# 3.Data advanced

## Combinatieleer

### Permutaties:

Verschillende soorten combinaties gemaakt uit een verzameling van n-elementen.

Hierin gebruik je telkens alle instanties van de deelverzamelingen.

Formule:

A picture containing text

Description automatically generated

Belangrijk!

Herhaling is niet mogelijk.

Volgorde is wel van belang.

### Variaties:

Verschillende combinaties gemaakt van p-elementen uit een verzameling van n-elementen.

Hierin worden niet telkens alle elementen gebruikt.

Formulte:

A picture containing text, watch, gauge

Description automatically generated

Belangrijk!

Herhaling is niet mogelijk.

Volgorde is wel van belang.

P-elementen 🡪 bv aantal verschillende groepjes van 2 die gemaakt kunnen worden uit een geheel van n-elementen

### Herhalingsvariaties:

Verschillende combinaties gemaakt van p-elementen uit een verzameling van n-elementen.

Hierin worden niet telkens alle elementen gebruikt.

Formule:

Text

Description automatically generated with medium confidence n 🡺 aantal elementen waaruit je kan kiezen.

Belangrijk!

Herhaling is WEL mogelijk.

Volgorde is wel van belang. 🡪 volgorde telt mogelijks mee for normal numbers

Je moet een getal van 3 cijfers (p) maken en je hebt de keuze uit 7 getallen (n)

### Combinaties:

Diagram

Description automatically generatedVerschillende combinaties gemaakt van p-elementen uit een verzameling van n-elementen.

Formule:

Text

Description automatically generated with low confidence

Belangrijk!

Herhaling is niet mogelijk!

Volgorde is niet van belang!

P <= n !!!!!!!!!!!!!!!!!!!!!

De eerste 5 personen die aankomen mogen binnen,

Het maakt niet uit in welke volgorde ze zijn binnengekomen.

Calendar

Description automatically generated

## Kansrekenen

### Termen:

* Universum (U): Verzameling van alle mogelijke uitkomsten.
* Gebeurtenis (A): Deelverzameling van het Universum.
* Kans (P): Waarschijnlijkheid dat een uitkomst zal plaatsvinden.
* Experiment: Hoe je aan je waardes komt.
* Doorsnede (A∩B): Verzameling van elementen die tot zowel A als B behoren.
  + maalteken
* Unie/vereniging (A∪B): Verzameling van elementen die tot A of B behoren.
  + plus
* Verschil (A∖B): Verzameling van elementen die tot A maar niet tot B behoren.
* Disjuncte verzameling: Verzamelingen die geen elementen gemeen hebben. (A∪B)

### Berekeningen:

* Algemene somregel:
* Complementenregel: Kans op universum = 1
* Verschilregel: A picture containing text, clock, watch, gauge

  Description automatically generated Kans van A – de gelijkenis met B

### Voorwaardelijke kans:

Gegeven gebeurtenis B, wat is de kans op gebeurtenis A.

Heel simpel zoek voorbeeld met knikkers.

Zie voorbeeld 14.

Berekening: gebeurtenis A/ Gebeurtenis B.

Originele kans \* kans na gebeurtenis

### Onafhankelijke gebeurtenissen:

Als gebeurtenis A gebeurtenis B niet beinvloed of omgekeerd.

### De wet van de totale kans:

Text, letter

Description automatically generatedDeeluniversums mogen niks met elkaar te maken hebben (disjuncte gebeurtenissen)

Bekijk voorbeeld 17 eens (opname)

Som van kansen is altijd gelijk aan 1

**BOOMSTRUCTUUR**

### Text, letter Description automatically generatedDe regel van bayes:

Boomstructuur weer!

Eerst wet van totale kans 🡪 Kans dat iets zich voordoet gegeven dat eerst iets anders zich voordoet.

## Data representatie

### Gegevens verzamelen:

Het verzamelen/analyseren van gegevens/cijfers.

#### Categorische gegevens:

* Nominaal = niet-geordende categorische systemen.
* Ordinale gegevens = geordende categorische systemen.

