Bruk av maskinlæring i klassifikasjon av kriminalsaker

Jan Roar Beckstrøm
Avdelingsdirektør/chief data scientist
Riksrevisjonens innovasjonslab
Nasjonalbiblioteket 3.3.2021

Utgangspunktet:

Se også:
https://www.riksrevisjonen.no/rapporter-mappe/no-2020-2021/undersokelse-av-politiets-innsats-mot-kriminalitet-ved-bruk-av-ikt/

ML brukt på tekst – langt ifra noe nytt

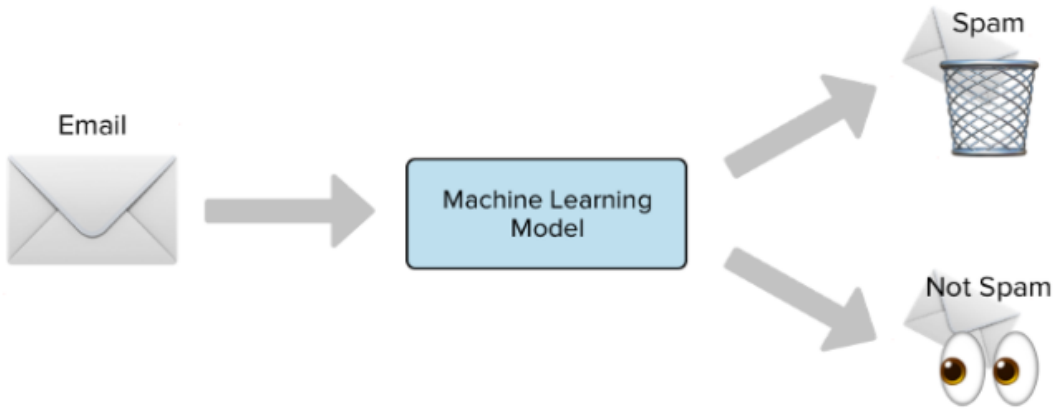
F.eks. lenge brukt i spamfiltre til epost

Spam Mail Detection Using Support Vector Machine.

Shreyak Follow

Aug 5, 2020 · 3 min read

In this blog, we are going to classify emails into Spam and Anti Spam. Here I have used SVM Machine Learning Model for that.



Kilde: https://becominghuman.ai/spam-mail-detection-using-support-vector-machine-cdb57b0d62a8

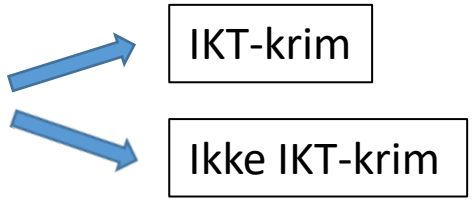
Det vi (ikke jeg da) har gjort

- Ligner relativt mye på spam-filter eksemplet

- Spørsmålet vi ønsket svar på:
  - «Av alle anmeldte kriminalsaker – hvor mange saker handler om IKT-krim?»

  - Et klassisk (binært) klassifiseringsproblem

    - 300 000+ saker → ML-algoritme



Alternative algoritmer testet

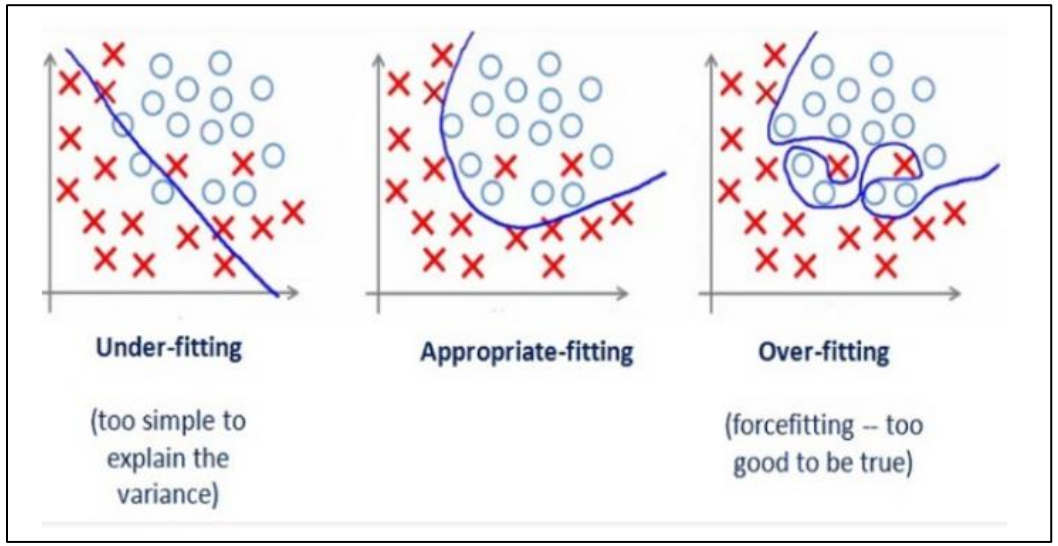
- Naive Bayes (dårlig resultat)
- Random Forest (sterkt overtrent)
- XGBoost (sterkt overtrent)
- **Support Vector Machine (SVM) – valgt**
  - SVM er en veiledet, black box ML-modell

