze niet gebruikt kunnen worden in de regressie. De factoren die dan over blijven zijn bouwjaar en kilometerafstand.

### Software keuze

Voor dit onderzoek zijn wij ook opzoek gegaan naar software die het beste bij ons en het onderzoek past. Hiervoor hebben wij gebruik gemaakt van een overzicht van gratis datamining software (Chatterjee, 2020). Hieruit is de volgende top 5 ontstaan;

- 1. Rapid Miner
- 2. Orange
- 3. Weka
- 4. Sisense
- 5. Revolution

Gezien de beperkte tijd van dit onderzoek hebben wij extra voorwaarde gesteld aan de software die we willen gebruiken.

- 1. Software ondersteunt regressie.
- 2. Software is makkelijk in gebruik.
- 3. Software geeft overzichtelijk de resultaten weer

Op basis van deze voorwaarde hebben wij gekozen voor Rapid Miner, zie tabel 4.

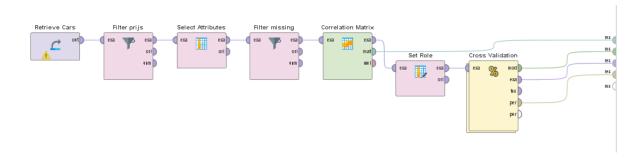
software x voorwaarden	Regressie	Makkelijk	Overzicht
Rapid Miner			
Orange			
Weka			
Sisense			
Revolution			

Tabel 4 Datamining software

### Resultaten

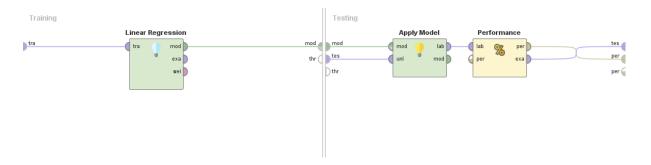
### **Eerste ronde**

Na het vooronderzoek zijn wij een model gaan maken in de verkozen software, zie figuur 1.



Figuur 1 eerste model

In dit model halen we alle gegevens eerst op. Vervolgens filteren we alle auto's onder de gevonden minimale en boven de maximale prijs uit de dataset (filter prijs). Daarna selecteren wij de kolommen die wij willen gebruiken (price, year, en odometer). Dan filteren wij alle auto's die deze gegevens missen uit de dataset (filter missing). We laten Rapid Miner ook een correlatie matrix maken. Deze matrix helpt ons later met het bepalen van de sterkte van de relatie tussen de factoren. Vervolgens zetten we de prijs als target en voeren wij de training en testing, zie figuur 2, uit in de cross validation zodat we zeker weten dat we niet gaan finetune voor perfecte resultaten.



Figuur 2 training en testing

Uit dit model zijn de resultaten gekomen te zien in tabel 5.

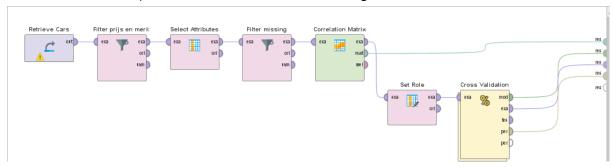
	alle auto's
Correlatie jaar/prijs	0,375
Correlatie kilometer/prijs	0,266
formule	597,915 * jaar - 0,022 * kilometer - 1185408,392
Root_mean_squared_errorgem (fout marge)	11459,404
Root_mean_squared_error div (stabiliteit)	1692,461
aantal auto's	366329

Tabel 5 resultaten eerste model

Uit deze resultaten blijkt dat de formule niet heel betrouwbaar of stabiel is, daarvoor zijn de cijfers te hoog. Daarnaast is de relatie tussen de factoren niet heel sterk, daar zouden de correlaties 0,5 of hoger moeten zijn. (Flotz, 2014) (Foltz, 2018)

### Tweede ronde

Omdat de resultaten van dit model niet sterk zijn hebben wij een tweede model gemaakt waar we ook filteren per fabrikant. Voor het model zie figuur 3 en de resultaten tabel 6.