#### Numerieke gegevens:

* Discrete gegevens = Gehele getallen (Bv schoenmaat etc)
* Continue gegevens = kommagetallen (Bv lengte etc)

### Gegevens voorstellen:

#### Frequentietabel categorische gegevens:

* Xi: de waardes die voorkomen
* Absolute frequentie (fi): Aantal keer dat een waarneming voorkomt in de gegevens.
* Relatieve frequentie (φi): Percentage van de waarneming tov het totaal.

#### Frequentietabel numerieke discrete gegevens:

* Cumulatieve absolute frequentie: Som van alle gegevens tot aan het huidige punt.
* Cumulatieve relatieve frequentie: Hoeveel % gelijk is aan het totaal?

### Belangrijke klassen-definities:

* Klassengrenzen [x,y[: zo zijn alle getallen groter dan en gelijk aan x en kleiner dan maar NIET gelijk aan y binnen deze klasse.
* Klassenbreedte: verschil tussen grootste en kleinste klassengrens.
* Klassenmidden: is de helt van de stom van de grootste en kleinste klassengrens.
* Klassenfrequenties:

### Chart, pie chart Description automatically generatedGrafische voorstelling:

* Cirkeldiagram (pieplot)
* Staafdiagram (barplot)
* Histogram : continue gegevens
* Boxplot
* Spreidingsdiagram

### Chart, bar chart, histogram Description automatically generatedJupyter code voor grafische voorstelling:

#### Circkeldiagram:

vooropleiding = slaagcijfers\_kort['vooropleiding']

data\_pie\_plot = vooropleiding.value\_counts()

Chart, histogram

Description automatically generateddata\_pie\_plot.plot.pie(figsize=(5,5), autopct='%3.2f%%', title = 'Cirkeldiagram vooropleiding')

#### Staafdiagram:

tweede\_zit = slaagcijfers\_kort['tweede zit']

data\_bar\_plot = tweede\_zit.value\_counts()

data\_bar\_plot.plot.bar(figsize=(5,5),title = 'Staafdiagram tweede zit')

#### Chart, box and whisker chart Description automatically generatedHistogram:

score\_OLOD1 = slaagcijfers\_kort['score OLOD1 (voor afronding)']

score\_OLOD1.plot(kind = 'hist', title = 'Histogram Score OLOD1')

#### Boxplot:

Chart, scatter chart

Description automatically generatedslaagcijfers\_kort[slaagcijfers\_kort['score OLOD1 (voor afronding)'].notnull()].boxplot('score OLOD1 (voor afronding)')

Scatterplot:

plt.scatter(slaagcijfers\_kort['score OLOD1 (voor afronding)'], slaagcijfers\_kort['uren gestudeerd'])

**Bij Pie en bar altijd eerst value counts in een var stoppen.**

### Gegevens samenvatten:

#### Kengetallen voor locatie:

##### Rekenkundig gemiddelde:

Alle waardes opgeteld / aantal waardes = gemiddelde

##### Gewogen gemiddelde:

Houd rekening met het feit dat sommige dingen zwaarder doorwegen.

Voorbeeld met studiepunten en examenpunten:

(Punten \* studiepunten)+(punten\*studiepunten)+…./totaal aantal studiepunten

##### Mediaan:

Rij van alle gegevens en daar de middeste van.

##### Kwartiel:

Als men een dataset in 3 gelijke delen verdeelt is een kwartiel een van die 3 delen.

Q1 is dan de mediaan van het eerste kwartiel

##### Modus:

Observatie van de hoogste frequentie.

#### Kengetallen voor locatie:

##### Spreidingsbreedte:

Grootste waarde – kleinste waarde

##### Interkwartiel afstand:

A picture containing diagram

Description automatically generatedQ3 – Q1

### Boxplot:

Word gemaakt met het volgende:

* Kleinste gegeven
* Eerste kwartiel Q1
* Mediaan Q2
* Derde kwartiel Q3
* Grootste gegeven

### Verband tussen variabelen:

A screenshot of a computer

Description automatically generated with medium confidence

### Categorische variabelen:

Varbiabelen waar die niet direct in cijfers uitgedrukt kunnen worden.

Text, chat or text message

Description automatically generatedBv: goed, middelmatig en slecht.

