[Prosjektnavn]

[Shridar Patil], [Dato]

*Malen inneholder en del beskrivende tekst om hva dere kan skrive om under hver overskrift. Dette er kun forslag til elementer: dropp det som ikke er relevant i ditt tilfelle, og legg til element du finner relevant. Det eneste kravet er at beskrivelsen er i tråd med livssyklusen til maskinlæringsprosjekt skissert i kurset. Det vil si, at prosjektbeskrivelsen følger strukturen reflektert i overskriftene nedenfor. Slett alle instruksjoner (tekst i kursiv) før innlevering.*

# **BESKRIV PROBLEMET**

## **SCOPE**

* *“Business objective”: Beskriv prosjektets mål og “business impact”.*

Målet med prosjektet er at det skal tilby en KI-tjeneste som planlegger dates basert på brukeren prompt der brukeren fritt spesifiserer ønsker til planleggingen for å få en så skreddersydd date plan som mulig for å øke den potensielle suksessen. Ettersom det er en GPT API som blir brukt vil dette produktet være like godt egnet for personer som skal møtes tidlig i et dating forhold og person som har vært sammen lenge, men vil ha forslag eller gi bort ansvaret for planleggingen til en ekstern tjeneste. Det er derfor et stort marked potensielle brukere ettersom det både kan passe «single» personer og etablerte parforhold.

* *Hvordan skal løsningen produsert i prosjektet brukes? Hvilke tilsvarende løsninger finnes i dag / hvordan løses problemet i dag? Hvordan ville du utført / utføres oppgaven manuelt, uten maskinlæring?*

Løsningen er generert ved hjelp av chat GPT sin API-løsning. Dette gir tilgang til chat GPT i utførbar kode slik at programmet kan kalle på GPT modellen og sende den input fra programmet før svaret fra GPT modellen kan dermed integreres tilbake igjen i programmet.

Per i dag finnes det ulike tilnærminger til denne problemstillingen. Den tradisjonelle løsningen er å manuelt planlegge daten selv. Et annet alternativ er å bruke generelle KI modeller som chat GPT eller gemini til å genere svar. Problemet her er at de ikke kan gi konkrete forslag ettersom de er generelle modeller som dermed ikke er spesialisert for denne typen oppgave. Et annet eksempel er <https://dateplanner.ai/> som er en større tjeneste so en slags «date manager» som har oversikt over flere dates og lignende.

Vår løsning søker å være en enkel men robust, ved at den kun fokuserer på å gi en skisse til en ide og ikke påtar seg mer ansvar enn å komme med forslag. Dermed blir programmet relativt enkelt, og videre krever ikke mange ytterligere funksjoner utover sin hovedoppgave som er KI modellen.

* *Hvordan skal ytelsen måles via “business metrics”?*

Ytelsen vil trolig her bli målt etter brukertilfredshet og bruk. Hva et godt date forslag er vil være subjektivt av natur og derfor er det vanskelig å bruke skåringsverktøy utover brukerens rapportering av tjenesten (hvor fornøyd de er, vil de anbefale det til en venn, tenker du å bruke tjenesten igjen?). Sammen vil dette vil data rundt tjeneste trafikken også være relevant for å se hvor godt produktet fungerer ettersom høyere etterspørsel vil vise til at produktet funker.

* *Hvis din maskinlæringsmodell skal være del av en større “pipeline” eller system, beskriv komponentene i dette. Vurder hvordan endringer i én del av systemet vil kunne påvirke andre.*

Modellen inngår i et større skript. Før tjenesten blir lansert har vi samlet inn relevant data om resturanter, aktiviteter og lignende. Disse har dermed blitt kjørt igjennom Open Ai sin embedding modell slik at vi har en tabell GPT modellen kan søke i, sammen med dette blir også værdata fra YR sin API også anvendt for å ha værmelding for de kommende dagene.