Figuur 3 tweede model

Fabrikant	Correlatie jaar/prijs	Correlatie kilometer/prijs	formule	Root mean squared error gem (fout marge)	Root mean squared error div (stabiliteit)	aantal auto's
all	0,375	0,266	597,915 * jaar - 0,022 * kilometer - 1185408,392	11459,404	1692,461	366329
ford	0,395	-0,278	579,512 * jaar - 0,022 * kilometer - 1084032,372	10879,465	357,915	92601
chevrolet	0,228	-0,189	366,878 * jaar - 0,018 * kilometer - 719998,047	13527,059	6046,634	73656
toyota	0,622	-0,46	871,745 * jaar - 0,020 * kilometer - 1736600,439	7543,948	1462,088	38159
nissan	0,546	-0,446	646,901 * jaar - 0,023 * kilometer - 1288693,521	6190,082	588,623	26435
ram	0,608	-0,438	1102,773 * jaar - 0,028 * kilometer - 2192538,651	9057,7	820,712	26176
honda	0,599	-0,422	825,715 * jaar - 0,014 * kilometer - 1648336,152	5741,065	3585,943	25585
jeep	0,534	-0,486	617,080 * jaar - 0,045 * kilometer -1220824,811	8400,859	1020,685	23951
gmc	0,612	-0,346	1147,537 * jaar - 0,018 * kilometer - 2285975,928	10135,17	1195,789	23126
dodge	0,351	-0,405	348,469 * jaar - 0,045 * kilometer - 683983,428	9545,162	1582,234	19344
bmw	0,603	-0,296	1181,669 * jaar - 0,014 * kilometer - 2359995,616	9267,069	2640,842	13468
hyundai	0,704	-0,287	922,763 * jaar - 0,005 * kilometer - 1847424,911	4529,546	1853,883	12182
mercedes-benz	0,32	-0,362	562,331 * jaar - 0,087 * kilometer - 1103909,232	18103,625	7755,302	11568
subaru	0,795	-0,642	913,832 * jaar - 0,026 * kilometer - 1823657,498	4125,57	183,909	10259
volkswagen	0,101	-0,216	71,120 * jaar - 0,037 * kilometer - 130069,058	8136,01	9714,896	10214
chrysler	0,517	-0,398	505,830 * jaar - 0,023 * kilometer - 1005568,133	5521,158	636,189	8866
kia	0,716	-0,564	903,643 * jaar - 0,021 * kilometer - 1807696,252	3839,763	733,499	8821
cadillac	0,444	-0,192	660,728 * jaar - 0,013 * kilometer - 1311678,939	12603,55	6444,811	7304
buick	0,423	-0,436	255,442 * jaar - 0,030 * kilometer - 500162,685	5964,597	1084,881	6357
lexus	0,591	-0,352	1375,109 * jaar - 0,013 * kilometer - 2746945,242	9060,26	6494,276	6292
mazda	0,647	-0,507	579,833 * jaar - 0,023 * kilometer - 115197,550	4705,844	646,139	5851
audi	0,276	-0,257	1025,842 * jaar - 0,066 * kilometer -2042044,022	15443,923	17219,778	5586
acura	0,64	-0,524	956,010 * jaar - 0,032 * kilometer - 1906551,536	6724,545	2239,384	4169
infiniti	0,834	-0,501	1473,744 * jaar - 0,010 * kilometer - 2948621,607	4881,939	612,295	3933
pontiac	-0,583	-0,408	-346,854 * jaar - 0,026 * kilometer + 703520,216	4600,611	578,552	3493
lincoln	0,194	-0,178	514,774 * jaar - 0,050 * kilometer - 1017080	13454,18	20700,547	3393
volvo	0,564	0,266	1170,149 * jaar + 0,020 * kilometer - 2342500,509	8714,818	1666,19	3088
mitsubishi	0,528	-0,379	526,318 * jaar - 0,015 * kilometer - 1047051,032	6428,725	693,229	2848
mini	0,748	-0,421	1060,222 * jaar - 0,013 * kilometer - 2120939,771	4310,659	652,737	2173
rover	0,475	-0,133	1410,07 * jaar - 0,004 * kilometer - 2813185,140	15391,249	3346,415	1840
saturn	0,321	-0,144	347,021 * jaar - 0,005 * kilometer - 691546,285	2539,589	1682,815	1790
mercury	-0,322	-0,17	-155,677 * jaar - 0,006 * kilometer + 316985,107	3560,447	1524,904	1760
jaguar	0,486	-0,477	507,789 * jaar - 0,089 * kilometer - 998416,273	8970,155	2723,055	1029
fiat	0,342	-0,695	138,655 * jaar - 0,100 * kilometer - 264490,771	3531,104	539,006	746
harley-davidson	0,349	0,243	502,682 * jaar + 0,040 * kilometer - 997295,463	7320,194	955,043	215
tesla	0,5	-0,28	3224,784 * jaar - 0,018 * kilometer - 6451870,947	11210,638	2760,873	207
alfa-romeo	0,506	-0,655	78,177 * jaar - 0,198 * kilometer - 125729,452	9601,498	6315,625	109
ferrari	0,601	-0,605	2460,243 * jaar - 1,414 * kilometer - 4801973,901	43072,577	26866,981	82
datsun	0,028	0,155	48,254 * jaar + 0,012 * kilometer - 83777,115	26983,624	40966,036	68
land rover	0,791	-0,681	806,620 * jaar - 0,050 * kilometer - 1604716,972	4416,624	3606,714	25
aston-martin	0,711	-0,66	2761,880 * jaar - 0,422 * kilometer - 5475157,527	24516,388	15815,084	22
porche						10
morgan						3
hennessey						2

