Titanic survival rate

Gruppe medlemmer:

Astrid Fredlund Bjander

Alexandra Davanger Berdal

Haniyeh Rezaee

Dato: 30.10.2025

BESKRIV PROBLEMET

OMFANG / SCOPE

Vi har utviklet et prosjekt som tar i bruk en binær klassifisering og implementert dette som en brukervennlig nettapplikasjon. Hovedmålet med prosjektet er å gi brukeren en sannsynlighetsbasert prediksjon for overlevelse på Titanic, basert på valgte variabler. Vi løser denne oppgaven ved å benytte maskinlæring på et Titanic-datasett (Titanic – machine learning from disaster). Dette datasettet inneholder passasjerinformasjon som alder, kjønn, billettklasse, antall søsken/ektefeller, antall foreldre/barn, og deres faktiske overlevelsesstatus. Ved hjelp av dette datasettet trener vi en modell til å predikere overlevelse basert på sammenhengen av variablene. Å ta i bruk vår nettapplikasjon er en tidseffektiv måte for Titanic interesserte til å se sannsynligheten for overlevelse, sammenlignet med å utføre alle statistiske beregningene manuelt.

Produktet er rettet mot folk som er interesserte i Titanic og gjerne nysgjerrige på egen sannsynlighet, eller generell sannsynlighet for overlevelse. Ved hjelp av enkle valg mottar brukeren en prediksjon. Nettapplikasjonen synliggjør hvordan enkelte variabler som kjønn eller rikdom (billettklasse), har påvirket overlevelsesraten. Nettapplikasjonen har også mulighet for

utvidelse dersom ønsket. Nå vil den gi en prediksjon basert på sammenhengen av alle variablene, men det er mulig å utvikle slik at vi kan få en prediksjon basert på bare en individuell variabel.

METRIKKER

Systemets suksess måles gjennom en kombinasjon av maskinlæringsmetrikker og software-metrikker. Maskinlæringsmetrikkene måler hvor nøyaktig og pålitelig selve prediksjonsmodellen er. Modellen har en treffsikkerhet på 81%, dette er en høy sannsynlighet og bygger tillit til applikasjonen. Vi tar i bruk en confusion matrix for å beregne presisjon og gjenkalling. Software-metrikkene måler nettapplikasjonens tekniske ytelse og brukeropplevelse. Ved å ta en latensavlesing fra nettapplikasjonen ser vi en tidsangivelse på 1.28 sekund og en indikasjon på 178 millisekund. Den totale tiden for nettapplikasjonen å avgi en prediksjon er da litt over ett sekund, noe som ikke er så gunstig med tanke på at det er et såpass "lett" modell.

DATA

Vi har brukt et datasett fra Kaggle, og disse dataene kommer fra den historiske passasjerlisten til RMS Titanic. Dette datasettet inneholder totalt 891 rader og som blir delt i et treningssett (med 712 rader) og et testsett (med 179 rader), der hver rad tilhører en passasjer. Hver passasjer har noen variabler som navn, alder, kjønn, hvilken klasse osv. Disse datasettene inneholder både kategorisk datatype (nominal) og numerisk (kontinuerlig og diskret). I maskinlæring må vi finne en passe mengde med data slik at vi unngår enten underfitting eller overfitting. Dataene våre er historisk data, og er hovedsakelig lukket. Det er liten sannsynlighet for at det blir lagt til mer data om overlevende eller avdøde da det er såpass lenge siden ulykken fant sted. Vi kan da trene maskinen vår på forskjellige metoder, men selve dataen vil ikke utvikle seg.

Problemet vårt er håndtert med Supervised Learning, der oppgaven her er en binær klassifisering. Vi bruker Random Forest-modell for maskinlæringen. Den tar inputvariablene og ser sammenhenger mellom dem og forteller til oss om vi hadde overlevd Titanic eller ikke.

Ettersom 705 av personene på Titanic som overlevde var 675 kvinner og barn så har kjønn og alder en stor påvirkning på om personen hadde overlevd eller ikke. Datasettet her har allerede fått label survived, hvis passasjerer har overlevd får de 1, og hvis ikke får de 0.