Disse inn dataene blir kombinert med brukerens input som embeedes og dermed blir brukt i et likhets søk ved hjelp av funksjonen «cosinus\_similiarty» fra sk\_learn. Dette gir dermed relevante forslag som legges til inputet til chat GPT for å optimalisere svaret. Svaret blir dermed gitt tilbake og dette sammen med brukerens prompt blir lagret i en brukerchat bygd på Gradio for at brukeren kan se hele samtalehistorikken for den gjeldende sessionen.

* *Beskriv stakeholders i prosjektet.*

Stakeholderen på utvikler siden er

* Forfatteren av denne rapporten

Tjenester brukt av produktutivkleren

* Open Ai sin API for chat GPT og embedding modellen
* YR sin API for henting av værdata

Brukeren selv som anvender tjensten

* *Beskriv en tentativ tidslinje for prosjektet med milepæler*

Delmål 1:

* Etablere en enkel terminal basert GPT modell som har et egnet systemprompt for oppgaven

Delmål 2:

* Koble terminal GPTen til en visuell nettside ved for eksempel bruk av Gradio bibloteket. Videre få meldingshistorikken til å lagres i denne

Delmål 3:

* Samle og embed en database som GPT modellen kan bruke til å ytterligere forbedre sine svar til brukeren

Delmål 4:

* Koble opp YR sin API for å få videre vær data for å kunne tilpasse forslagene ytterligere

Delmål 5:

* Polish?

* *Definer hvilke ressurser, for eksempel beregningsressurser og personell, som behøves for å gjennomføre prosjektet.*

Prosjektet behøver følgende ressurser

* API fra Open AI som utfører embedding og forespørsler til chat GPT. Dette forgår på deres datasentre og utvikler betaler en som for antall tokens brukt.
* YR sin API lagrer værdata som kan hentes ved en API requrest. For å få denne dataen er det derfor nødvendig at denne tjenesten fungerer.
* Hosting tjenste: For å kjøre nettsiden med date planleggeren, må det være en dedikert datamaskin som kjører denne slik at den holdes åpen kontinuerlig og kan håndtere eventuelle forspørsler fra brukerene.

## **METRIKKER**

* *Beskriv minimum “business metric”-ytelse for at prosjektet skal være en suksess.*
* *Beskriv maskinlærings- og software-metrikker eller mål som skal brukes til å måle hvorvidt systemet/løsningen fungerer. Her bør du bruke metrikker som enkelt kan regnes ut. Treffsikkerhet (accuracy), kvadratisk avvik (“mean squared error”), latency, throughput, etc. Forklar hvordan disse henger sammen med “business objective” skissert over.*

# **DATA**

*Beskriv hvilke data og labels som skal brukes. Hva slags type data er dette? Hvor får du tak i data / hvordan kan data samles inn? Hvor mye data er tilgjengelig nå og hvor mye data estimerer du at behøves? Hvis problemet skal håndteres med “supervised learning”-metoder, beskriv hvordan du skal få tak i labels. Hvordan sikre at labels er tilstrekkelig konsistente?*

Data som brukeres er primært strenger eller tekst objekter mellom brukeren og KI modellen. Dette vil være forespørslene og svarene modellen får underveis i prosseseringen av informasjonen.

Videre brukes også en embedded fil. Dette er en fil der teksten har blitt tokenifisert til en vektor representasjon. Dette er en tabell fylt med kolonner og rader av tall som modellen kan bruke til å representere teksten. To slike filer kan dermed bli undersøkt for likheter og dermed brukes til å finne tokens som ligner og dermed sannsynlig matcher.

YR sine data er videre en tabell med både numeriske og kategoriske data som forteller om været forespeielet. Denne tabellen er langt mindre og organiserer dataen for den forspeilede værmeldingen for de kommende dagene for temperatur, vind, sannsynlighet for regn og skydekke etc.

*Beskriv eventuelle personvernhensyn eller andre relevante etiske betraktninger.*

Ingen data blir lagret av utvikleren under denne prossen. Det er derimot relevant å nevne at dataene blir til sendt Open Ai. Det er derimot ingen sensetiv data som anvendes i denne prossesen. Brukeren står fritt til å begrense hvor mye informasjon de vil oppgi og det oppgis sjeldent informasjon so er personsenstivt. Videre er det ingen bruker, mail eller lignende som brukes som kan knyttes hver session til en enkeltbruker.

