Misbrug af data og A.I bruges til avancerede profileringer

Studieretningscase 2019



**Lavet af Marcus Presutti**

**Klasse - 3.F**

**Dato: 19/08-2019**

**Lærere:**

**Christian Reinhold (CHR)**

**&**

**Lars Barlebo (LMBA)**

Indholdsfortegnelse

[**Indledning 3**](#_Toc17138373)

[**Redegørelse 3**](#_Toc17138374)

[**Machine Learning 4**](#_Toc17138375)

[**Clustering 4**](#_Toc17138376)

[**Script 5**](#_Toc17138377)

[**Algoritme & matematiske udregninger 6**](#_Toc17138378)

[**Analyse 9**](#_Toc17138379)

[**Menneskerettigheder 10**](#_Toc17138380)

[**Data Portability 11**](#_Toc17138381)

[**Vurdering 11**](#_Toc17138382)

[**Konklusion og diskussion 11**](#_Toc17138383)

[**Bibliografi 12**](#_Toc17138384)

[**Kilder 12**](#_Toc17138385)

[**Bilag 14**](#_Toc17138386)

# 

# Indledning

De digitale fodspor, som vi efterlader, når vi ‘liker’ og kommenterer posts på nettet, er i virkeligheden vigtige data, som fortæller en hel del om vores identitet og opførsel. Disse data bruges til at lave psykometriske profileringer af os. Denne form for data-analystik bruges ofte til at lave harmløse reklamer, som appellerer til sine målgrupper. Desværre har der i forbindelse med der amerikanske præsidentvalg været stort fokus på misbrug af denne data. Når man ved nok om en person, kan man udnytte individuelles svagheder, og manipulere folk til at gøre nogle bestemte handlinger, til andres fordel. Denne slags data-misbrug kan ses i form af propaganda i medierne. Vdr. skandalen med Cambridge Analytica (CA) i 2016 så man også, at reklamer og kampagne-videoer var direkte rettede mod let påvirkede og mentalt svage mennesker (*persuadables*). Denne metode påvirkede markant resultaterne under, eksempelvis, præsidentvalget 2016. Succesen skyldtes altså, at CA havde omkring 5.000 datapunkter på over 50. mio. amerikanske vælgere, som blev brugt til at lægge strategier i valgkampagnen (The Great Hack, 2019)

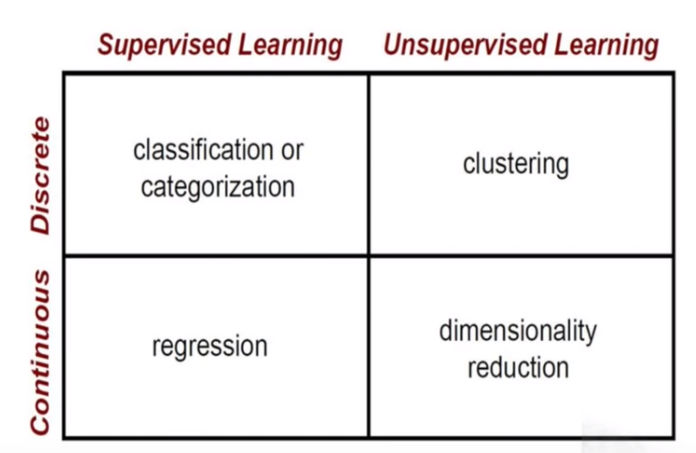
Jeg vil i denne opgave fokusere på, hvordan det overhovedet er muligt at lære nok om folk, på baggrund af fx almen Facebook-aktivitet, til at man kan forudse folks opførsel, og påvirke, hvordan de stemmer; med fokus på klassificering- og clustering-algoritmer. I den sammenhæng vil jeg også dække, hvorfor data skal beskyttes mere, da der er alt for let adgang til det; men også hvordan teknologien bag dette kan / vil blive brugt i fremtiden.

