

# Energikarakter

David Bao Bui, Christoffer Toftevang, Henrik Bloch Helmers, [02.11.2025]

## 1: BESKRIV PROBLEMET

Vårt mål i dette prosjektet er å bruke maskinlæring til å forutsi forskjellige hus sitt energimerket. Vi mener at dette er et relevant problem siden nå til dags så er det vanskeligere å få energimerke sitt, samtidig som det har blitt dyrere. I år 2025 sank støtten for energikartlegging/ rådgivning fra en fastpris på 5000kr til å bare dekke 25% av totalsummen opp til 500kr.

Hvordan produktet kommer til å funke er ved at kunden setter inn informasjonen til huset som vil gjøre at produktet vil returnere en energimerke karakter til kunden.

I dag så bruker de fleste Enova for å sjekke energimerke manuelt, men de krever at du logger inn og at du bare kan sjekke dine egne boliger. Vår løsning trenger du ikke å logge inn og er ikke en Enova konkurrent, men et støtteverktøy som kan bli brukt hvis du ønsker å finne energimerking forttere eller for flere hus.

## METRIKKER

Feilundersøkelse av modellen ble gjort gjennom confusion matrix og klassifikasjonsrapport (precision, recall og F1-score). Ut i fra dette kunne vi se hvilke klasser som var vanskeligst å skille mellom. I vårt tilfelle så vi at de midtre klassene (3-5), var de vanskeligste klassene å skille seg mellom.

Informasjonen vi samlet fra confusion matrix og klassifikasjonsrapporten ble brukt til å endre hyperparameterene til modellen som `n_estimators`, `max_depth` og videre utvikling av inputverdier.

## 2: DATA

Vår data kommer fra Enova sitt energimerke datasett, hvor de har informasjonen vi trenger for å få modellen til å returnere et resultat. Dataen inneholder blant annet postnummer, årstall, hustype, fossilandel og materialvalg til huset, som modellen bruker for å komme til et resultat. Labelsene som blir brukt for å måle er hovedsakelig karakterer fra A til G, men vi har bestemt oss for å velge tall fra 1-7 istedenfor på grunn av at vi fant ut at treffsikkerheten økte, og gjorde det enklere å jobbe med modellene.

Dataen brukt i vår løsning er også ikke basert på enkeltpersoner, men bare bygninger som vil si at den er ikke-personidentifiserbar. Dette vil si at sensitiv informasjon som navn og fødselsdatoer ikke blir tatt i bruk.

### 3: MODELLERING

Modellen: <https://huggingface.co/spaces/chrtof13/energikarakter-modell>

I prosjektet vårt har vi valgt å bruke to forskjellige maskinlæringsmodeller for klassifisering av energikarakter: SGDClassifier og RandomForestClassifier. Grunnen til at vi valgte disse modellene var å sammenligne en enkel lineær modell med en mer kompleks modell for å se hvilke som egner seg best til vårt brukstilfelle.

Vi begynte med å trene SGDClassifier. Dette er en enkel modell som kan gi oss innsikt i om vi trenger å benytte oss av en mer avansert modell. Denne modellen fikk en accuracy på ca 46%, noe som gav oss grunnlaget for å vurdere en mer avansert modell.

Vi gikk over på RandomForestClassifier, som kombinerer mange beslutningstrær for å redusere varians og øke nøyaktigheten. Denne modellen gav oss en betraktelig bedre nøyaktighet på rundt 62% accuracy etter justering og forskning på forskjellige hyperparametere. Dette viste at RandomForestClassifier var den riktige modellen for oss basert på at vi trenger å finne sammenhenger mellom variabler som ikke er lineære.

For å forstå hvordan modellen “tenker” og hvor godt modellen faktisk fungerer, undersøkte vi feature importance. I starten av prosessen fant vi ut at “beregnetleverttotaltkwhm2” hadde desidert størst innvirkning på modellens resultater, dette skyldes at energikarakteren er direkte basert på denne verdien, det er en del av selve beregningen av energikarakteren. Å bruke denne verdien vil derfor være et eksempel på data leakage, der modellen får tilgang til informasjon som ikke ville vært tilgjengelig i en prediksjonssituasjon.

Vi ønsket en modell som lærte seg mønstre basert på faktiske bygningsegenskaper, derfor fjernet vi “beregnetleverttotaltkwhm2” fra datasettet. Vi valgte likevel å beholde “beregnetfossilandel” i modellen, selv om denne verdien typisk beregnes av fagfolk basert på bygningens oppvarmingssystem. Det er verdt å merke seg at i et praktisk bruksperspektiv kunne denne featuren vært ekskludert, da den ikke nødvendigvis er tilgjengelig for privatpersoner som ønsker å estimere energikarakteren på sitt eget hus.

## 4: DEPLOYMENT

Når modellen er ferdig trent, så skal vi sette den i drift som en webapplikasjon i huggingface. Kunden kan da bruke siden til å oppgi postnummer, årstall, hustype, fossilandel og materialvalg til huset. Modellen vil da returnere en liste av predikert karakter som viser karakteren huset har fra 1-7, hvor 1 tilsvarer en G og 7 tilsvarer en A.

Predikasjonene blir loggført sammen med inputverdiene og overvåkes ved å se på latency og tilgjengelighet. Etterhvert hvis det kommer mer informasjon fra flere datasett så vil vi trene modellen på nytt for å integrere den nye informasjonen.

På grunn av at vi jobber med prosjektet via github så har vi muligheten til å rolleback til tidligere versjoner hvis det trengs.

Vår plan videre er å fortsette å overvåke modellen, og endre hvis det trengs. Samtidig så må vi prøve å jobbe mer med modellen for å få høyere accuracy, bredere datagrunnlag og muligens implementere aktiv læring.

## 5: REFERANSER

*Enova SF. (n.d) Energimerkesystemet - Offentlig data: Fil\_EnergimerkerFileByYear [API-dokumentasjon]. (Versjon V1). Enova. Hentet 2. November 2025, fra: <https://data.norge.no/nb/datasets/6e841199-64b4-36d4-afd2-3d8054e2f96c/energimeking-av-boliger-og-yrkesbygg?tab=overview>*

*GitHub-repo for ML-delen - DAT158. laget av: Alexander Lundervold og Steffen Maeland. Hentet 2. November 2025, fra: <https://github.com/HVL-ML/DAT158>*

*OpenAI. (2025). ChatGPT (GPT-4.5) [LLM]. <https://chat.openai.com>*

*Anthropic. (2025). Claude (Claude 3.5 Sonnet) [LLM]. <https://claude.ai/>*