- Test alltid flere alternativer → ikke gitt hva som er det beste for deg

En note om «overtrening»

- Overtrening et kronisk problem når du driver med veiledet ML

- Enkelt sagt:
  - Du lager (trener) en klassifiseringsmodell på basis av noen kjente (trenings-)data
  - For å predikere klassetilhørighet for nye, ukjente data
  - Alltid risiko for at modellen er «for godt» tilpasset treningsdata → gir dårlig prediksjonskraft



Kilde: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Trene en modell? Fint, men…

- Hjelp vi har ikke treningsdata!
- Vel, da må lage treningsdata…

Manuell klassifisering av 1072 saker

1. Uthenting av dokumenter fra totalt 334 544 saker og til sammen 396 917 anmeldelsesdokumenter

2. Uttrekk av 1072 tilfeldig valgte saker, for manuell gjennomgang og klassifisering

Resultat: 1072 manuelt klassifiserte saker. Dette ble treningsdata

ML-klassifisering

- Datagrunnlag → anmeldelsestekst, saksbeskrivelse mv.
- Standard datavasking: tokenization, stemming etc.
- Noen ord vil være spesifikke/opptre oftere for IKT-krim saker
- Ord er således variablene som definerer prediksjonen

Valg av ord/variabler/features

- Noen ord er altså mer vanlig for IKTkrim, enn for ikke-IKTkrim
- Vekting av ord vha. TF-IDF (basert på treningsdata)
  - (Term Frequency – Inverse Document Frequency)
- Tok 150 ord med høyest vekt fra hver klasse, fjernet ord som var felles (fra de 1072 treningssakene)
- Valgte 70 ord (variabler) som hadde størst forskjell i vekt fra de to klassene

Om synonymer – en greie…

- En del ord forekommer sjelden, men er veldig IKT-spesifikke (f.eks. «hack»)
- Viktige ord (samlet sett), men de får i utg.pkt. liten betydning ettersom de er sjeldne
- Laget derfor en «synonymvariabel» som samlet disse ordene

«Synonym» for å fjerne standardtekst

- Ved OCR scanning får du med standardtekst fra skjemaer
- Laget en «synonym»-variabel som inneholdt standardtekst (setninger)
- «Hvis skjema er brukt, fjern tekst som definert i synonymvariabel»

Synonymer - Generell erfaring

- Viktig å ha en plan for håndtering av synonymer
- Det vil ofte ha stor betydning for hvor god modellen din blir

Trening-test-validering av modell

- Support Vector Machine
- 70+ variabler
- 1072 manuelt klassifiserte saker som treningsdata
- Brukte 5-fold kryssvalidering (944 saker til trening/test, 137 saker for å validere modellen)