# Redegørelse

Ordet, ‘data’, bliver ofte brugt, når man snakker om information. Man kan umiddelbart dele disse op i to kategorier; kvantitativ- og kvalitativ. Kvalitativt data skal beskrives, mens kvantitativt data kan måles, i form af tal, hvilket er optimalt for computere. Kvantitativt data kan fx være hvor høj man er, hvor meget næring, som er i éns morgenmad, eller hvad en pose chips koster. Man kan yderligere sortere kvantitativt data i grupperinger (*discrete data*) og kontinuer data. Når man vil finde ud af, hvad folk sandsynligvis vil stemme på, vil man finde mønstre i dataene, som foreslår dette. Undersøgelser har vist, at man kan bestemme en persons ideologi (rød eller blå blok), ud fra deres Facebook (like-)historik, med 90% sikkerhed (Bæk K, 2017)**.** CA tog dette til et anden niveau i valgkampen. Ved at se på data fra personlighedstests, samt almen like-historik, kunne de vurdere, hvem der havde størst sandsynlighed for at kunne ændre politisk holdning. Disse mennesker, *the persuadables*, blev hurtigt målgruppen i valgkampen, da CA (som arbejdede for Trump), viste hvordan man skulle påvirke amerikanerne optimalt.

Ifølge amerikansk lov, er dette slags samarbejde ulovligt i valg-kampe. Desuden virker det grotesk, at der findes så meget data om én, som er til rådighed for andre mennesker. Valgkampe er én af de mange måder, data kan blive brugt imod os. Derfor blev der i EU i 2018 indført en datasikkerhedslov, GDPR (General Data Protection Regulation). Dette strammede reglerne for, hvad man må med data, og hvordan man indsamler det. Dette inkluderer data portability, som giver brugerne flere rettigheder og kontrol over sin egen data

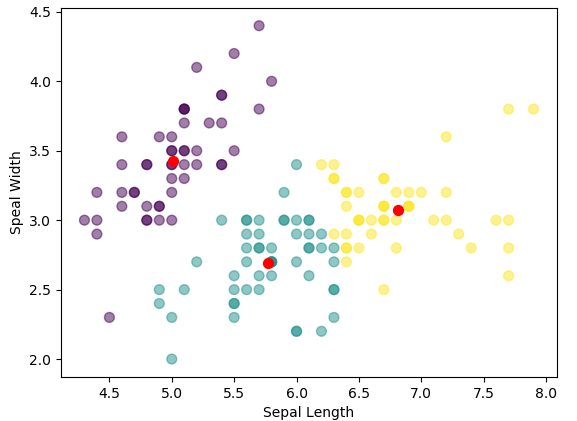
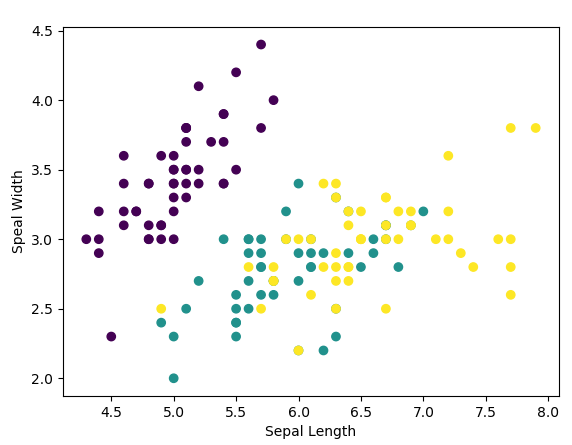
## Machine Learning

Teknologien, som er grundlag for hele problematikken, er machine learning. Det vil altså sige, at computere selv er i stand til at lære, uden eksplicit at programmere, hvordan den skal lære.Hovedmekanismen i machine learning er, at computere kan finde mønstre i data, og dermed identificere og forudse ting. Ved brug af *unsupervised learning,* kan den selv forudse, hvilken gruppe, nogen hører ind under. Man kunne eksempelvis gruppere folk efter ideologi eller efter hvor lette folk er at påvirke. Modsat *supervised learning*behøver den ikke at kende til hvilke grupper, som eksisterer i forvejen; de opdages mere eller mindre af sig selv. En måde at gruppere folk ved brug af *unsupervised* learning, er *‘clustering’*.