Tabel 6 resultaten tweede model

Uit deze resultaten blijkt dat ook voor de meeste fabrikanten de formule vrij onbetrouwbaar of onstabiel is. Daarnaast zijn er drie fabrikanten die in deze dataset te weinig voorkwamen om door het model heen te kunnen.

# Conclusie en aanbeveling

Het doel van dit onderzoek was het beantwoorde van de volgende vraag:

"Hoe kan bedrijf x de realistische verkoopprijs van een auto voorspellen door middel van historische gegevens?"

Uit het onderzoek blijkt dat het met deze gegevens er geen accurate formule kan worden gegenereerd. Dit kan doordat er maar twee factoren zijn die meegenomen kunnen worden in het proces, terwijl er in werkelijkheid meer factoren meespelen bij de prijs van een auto. Het kan ook zijn dat regressie niet de juiste manier is om deze vraag te beantwoorde maar dat een andere methode een beter antwoord kan leveren. Echter is het wel zo dat specifiek voor het merk RAM deze methode lijkt te werken. RAM is het enige merk dat bij zowel de stabiliteit als fout marge onder de duizend zit.

Wij raden aan om voor deze vraag nog meer onderzoek te doen. Eventueel zou de dataset uitgebreid moeten worden met meer gegevens die als nummer opgeschreven kunnen worden. Deze gegevens zouden de formule kunnen verbeteren al ontstaat er wel een kans voor te veel gegevens, waar voor gewaakt moet worden. Daarnaast moeten andere mogelijkheden worden bekeken zoals classificatie.

### Bronnen

- Adesa. (2016, juli 11). 7 factoren die de prijs van een tweedehands auto bepalen.

  Opgehaald van cms.adesa.eu:

  https://cms.adesa.eu/nl-nl/blogs/blog/2016/07/11/7-factoren-die-de-prijs-van-ee
  n-voertuig-bepalen
- autogeschiedenis.nl. (sd). Waardoor wordt de waarde van een tweedehands auto bepaald? Opgehaald van utogeschiedenis.nl: http://www.autogeschiedenis.nl/auto-kopen/waarde-tweedehands-auto-bepal en.htm
- Bassett, A. (sd). The Three Key Things that Affect the Price of a Used Car. Opgehaald van instamotor.com:

  https://instamotor.com/sell-car/car-value/3-key-things-that-affect-the-price-of-a-used-car
- Chatterjee, M. (2020, maart 16). Top Free Data Mining Tools of 2020. Opgehaald van mygreatlearning.com:

  https://www.mygreatlearning.com/blog/top-data-mining-tools/
- Chin, S. (2018, juli 12). Data Mining: How You're Revealing More Than You Think.