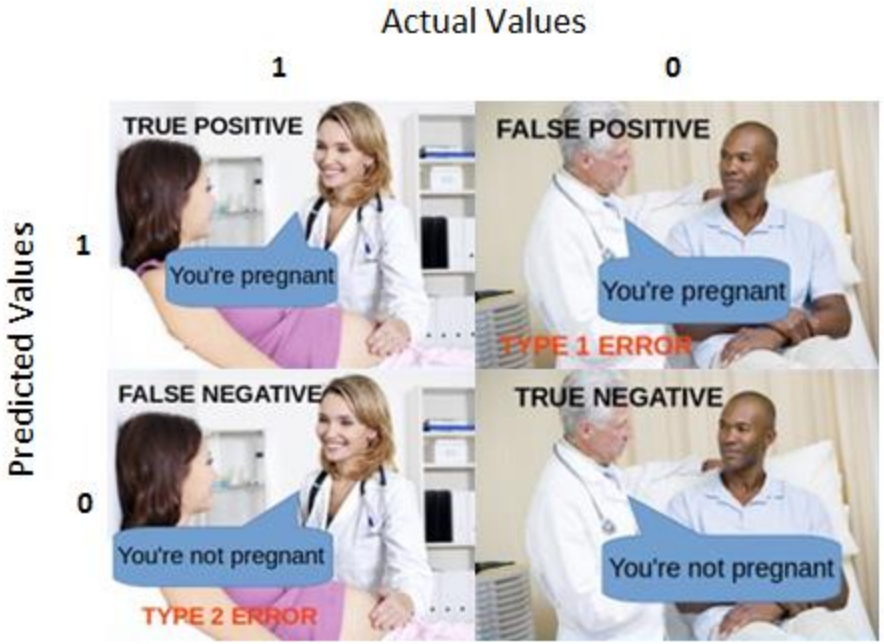
Ferdigtrent modell

- Kjørt på hele populasjonen (300 000+ saker)
- De sakene som ligner mest på IKT-krim tekstmessig…
- Legger da modellen i «IKT-krim boksen»

Resultater…?

- Finnes ulike mål på hvor god en modell er

- Og ingen modell er perfekt (Jf. «modell»)



https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

- F.eks. graviditet – hva er verst?
  - Å få beskjed om at du er gravid – uten at du faktisk er det? (falsk positiv)
  - Å få beskjed om at du ikke er gravid – når du faktisk er det? (falsk negativ)

- Så: Hvilket mål skal du bruke?

Matthews korrelasjonskoeffisient (MCC)

- MCC = 1
  - Perfekt samsvar mellom prediksjon og realitet (alle saker klassifisert riktig)
- MCC = 0
  - Modellen er like god (dårlig) som myntkast
- MCC = -1
  - 100 % av sakene er feilklassifisert

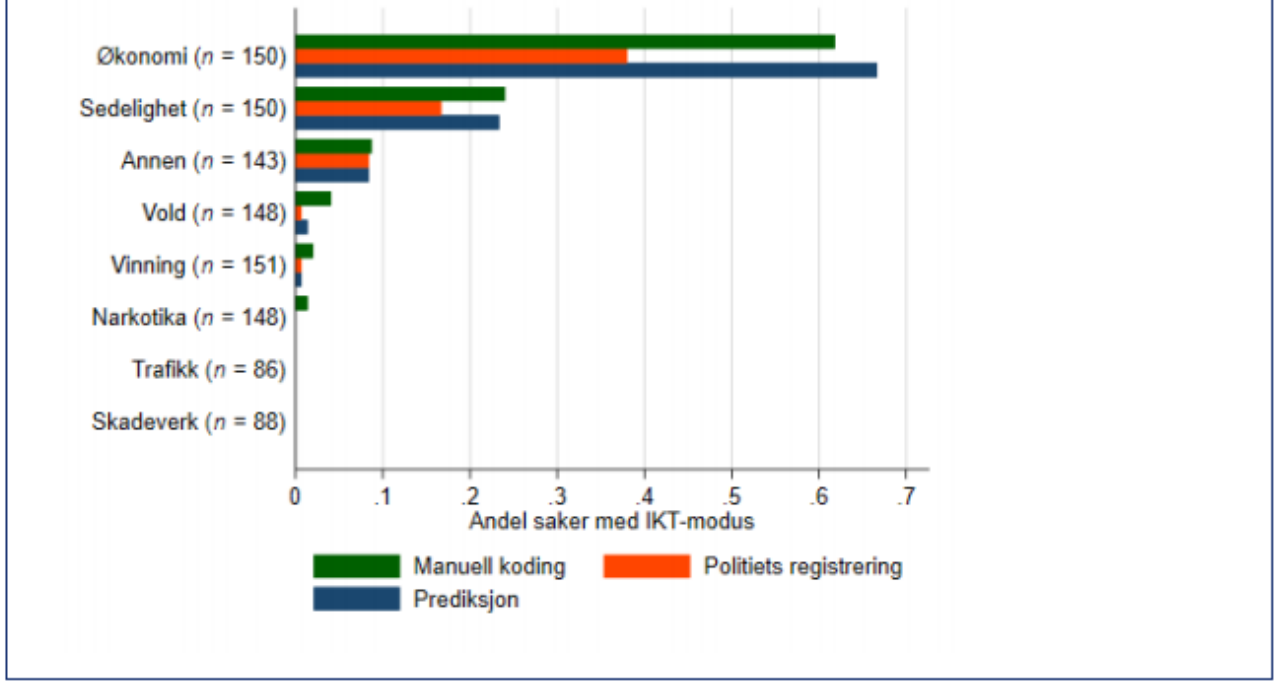
$$MCC = \frac{TN \times TP - FP \times FN}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}}$$