### Clustering

Den mest populære clustering-algoritme kaldes *k-means-clustering*. Supervised-counterpart til denne er *k-nearest-neighbors*, hvor grupperne allerede kendes, når de læses i træningsdataen:

Figur 1 - Datatyper og maskinlærings-metoder

**K-means-clustering**  **K-nearest-neighbors**

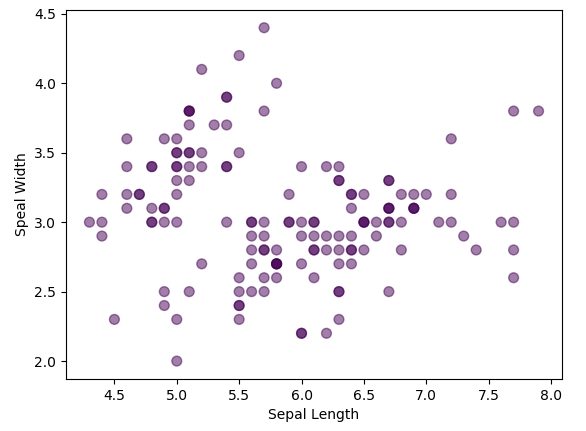
Figur 2.2

Figur 2.1

Hver farve repræsenterer et cluster. Fig. 2.2 viser hvordan clusterne rigtigt er fordelt, mens fig. 2.1 viser den gennemsnitligt rigtige cluster. (Man har her valgt, at den skal deles op i 3 clusters)

K-means-algoritmen går ud på, at man deler dataene op i *K* grupper, som virker mest naturligt, afhængigt af dataenes densitet. På fig. 2.3 viser de mørkere punkter, hvor dataene ligger tættere op ad hinanden. Man må selv vurdere, hvor mange grupper / ‘clusters’ der skal være, mens algoritmen vurderer, *hvor* grupperne skal placeres

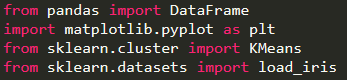
.

****

Figur 2.3

(**Figur 3**)

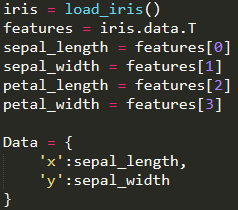
#### Script

**Pandas**-modulet bruges til data-analystik. Jeg gør brug af DataFrame, for at have ene relevant datastruktur til grafen. 

Figur .1

**Matplotlib** bruges til at vise grafen i et plot.

**Sci-Kit Learn (sklearn)** leverer både dataset i numpy-array-format, samt machine-learning-algoritmer.

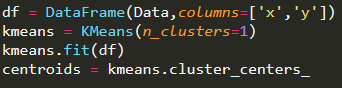


Figur 3.2

Først loades iris-datasættet. Datasættes er bygget som et dictionary-format, hvor hver key står for henholdsvis data, beskrivelse af datasæt m.m.

Ved at sætte et ‘T’ på data-key’en, separerer man hver data-column i hver sin liste.

Én kombination af data, som man kan sammenligne er stilkelængden og stilkebredden.

**DataFrame()** ændrer datastrukturen. 

Figur 3.3

**KMeans()** sætter K clusters tilfældigt, som senere bliver ændret.

**fit()** udfører hele K-means algoritmen, hvor der rykkes rundt på tyngdepunkter (*centroids*) og clusters

**cluster\_centers\_** angiver koordinaterne til hvert tyngdepunkt



Figur 3.4

Ved at gøre brug af matplotlib-modulet, kan vi sætte dataene ind i et plot, og samtidigt vise, hvor tæt data er på hinanden (densiteten). Samtidigt placeres tyngdepunkterne som en rød prik.



Figur 3.5

Hver akse får tildelt en titel på plottet

Når alt er ’commitet’ til plottet, kan man vise grafen i et separat vindue.

#### 

#### Algoritme & matematiske udregninger

For at kunne tildele data til et cluster, skal man kende densiteten, altså tætheden af dataene. Man kan udregne, hvordan hvert cluster skal fordeles, ved brug af følgende algoritme (Lavrenko, 2014).