*Hvordan skal data representeres for maskinlæringsmodellene? Beskriv eventuelle behov for rensing av data, feature engineering og skalering.*

GPT modellen mottar embedded data fra brukeren og for databasen med relevante forslag for dating. Dette gjøres slik at det kan utføres matching mellom inputet og databasen for å kunne gi mer relevante forslag. Utvoer dette arbeider programmet med data i den samme formen som den blir opgitt.

# **MODELLERING**

*Beskriv hvilken eller hvilke maskinlæringsmodeller du skal utforske. Beskriv hvordan du planlegger å estimere baseline-ytelse og baseline-oppførsel. Husk at dine første modeller bør være enkle. Basline-ytelse kan avdekkes via modeller eller enkle ikke-maskinlæringsbaserte løsninger, ved å lete etter resultater oppnådd av andre på samme oppgave, eller ved å estimere “human level performance” der det er relevant. Beskriv hvordan du skal undersøke feil-prediksjoner og “feature importance”, og hvordan dette skal brukes til å forbedre modellene.*

Ettersom det skal brukes en LLM modell i dette prosjktet ble det ikke utforsket andre modeller. Det ble gjort kjappe søk av de største LLM modellene som er tilgjenglig på markede, og chat GPT APIen som er tilgjenglig er var enkel og begynne å integerere i et slik prosjekt og ble derfor valgt.

Baseline ytelse for meg ble derfor å bruke standard chat GPT modellen med samme prompt og sammenligne resultatene mellom dette og min egen model.

# **DEPLOYMENT**

*Hvordan skal modellen(e) settes i drift? Hvordan skal prediksjonene brukes? Hva er dine planer for monitorering og vedlikehold av maskinlæringssystemet? Hvis relevant, hvilke planer har du for å forbedre systemet etter at det er satt i drift?*

Modellen skal settes i drift ved å brukes i en nettside som er bygd på Gradio. Denne vil bli satt til public og dermed vær tilgjenglig for andre å få tilgang på. Dermed kan brukere teste nettsiden. Systemet lagrer ikke data etter fanen lukkes og det blir derfor lite aktivt monitorering.

Open Ai har klare retingslinjer som reduserer uakseptabel oppførsel fra modellen og vi som utvikler belager oss på at disse er tilstrekkelige. Dette er ment som et skoleprosjekt og det er derfor begrensinger på hva vi rakk å implementere for dette prosjktet.

* Dette er også grunnen til at prosjektet er begrenset til Bergen ettersom ønsket var at modellen ikke skal bli for krevende på ressurser. Dette ville ha gitt økt kostnad av tokens og større filer for databasen som dermed hadde gjort tjensten ytterligere tregere hvis ikke mer data ressurser blir gjort tilgjenglig som igjen øker prisen ytterligere.

**Eventuelle utbedringer:**

Om dette var et prosjekt som skulle på markedet hadde vi utført følgende utbedringer eller disktuert om de skulle implementeres for et mer ferdig og fullstendig produkt:

* Personlige brukere, som lagrer meldings historikk
  + Kan også oppgi mer informasjon eller kontekst som modellen kan bruke for enda bedre tilbakemeldinger
    - Preferanser for aktiviteter
    - Plassering
* Åpne opp for web crawling som ville gitt gpt modellen mer informasjon som også ville vært enda mer relevant enn en standard database
  + Obs vil øke kostnadene med høyere antall tokens, og evt mer embedding
  + Videre enda mer krevende i forhold til personvern og tilatelser til crawling
  + Vil også gjøre programmet tyngere å kjøre
* Lage en rapporterings tjenste for brukeren
  + Tilbakemeldinger på dårlig eller upassende respons
  + Kan dermed videre optimalisere tjensten
  + Ønske om funksjoner ikke tilgjenglig per nå

# **REFERANSER**

*List opp kilder du har brukt til å utarbeide prosjektbeskrivelsen. Listen av referanser bør sannsynliggjøre at prosjektet kan lykkes (“feasability”)*