- Resultat MCC for saker om IKT-krim: 0,82

Resultater forts:

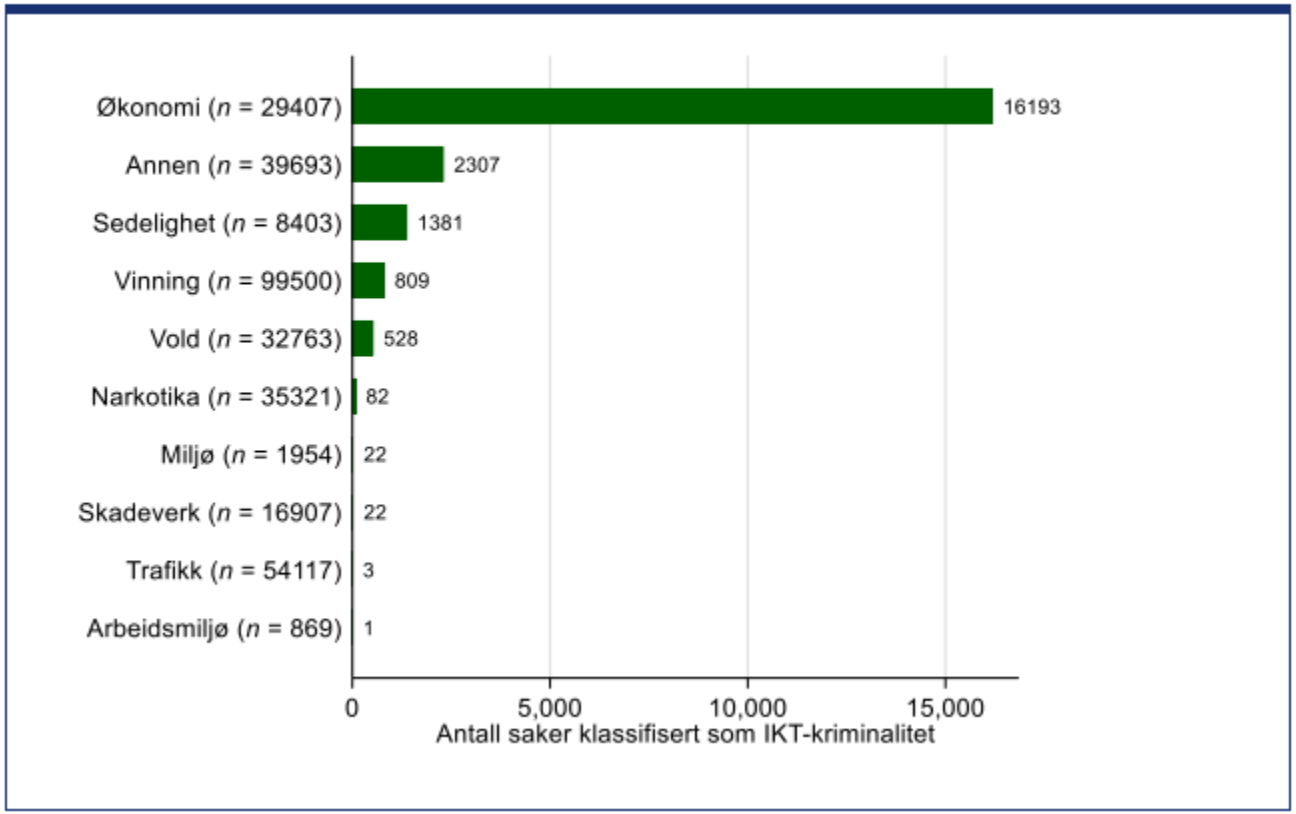
- 21 500 saker av totalt 334 544 ble kategorisert av modellen som IKT-krim.

Figur 1 Andel IKT-kriminalitet for utvalg av saker ifølge manuell koding, politiets modusregistrering og maskinlæringsmodellens prediksjon etter kriminalitetstype* (N=1072)



Og noen flere resultater…

Figur 2 Antall saker registrert i 2018 klassifisert som IKT-kriminalitet av maskinlæringsmodellen etter kriminalitetstype (N=318934)



En kort oppsummering

- Maskinlæring brukt på tekst – et felt i rivende utvikling
- Veldig gøy å jobbe med☺
- Finnes hyggelige folk å snakke med om slikt
- Norsk er et lite språk
  - Vi må fikse ting sjæl?