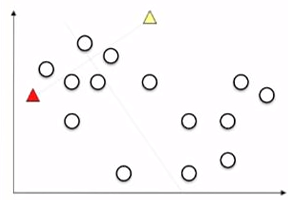
1. Input:

K = antal tyngdepunkter (‘*centroids’*)

n = antal datapunkter

Set punkter, x1 … xn

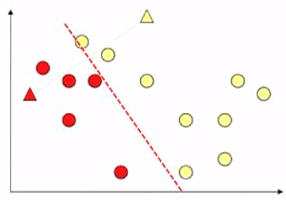
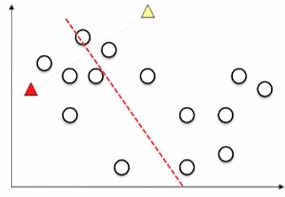
2) Placer tyngdepunkter tilfældigt på koordinatsættet



Figur .1

3) For hvert datapunkt, xi :

* Find nærmeste tyngdepunkt
* Sæt punktet xi til en gruppe (*‘cluster*’), j

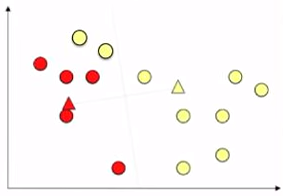


Figur 4.2

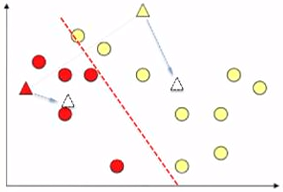
Figur 4.3

4) For hvert ‘cluster’, j = 1 … K :

* Sæt et nyt tyngdepunkt, som svarer til gennemsnittet af alle datapunkternes positioner, xi, i nuværende cluster, j

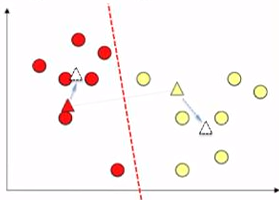


Figur 4.



Figur 4.4

5) Gentag fra *step 3* indtil ingen punkter ændrer sig

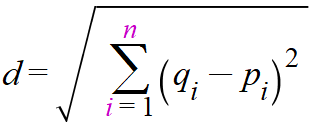


Figur 4.6

I algoritmen bruges 2 udregninger: distancen mellem datapunkt og tyngdepunkt, og den gennemsnitlige position mellem vektorpunkter:

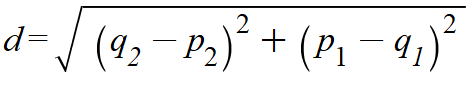
**Distanceformlen:**

Distance mellem 2 punkter kan teoretisk set foregå i, ikke blot 2D og 3D, men også i 4+ dimensioner. Derfor kan man kigge på den universelle distanceformel, som er afhængig af *n* dimensioner.



*Man tager kvadratroden af summen af forskellen af koordinater i hver n dimension, opløftet i anden. (q = punkt1; p = punkt 2)*

Man kan også kigge på den almene, som jeg bruger i sammenhæng med min egen clustering, som er i 2 dimensioner.

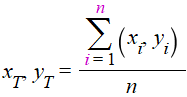


*Her sker ligesom i ovenstående, men der skal kun lægges 2 værdier til hinanden, så summeringstegnet bliver sjældent brugt, når man kigger på distancer i 2 dimensioner*

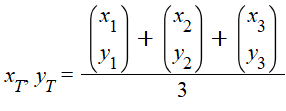
**Gennemsnitsposition**

For at finde tyngdepunktet, skal man kende gennemsnitspositionen i alle dimensioner. Dvs. alle x-koordinater lagt sammen, og delt med antal datapunkter, alle y-koordinater lagt sammen osv.

