Soppklassifiserings app

Johan Levent Gencher / Gruppe 16, 15.11.2024

Link: https://mushroom-classification-app-gyhlspddxfdkxhhnwcigcd.streamlit.app/

Scope

- **"Business objective"**: Målet med prosjektet er å utvikle en maskinlæringsbasert app som kan finne ut om en sopp er giftig eller spiselig basert på soppens egenskaper.
- **Bruk av løsningen**: Løsningen vil være en nettapplikasjon der brukere kan velge egenskaper for en sopp og få et umiddelbart svar på om den er spiselig eller giftig.
- **Måling av ytelse**: Ytelsen til modellen skal måles ved hjelp av nøyaktighet og treffsikkerhet. "Business metric" vurderes som akseptabel når modellen oppnår minst 95% nøyaktighet.
- **Systemkomponenter**: Modellen er en del av et større system som inkluderer en frontend bygd med Streamlit for interaktiv bruk.
- **Stakeholders**: Soppplukkere, forskere innen mykologi og allmennheten som er interessert i matsikkerhet.
- **Tidslinje**: Prosjektet ble gjennomført i løpet av tre dager, med milepæler som inkluderer modelltrening, testing og implementering i en webapplikasjon.
- **Ressurser**: Python-programmering, bibliotekene scikit-learn, pandas, streamlit og UC Irvine Machine Learning Repository ble brukt. Beregningsressursene var tilstrekkelige med lokal maskin.

METRIKKER

- **"Business metric" minimum**: Minimum nøyaktighet på 95% for at løsningen skal anses som vellykket.
- **Software-metrikker**: Nøyaktighet (accuracy), presisjon (precision), recall og f1-score. Disse metrikkene ble brukt for å evaluere modellens ytelse og sikre at modellen oppfyller forretningsmålene.
- **Sammenheng med "business objective"**: Høy nøyaktighet er viktig for å redusere risikoen for feilklassifisering som kan føre til helseskader.

DATA

- **Datakilder**: Datasettet ble hentet fra UC Irvine Machine Learning Repository og inneholder soppklassifiseringer basert på 22 egenskaper.
- **Datatype**: Kategoriske data med egenskaper som soppform, farge, lukt og habitat.
- **Antall data**: Datasettet inneholder totalt 8124 eksempler, hvorav 4208 er spiselige og 3916 er giftige.
- **Personvernhensyn**: Ingen personvernhensyn var relevante for datasettet.
- **Datarepresentasjon**: Data ble representert som kategoriske variabler som ble kodet numerisk ved hjelp av LabelEncoder for bruk i maskinlæringsmodellen.
- **Dataprosessering**: Det var nødvendig å håndtere manglende verdier i kolonnen stalk-root ved å fylle inn med en kategori missing.

MODELLERING

- **Utforskede modeller**: En beslutningstre-modell (Decision Tree Classifier) ble brukt for sin enkle implementasjon og forståelige beslutningslogikk.
- **Baseline-ytelse**: Første baseline ble målt ved å bruke en enkel tren/test-splitt med en nøyaktighet på 100%
- **Evaluering av feil**: Feilprediksjoner ble analysert ved å bruke klassifiseringsrapporter og forvirringsmatriser.
- **Feature importance**: Modellen hadde en nøyaktighet på 100%, noe som kan tyde på at dataen den trente på var for lik, eller at den vil fungere dårligere på ny data.

DEPLOYMENT

- **Driftsetting**: Modellen ble satt i drift som en webapplikasjon ved bruk av Streamlit, slik at brukere kunne interaktivt klassifisere sopp ved å velge egenskaper i grensesnittet.
- **Plan for monitorering og vedlikehold**: Vedlikehold vil innebære oppdateringer til modellen hvis nye data blir tilgjengelig, samt ytelsestesting for å sikre at modellen fungerer optimalt over tid.
- **Forbedringsplaner**: Fremtidige forbedringer kan inkludere integrasjon av en mer kompleks modell og bruk av flere datakilder for å styrke generaliseringen.

REFERANSER

- UC Irvine Machine Learning Repository (Mushroom Dataset).
- scikit-learn dokumentasjon for modelltrening og evaluering.
- Streamlit dokumentasjon for webapplikasjonsutvikling.
- Chatgpt KI struktur og feilkontroll i webapplikasjonsutvikling.