### Numerieke variabelen:

Deze variabelen worden in cijfers

Uitgedrukt. Bv: schoenmaten, lengte,…

#### (On)Afhankelijke variabelen:

Alle variabelen hebben dezelfde verhoudingen. (Zie ppt)

#### Positief/negatief correleren:

Toename 1 variabele 🡪 toename andere variabele 🡺 positief correleren

Toename 1 variabele 🡪 AFNAME andere variabele 🡺 negatief correleren

#### Chart, scatter chart Description automatically generatedScatterplot:

Een weergave van numerieke variabelen.

Hierbij krijgt een variabele een x en y waarde.

##### Linear/non-linear:

Als tussen de punten een ietwat rechte lijn kan trekken.

Is het een linear verband. Voorbeeld: non-linear 🡺

#### Uitschieters:

Zorgen voor een fout verband tussen variabelen.

#### Verstrengelende factoren:

Correlatie impliceert geen causaliteit. Het feit dat 1 ding voorvalt en het andere ding gebeurt heeft niet altijd een rechtstreeks verband. Het kan zijn dat beiden voorvallen een gezamelijke oorzaak hebben.

### Missing values:

#### Reden:

* Respondenten antwoorden niet
* Antwoord niet opgenomen in dataset

#### Soorten missing values:

* MCAR: waardes die compleet at random missen, deze waardes hebben niets met elkaar te maken. Regematig nulwaardes in verschillende groepen!
* MAR: Regelmatig nulwaardes binnen eenzelfde groep (foutieve waardes?) maar je kan er vaak wel een lijn door trekken.

### Outliers:

Uitschieters tov het grootste deel van de waarnemingen.

#### Oorzaak:

* Onjuiste waarde
* Juiste waarde (onwaarschijnlijk)

#### Detecteren:

* Boxplot
* Z-score

### Populatie:

#### Steekproef:

* Census 🡪 totale populatie
* Steekproef/Sample 🡪 geselecteerd deel van de populatie
  + Selecte steekproef: proefpersonen worden specifiek geselecteerd.
  + Aselecte steekproef: random personen worden gekozen. (alle elementen hebben een even grote kans om voor te komen)

### De dichtheidsfunctie:

De “ideale” beschrijving van de volledige dataset.

Wordt gebruikt bij grote datasets, als staafdiagrammen en freq tabellen te onduidelijk worden.

De dichtheidsfunctie is de vloeiende curve door het histogram.

Chart, histogram

Description automatically generated

#### Soorten dichtheidsfuncties:

* Symmetrische verdeling:
  + Linkertak en rechtertak vormen elkaars spiegelbeeld/
  + Mediaan en gemiddelde zijn ongeveer gelijk
* Rechts-scheve verdeling:
  + Gemiddelde is groter dan mediaan
  + Rechterkant loopt langer uit
* Links-scheve verdeling:
  + Gemiddelde kleiner dan de mediaan
  + Linkerkant klimt langer omhoog

### De normaalverdeling:

🡪 een veel voorkomende dichtheidsfunctie

Een normaalverdeling heeft een typische klokvorm. Deze is symmetrisch rond het gemiddelde.

A picture containing text

Description automatically generated

## Machine learning

### Wat is machine learning:

Machine learning leert machines zelfstandig taken uitvoeren.

#### Soorten machine learning:

* Supervised learning:
  + Classificatie
  + Regressie
* Unsupervised learning:
  + Clustering

### History:

Jaren 50 als schaakrobot

Perceptron 1958 = 1ste kunstmatig neuraal netwerk

### ML problemen herkennen:

ML is het proces waarbij een computerprogramma in staat is te leren hoe een taak uit te voeren en dit door ervaring. Ervaring = data

Diagram

Description automatically generated Diagram

Description automatically generated

Rule based : je past steeds nieuwe regels zelf toe.

Bij ML start je met het toepassen van enkele regels. Deze regels worden automatisch geupdate op basis van nieuwe data.

1. Verzamel een grote dataset.
2. Gebruik een algoritme dat “zelfstandig” een verband vind.
3. Update dit verband voortdurend met nieuwe data.