(n = antal punkter)



I tilfælde af, at der er 3 punkter, og 2 dimensioner, kan man skrive følgende:



# Analyse

I 2014 fandtes en app på Facebook, kaldt *‘This is your digital life*’. Dette var en personlighedstest, som var blevet godkendt af Facebook, i form af Terms of Conditions. Testen blev taget af omkring 320.000 amerikanere, men samlede samtidigt data fra hver ven fra alles venneliste. Appen opsamlede i alt data fra over 50 mio. mennesker, hvilket er absurd, da vennerne ikke har noget, at skulle have sagt. Det var sådan, at Cambridge Analytica fik info om så mange mennesker, da de havde købt den data af appens ejer. Med alt den data, er der over 5000 datapunkter på enhver amerikansk vælger (The Great Hack, 2019). Hvis man kan genkende en iris-blomst på en kombination af to variabler / datapunkter, så overvej præcisionen ved at kombinere 5000 variabler. Den slags data kan næsten bruges som et våben, så farligt det er. Man kan lære utroligt meget om folk, og med de rette metoder. Man kan teoretisk set sagtens bruge den almindelige k-means-clustering algoritme, selvom den ser begrænset ud til 2 variabler/dimensioner af data ad gangen. Algoritmen, som nævnt, giver fuld mulighed for at udvide i uendelige dimensioner. Enten bliver resultatet yderst præcist, ellers bliver det bare kludret. Det meget afhængigt af, hvor mange clusters man bruger. Umiddelbart vil nøjagtigheden falde eksponentielt, jo flere clusters, der laves.

Hvordan kan dette virkelig være muligt, at nogen besidder alt denne data?

CA har remedierne til at analysere dataene, og bruge den til hvad end de lyster. “Knowledge is power”, som man siger. Det er ikke mystisk, at CA har taget størstedelen af æren for valgsejren. Det lyder yderst forkert, og etisk ukorrekt, at store virksomheder kan få fat i den data. Dette viser blot, at man som forbruger ikke tager sig af ‘Terms of Condition’, da nogle ting slet ikke bør være tilladte til at begynde med. Loopholes i kontrakten gør, at folk bag fx personlighedstesten, kan ”snyde” sig frem til data. Ingen love var brudt i den sammenhæng, men integriteten mellem bruger og, specielt, Facebook var blevet svækket. Mange andre tjenester end Facebook lider af samme problem.

Grunden til postyret er, at reglerne om data-indsamlings-krav var skrevet før moderniseringen af internettet og muligheden for at opsamle store mængder data. Det skal forstås, at for 20 år siden, havde vi ikke samme muligheder for kommunikation, som i dag.

Af denne grund er indførslen af GDPR en historisk begivenhed. De strengere regler betyder først og fremmest, at det er sværere at finde huller (‘*loopholes*’) i Terms of Conditions, som kan bruges til at udnytte brugere. Brug af loopholes er i sig selv umoralsk, men ikke nødvendigvis ulovligt. En anden vigtig ting er også, at man selv har mere kontrol over, hvilke personlige data, som lægges ud på nettet / bliver indsamlet. Selvom vennerne til dem, som tog personlighedstesten, ikke havde taget testen, kan man formode, at disse venner deler mange karakteristika med test-tageren. På den måde får de mere data. Dette er en anden clustering-metode (dog en lidt vag fremgangsmåde), kaldes ‘look-alike-clustering’. Derfor bør vennelister også være private, da de også kan blive brugt i sammenhæng med data-misbruget (EU, 2016).

Man kunne sagtens lave mere overbeskyttende regler, men det ville sandsynligvis ødelægge en stor gren af ‘reklame-industrien’, da de er afhængige af en hvis mængde data. Hovedmålet ved GDPR er, at éns *persondata* er beskyttet. Dette er første skridt mod data-frihed, som af flere ses som værende en menneskerettighed.