  Opgehaald van YouTube: https://www.youtube.com/watch?v=EH3bp5335IU
- Clercq, P. I. (2017, januari 22). Tweedehands auto kopen? Hier moet je op letten.

  Opgehaald van rtlnieuws:

  https://www.rtlnieuws.nl/geld-en-werk/artikel/174961/tweedehands-auto-kopen-hier-moet-je-op-letten
- Flotz, B. (2014, december 2). Statistics PL15 Multiple Linear Regression. Opgehaald van YouTube:

  https://www.youtube.com/playlist?list=PLIeGtxpvyG-IqjoU8liF0Yu1WtxNq\_4z-
- Foltz, B. (2018, mei 11). Statistics PL14 Simple Linear Regression. Opgehaald van YouTube:

  https://www.youtube.com/playlist?list=PLIeGtxpvyG-LoKUpV0fSY8BGKIMIdmfC i
- Han, J., Kamber, M., & Pei, J. (2012). In Data Mining Concepts and Techniques (pp. 5-8; 327; 444; 544). Amsterdam: Elsevier.
- Heijst, L. v. (2018, november 1). Regressieanalyse uitvoeren en interpreteren.

  Opgehaald van scribbr.nl: https://www.scribbr.nl/statistiek/regressieanalyse/
- Houthoofd, D. (2015, december 11). Data mining vs. process mining: wat is het verschil? Opgehaald van Horsum.be:

- https://www.horsum.be/nl/blog/data-process-analytics/data-mining-vs-process-mining-wat-het-verschil#:^:text=Data%20mining%20analyseert%20statische%20informatie,op%20de%20meest%20recente%20gegevens.
- McNulty, E. (2015, januari 8). What's the difference between supervised and unsupervised learning? Opgehaald van dataconomy.com:

  https://dataconomy.com/2015/01/whats-the-difference-between-supervised-an d-unsupervised-learning/
- Oracle. (2017). Data mining Concepts. Opgehaald van docs.oracle.com: https://docs.oracle.com/database/121/DMCON/title.htm
- Threewitt, C. (2019, augustus 22). 10 Factors That Affect Your Car's Resale Value.

  Opgehaald van auto.howstuffworks.com:

  https://auto.howstuffworks.com/buying-selling/car-resale-value.htm
- Truebil. (sd). 7 Factors That Influence the Price of a Used Car. Opgehaald van truebil.com:

  https://www.truebil.com/blog/7-factors-that-influence-the-price-of-a-used-car
- Tuil, K. v. (2011, augustus 26). Wat is datamining? Opgehaald van computerworld.nl: https://www2.computerworld.nl/business-intelligence/74941-wat-is-datamining
- Veldman, J. (sd). Welke auto's zijn het meest waardevast en waarom? Opgehaald van bynco.com:

  https://bynco.com/nl/blog/welke-autos-zijn-het-meest-waardevast-en-waarom/
- Vig, P. (2016, juli 27). Factors which affect used car valuation price. Opgehaald van autoportal.com:

  https://autoportal.com/articles/factors-which-affect-used-car-valuation-price-6
  446.html
- webuyanycar.com. (2013, januari 16). 8 Factors That Affect the Resale Value of Your Car. Opgehaald van webuyanycarusa.com:

  https://www.webuyanycarusa.com/blogs/view/8\_factors\_that\_affect\_the\_resale\_value\_of\_your\_car
- wewantanycar. (sd). Factors affecting your car value. Opgehaald van wewantanycar.com: https://www.wewantanycar.com/guides/factors-affecting-car-value