### Gebruik:

* Netflix
* Zelfrijdende autos
* Slimme thermos
* Facial recognition

### Het ML process:

1. Bepalen welk type ML probleem we hebben.
2. Gebruik de data die je hebt.
3. Pas een standaard algoritme toe op jouw data.

#### Type ML - problemen:

Kies uit:

* Supervised:
  + Zowel input als output is aanwezig
* Unsupervised:
  + Het algoritme heeft enkel data dat voorzien is van input (geen output!)

#### Data:

Je zal grote hoeveelheden data nodig hebben voor ML. Deze data kan je best omzetten naar numerieke data aangezien deze gemakkelijker is om te gebruiken.

#### Algoritme:

Het type algoritme hangt af van het type probleem. Je kan ook meerdere bestaande algoritmes uitvoeren op je data.

### Supervised Machine learning “classificatie”:

#### Classificatie:

Een probleem toekennen aan een categorie. Het heeft een beperkt/bepaald aantal uitkomsten.

Deze categorieën zijn op voorhand bepaald.

Text

Description automatically generated

#### Classifier:

Het algoritme dat een indeling uitvoert. Dit algoritme geeft een label aan het object.

Dit algoritme kunnen we ons voorstellen als wiskundig & statistisch algoritme.

Bijvoorbeeld: Naive Bayes algoritme, Support vector machine algoritme,…

#### Features:

Eigenschappen – kolommen in onze datasets. Het is belangrijk dat deze data features representatief zijn en van numerieke aard. Zelfs tekst en fotos moet je proberen omzetten naar een numerieke grootheid.

#### Training:

De classifier gebruikt een deel van de data waarvan de categorie reeds gekend is. Van deze “training data” leert de classifier hoe nieuwe data te classificeren. Het probeert regels en patronen af te leiden van deze trainingsdata.

#### Testing:

Test het model door gebruik te maken van test data die niet gebruikt is als trainingsdata.

### Naive Bayes Algoritme:

Voorbeeld: Politieagent-jogger

Kan zowel bij ordinaal als bij nominale data toegepast worden.

Aan de hand van data gaat het systeem bepalen welke van de 2. Het systeem heeft sowieso input & output.

1. Globale ouput bekijken : kans op politie 10% en jogger 90%
2. Checken aan welke voorwaarden de persoon voldoet
3. Berekent de voorwaardelijke kans

Waarom naive: Bekijkt alle voorwaarden apart. Handboeien? Bekijkt toch nog walkie & badge etc.

### Chart, scatter chart Description automatically generatedSupport vector Machine:

Enkel bij binaire data.

Voorbeeld: Reclame of geen reclame

Tekeningen worden omgezet naar nummers.

Het algoritme trekt een vlak tussen de twee en neemt dit

Als testdata voorbeeld.

### Diagram Description automatically generatedDecision tree:

Boomstructuur.

Lijkt me if structuur.

Chart, scatter chart

Description automatically generated

### Linear Discriminant Analysis:

Zet alle eigenschappen uit in een n-dimmensionale ruimte. Gaat proberen een rechte te definieren zodat alle data duidelijk te zien is. De rechte is dan je ‘model’.

### Nearest neighbor:

Chart, bubble chart

Description automatically generatedWord hetzelfde geklasseerd als diegene bij wie jij het dichtst ligt. Je labels zijn op voorhand gekend.

### Chart, scatter chart Description automatically generatedSupervised Machine learning “regressie”:

Onbeperkt aantal uitkomsten

= continue variabele.

Voorbeeld: lineare regressie

y

Y = a + bx

Voorspelde waarde zijn de waarden op de lijn zelf.

A

### Chart, scatter chart Description automatically generatedClustering:

Unsupervised machine learning.

🡪 input gegeven maar output niet.

Doel: items groeperen 🡪 clusters zijn nog niet gekend (niet zoals nearest neighbors)

Gegroepeerd op gelijkenissen.

### Recommendations:

#### Doel:

* Klanten trouw laten zijn
* Aanbevelingen doen

#### Collaborative filtering:

Op basis van items die jij een score hebt gegeven gaat het systeem kijken naar andere gebruikers die dezelfde scoren hebben gegeven. Hierbij bekijken ze dan dingen die jij ook interessant kan vinden en geven je dan een aanbeveling.