## Menneskerettigheder

Det er værd at kigge på, at vi dog stadig bliver røvet for vores data. Det giver kun mening, at vores information er vores ejendel, og ikke et produkt, som andre tredje-parter kan handle med. De originale ejere af dataene bliver nemlig ikke betalt direkte. Af samme årsag har man copyright-selskaber, som sørger for, at folks, eventuelt, musik ikke bliver taget af andre. Vi har lignende, ejendomsretten, som sikrer, at vi må beholde, hvad end vi har købt, og ikke må få det stjålet (FN, 1948). Rent juridisk er det stadig lovligt at dele meget data, på trods af GDPR, men det virker umoralsk, da vi, som mennesker, har rettighed til at beholde vores data. En stor ulempe er dog, at alt data-handel foregår bag vores ryg. Det er en dyr og besværlig proces (se figur 6), men alligevel er det pengene værd, fordi vi bliver påvirkede af reklamerne. 

Figur 5

Selvom reklamer er irriterende, har de aldrig været et kæmpe problem i vores hverdag. De er skyld i, at mange ting er ‘gratis’ (i form af penge). Det kan dog blive et problem, hvis det påvirker det demokratiske system i land, som USA, hvor frihed betyder meget for del almene borger. Når reklamerne pludseligt vises i form af propaganda, manipulation og Fake News, føler man, at sin data er blevet udnyttet. Det er dog usandsynligt, at man indser, hvis man bliver manipuleret. Propagandaen, i form af overmanipulerede reklamer, om at skabe hade mod immigranter eller en politisk kandidat, er rettet mod grupper / clusters, som er vurderet til at blive let påvirket af disse ting. Med en blanding af avancerede algoritmer og dygtige data-analytikere, har man fundet emner, som rammere hårdere på folk i visse områder, aldersgrupper m.m.

Misbrug af data og machine learning er grunden til, at mange frygter A.I og dets potentiale. Man føler sig heller ikke nær så sikker, vidende, at en computer ved mere om éns behov, end man selv gør. Machine Learning har potentiale for store ting i fremtiden, men er stadig afhængig af både data og folk, som skal finde data, som maskinen kan bruge. Derfor er det første problem, som man bør løse, være, at få bedre datasikkerhed.

## Data Portability

Et stort problem, som automatisk opstår, når man giver sin data til tjenester som Facebook, Google, Twitter osv., er, at de opbevarer alt din data i ‘siloer’. Man har altid kunne spørge tjenester om at få sin egen data udleveret, men det har aldrig været en let proces. Indtil i 2018, hvor GDPR indførte et andet smart koncept, som netop kigger på nogle menneskerettigheder. Det går ud på, at data ikke skal ejes af virksomheder, men skal kun være tilgængelig for virksomhederne, hvis ejeren bruger deres service. Hvis man ikke vil bruge Facebook mere, har man altså rettighed til at både downloade alt data, som de har på én, men også overføre den data til andre tjenester. Dette skaber også mere konkurrence mellem tjenesterne, om hvem, som kan få brugernes opmærksomhed. I teori, bør vores data være gemt i en personlig silo, som web-tjenester kan få fat i, med *vores* samtykke, og ikke os, som skal have fat i dem, for at få fat i vores egen data.

Fra et teknisk standpunkt skal de bruge universelle filtyper, så data let kan sendes og fortolkes mellem tjenesterne. Filtyper som JSON, og xml-filer er populære database-formater. Det kan også diskuteres, om hvert menneske skal have sin egen server, som de kan opbevare alt sin personlige data på.

# Vurdering

### Konklusion og diskussion

Machine Learning kan være et våben i de forkerte hænder, men er ingenting uden ordentligt data, hvilket skaber en masse misforståelse. Den bedste machine learning algoritme er den, som skal bruge så lidt træningsdata som muligt, for at være effektiv. Hvor effektivt en algoritme kan blive, er også begrænset. Machine learning er ligesom en pistol; den ser skræmmende ud, men er ubrugelig uden ammunition.

Angående regler for databeskyttelse og -frihed bevæger os hen i den rigtige retning. Der er blevet reguleret på de vigtigste områder, så som mere kontrol over egen data, og regler vdr. betingelseskontrakter. Det kan tydeligt ses på hjemmesider, at der er mere fokus på, at spørge brugeren om tilladelse, inden de kommer ind på siden. Resten ser man ikke nær så tydeligt, men er betrykkende, at man ved, at der er blevet en forskel. Vi skal ikke nødvendigvis være helt tilfredse, og lade situationen passere. GDPR gælder kun i EU, hvilket betyder, at andre steder i verden stadig er påvirkede af ældre love. Dog bliver andre steder påvirkede, hvis de skal kommunikere med steder i EU.