Vår data er ganske konsistent, og vi har all den informasjonen som vi trenger for å predikere overlevelse. Vi har noen rader som mangler informasjon for alder eller lugar, men vi kan gjør dataen mer konsistent ved å enten slette de variablene, eller erstatte dem med estimerte verdier (imputing). Dataene våre er offentlig informasjon, og vi må ikke ta et spesielt hensyn til personvern selv om dataene baserer seg på ekte mennesker. Listen over passasjerer på Titanic er offentliggjort, og Titanic ulykken er nå blitt en historisk tragedie.

Maskinen har mest lyst på numeriske verdier så derfor omgjør vi de kategoriske verdiene som kjønn og påstigning til numeriske verdier ved bruk av one-hot encoding. For eksempel gjør vi kjønn binært ved å si at 1 representerer kvinner og 0 representerer menn. Vi må også fjerne eller bruke imputing for de dataene som mangler som alder og lugar. Alder har en viktig påvirkning på utfallene (om vi hadde overlevd eller ikke) så det er viktig å ikke slette den, og heller bruke imputing og sette en gjennomsnitts alder for de manglende rader. Vi dropper de variablene som vi har ikke noe bruk for eller som har alt for mange manglende verdier som for. eks Lugar, Billett nummer, Navn og PassajerId;

MODELLERING

Til datasettet fokuserte vi først på logistisk regresjon for en enkel baseline modell. Siden logistisk regresjon egner seg godt for binær klassifisering, testet vi denne modellen først på datasettet. Vi skal klassifisere og finne ut sannsynligheten for overlevelse, der vi har to utfall: overlevelse eller død. Grunnen til valg av logistisk regresjon for baseline ytelse, er at den er enkel på slike datasett, og lett å tolke. Logistisk regresjon ga en nøyaktighet på: 79.89%

Så testet vi Random forest classifier, for en mer fleksibel løsning, da Random forest classifier fanger opp ikke-linære mønstre slik som kjønn, alder og klasse. Dermed fant vi at random forest

klassifier ga en høyere nøyaktighet på: 81.56%. Noe som gir oss en forbedring på baseline ytelsen.

For å undersøke feil-prediksjon vil vi ta utgangspunkt i en confusion matrix og klassifikasjonsrapport. Dette kan gi oss et bedre bilde over prediksjonene og hvilke data som blir klassifisert feil. Også ved å analysere feature importance kan vi eventuelt finne ut hvilke faktorer som kan ende opp med å påvirke prediksjonene mest. Slik som for eksempel, kjønn, billettpris eller klasse. Ved å utføre dette til veien videre, kan vi da få et utgangspunkt i hvordan vi kan forbedre modellen, og forstå de ulike faktroene som kan være med på å påvirke modellen vår.

DEPLOYMENT

Modellen kan settes i drift og brukes som en måte for å beregne sannsynligheten for overlevelse, basert på kjønn, alder, billetter, klasse. osv. Prediksjonene kan brukes til å bergene og analysere hvilke faktorer som påvirker overlevelse, eller for å vise sjansen for overlevelse. Siden Titanic ulykken skjedde for over 100 år siden, er det lite sannsynlig at det vil trengs vedlikehold og monitorering. Dette gjelder også for oppdatering av nye data, ettersom at all relevant informasjon fra ulykken allerede er finnes.

For å forbedre systemet etter at det er satt i drift, er det mulig å videreutvikle brukerinteraksjonene. Nå er nettapplikasjonen slik at vi må fylle inn alle variablene for å få en binær verdi, enten overlevd eller ikke. Vi kunne lagt til muligheten til å velge bare en variabel, og heller se en sannsynlighet basert på denne. Ved å utvide applikasjonen slik vil den også være mer lærerik for brukerne da du får dypere innsikt i hvordan, kjønn eller rikdom påvirket overlevelse.

5: REFERANSER

https://www.kaggle.com/competitions/titanic/data?select=train.csv