Hvis GDPR kun gælder i EU, skal vi så stadig formode, at der kan ske en forskel i næste amerikanske præsidentvalg, sammenlignet med valget i 2016? Umiddelbart har hele CA-skandalen sat ekstra lys på, at data ikke bliver udnyttet i samme sammenhæng. Desuden er mange store firmaer centraliseret i Europa, og formodentligt bliver der også strammere regler i USA og andre steder. Man kan dog heller ikke udelukke muligheden, da machine-learning-teknologien kun er blevet bedre siden da. Selv 100 datapunkter på hver amerikaner, eller endda kun måske 25 kan havde en vigtig betydning. Man skal bare passe på med at undervurdere vigtigheden ved data.

# Kilder

Referencer

Bæk K, J., 2017. *PLOS ONE.* [Online]   
Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184562  
[Senest hentet eller vist den 15 August 2019].

EU, 2016. *gdpr-info.* [Online]   
Available at: https://gdpr-info.eu/  
[Senest hentet eller vist den 15 August 2019].

FN, 1948. *FN's Verdenserklæring om Menneskerettighederne.* s.l.:s.n.

Lavrenko, V., 2014. *K-means clustering: how it works,* s.l.: s.n.

*The Great Hack.* 2019. [Film] Instrueret af Karim Amer, Jehane Noujaim. USA: Netflix.

Illustrationer

**Figur 1**

<https://www.youtube.com/watch?v=9991JlKnFmk>

**Figur 2.1 - 2.3**

*Grafer er resultat af scripts fra følgende kilder:*

2.1) <https://github.com/vprusso/youtube_tutorials/blob/master/machine_learning/iris_classification/part_2.py>

2.2 & 2.3)

<https://datatofish.com/k-means-clustering-python/>

**Figur 3.1 - 3.5**

*Kode ligger I bilag*

**Figur 4.1 - 4.6**

Lavrenko, V., 2014. *K-means clustering: how it works,* s.l.: s.n.

**Figur 5**

*The Great Hack.* 2019. [Film] Instrueret af Karim Amer, Jehane Noujaim. USA: Netflix

Andet

### Datasæt

Iris

<https://archive.ics.uci.edu/ml/datasets/Iris>

### Kode-snippet

Clustering i python

<https://datatofish.com/k-means-clustering-python/>

Iris datasæt i python

<https://github.com/vprusso/youtube_tutorials/blob/master/machine_learning/iris_classification/part_2.py>

### Baggrundsviden

K-means clustering algoritme & forståelse af machine-learning algoritmer

<https://www.youtube.com/watch?v=9991JlKnFmk>

<https://www.youtube.com/watch?v=_aWzGGNrcic>

Cambridge Analytica & Facebook drama

<https://medium.com/@privacyint/what-zuckerberg-forgot-to-mention-profiling-7b7c596b9823>

<https://www.dr.dk/nyheder/viden/teknologi/overblik-saadan-fik-og-brugte-cambridge-analytica-50-millioner-profiler>

<https://www.theguardian.com/commentisfree/2018/mar/23/plenty-more-like-cambridge-analytica-data-facebook>

<https://www.refinery29.com/en-us/2019/07/238738/what-happened-after-netflix-documentary-the-great-hack>

<https://en.wikipedia.org/wiki/Cambridge_Analytica#Elections>

Social profilering

<https://www.researchgate.net/publication/293089145_Towards_semantic_dataset_profiling>

GDPR

<https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en>

<https://gdpr-info.eu/>

Data:

<https://blog.minitab.com/blog/understanding-statistics/understanding-qualitative-quantitative-attribute-discrete-and-continuous-data-types>

Menneskerettigheder

<https://amnesty.dk/om-amnesty/fns-verdenserklaering-om-menneskerettigheder>

<https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf>

# Bilag

Script

