Evaluering av ML-modeller

ELMED219: Momentliste E01–E10

ELMED219

Vår 2026

Oversikt

1. **Confusion Matrix** (forvirringsmatrise)
   - E01–E02: TP, TN, FP, FN
2. **Grunnleggende metrikker**
   - E03: Accuracy (nøyaktighet)
   - E04: Precision (presisjon)
   - E05: Recall / Sensitivity (sensitivitet)
   - E06: Specificity (spesifisitet)
3. **Avanserte metrikker**
   - E07: F1-score
   - E08: ROC-kurven og AUC
   - E09: Ubalanserte datasett
   - E10: TRIPOD-retningslinjer

E01: Tolke en Confusion Matrix (forvirringsmatrise)

Hva er det?

Tabell: predikerte vs. faktiske klasser

Faktisk \ Predikert	Positiv	Negativ
Pos	**TP**	**FN**
Neg	**FP**	**TN**

Avlesning:

- **TP**: Korrekt positiv
- **TN**: Korrekt negativ
- **FP**: Falsk alarm
- **FN**: Oversett tilfelle

Eksempel: Kreftscreening

- TP = Kreft oppdaget korrekt
- FN = Oversett kreft (*farlig!*)

E02: TP, TN, FP, FN i medisinsk kontekst

Term	Definisjon	Medisinsk eksempel
TP	Syk pasient, modell sier syk	Kreft korrekt identifisert
TN	Frisk pasient, modell sier frisk	Frisk person bekreftet frisk
FP	Frisk pasient, modell sier syk	Unødvendig biopsi
FN	Syk pasient, modell sier frisk	Oversett kreft

Kritisk spørsmål i medisin

Hva er verst: **Falsk positiv** (FP) eller **Falsk negativ** (FN)?

- Screening: FN er farligere (oversett sykdom)
- Invasiv behandling: FP kan være farligere (unødvendig risiko)

E03: Accuracy (nøyaktighet)

Definisjon

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{Korrekte prediksjoner}}{\text{Alle prediksjoner}}$$

Eksempel:

85	10
5	900

$$\text{Accuracy} = \frac{85 + 900}{85 + 900 + 10 + 5} = \frac{985}{1000} = 98.5\%$$

Advarsel

Høy accuracy kan være misvisende ved ubalanserte datasett!

E04: Precision (presisjon / PPV)

Definisjon

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Sanne positive}}{\text{Alle predikerte positive}}$$

Spørsmål: "Av alle modellen sier er positive, hvor mange er faktisk positive?"

Høy precision viktig når:

- Kostbar oppfølging av positive
- Vil unngå falske alarmer
- Eksempel: Sjelden sykdom

Eksempel:

$$\text{Precision} = \frac{85}{85 + 5} = 94.4\%$$

"94% av flaggede pasienter er faktisk syke"

E05: Recall / Sensitivity (sensitivitet)

Definisjon

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Sanne positive}}{\text{Alle faktisk positive}}$$

Spørsmål: "Av alle som faktisk er positive, hvor mange fanger modellen opp?"

Høy recall viktig når:

- Alvorlig konsekvens av å overse (FN)
- Screening for farlig sykdom
- Eksempel: Kreftscreening

Eksempel:

$$\text{Recall} = \frac{85}{85 + 10} = 89.5\%$$

"Vi fanger opp 89.5% av alle krefttilfeller"

Andre navn

Recall = Sensitivity = True Positive Rate (TPR)

E06: Specificity (spesifisitet)

Definisjon

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{\text{Sanne negative}}{\text{Alle faktisk negative}}$$

Spørsmål: "Av alle som faktisk er negative, hvor mange identifiserer modellen korrekt?"

Høy specificity viktig når:

- Vil unngå unødvendige inngrep
- Kostbar/risikabel behandling
- Eksempel: Bekreftelsetester

Eksempel:

$$\text{Specificity} = \frac{900}{900 + 5} = 99.4\%$$

"99.4% av friske korrekt identifisert som friske"

Trade-off

Høy sensitivity ↔ Lav specificity (og omvendt). Justering av terskelverdi påvirker begge.

E07: F1-score

Definisjon

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Harmonisk gjennomsnitt av precision og recall

Hvorfor harmonisk gjennomsnitt?

- Straffer ekstreme verdier hardere enn aritmetisk gjennomsnitt
- Precision = 100%, Recall = 0% $\rightarrow$ F1 = 0% (ikke 50%!)

Når bruke F1?

- Ubalanserte datasett
- Når både FP og FN er viktige
- Sammenligne modeller

Eksempel:

$$F1 = 2 \cdot \frac{0.944 \cdot 0.895}{0.944 + 0.895} = 0.92$$

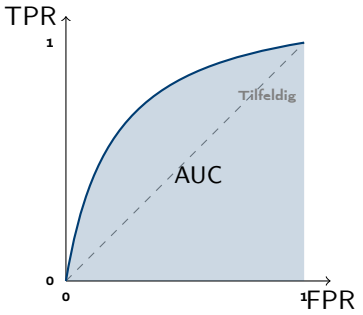
E08: ROC-kurven og AUC

ROC-kurve:

- **R**eceiver **O**perating **C**haracteristic
- Plotter TPR vs. FPR for ulike terskler
- Viser trade-off mellom sensitivity og specificity

AUC (Area Under Curve):

- Arealet under ROC-kurven
- Verdi mellom 0 og 1
- AUC = 0.5: tilfeldig gjetting
- AUC = 1.0: perfekt klassifikator



Tommelfingerregel

AUC $>$ 0.9: Utmerket AUC 0.8–0.9: God AUC 0.7–0.8: OK AUC $<$ 0.7: Svak

E09: Når accuracy er utilstrekkelig

Problemet med ubalanserte datasett:

Eksempel: 1000 pasienter

- 950 friske, 50 syke
- Modell predikerer ALLE som friske

Resultater:

$$\text{Accuracy} = \frac{950}{1000} = 95\%$$

$$\text{Recall} = \frac{0}{50} = 0\%$$

Høy accuracy, men ubrukelig modell!

Løsninger:

- Bruk F1-score eller AUC
- Fokuser på recall for screening
- Rapporter alle metrikker
- Sammenlign med baseline

Teknikker for ubalanse:

- Oversampling (SMOTE)
- Undersampling
- Class weights
- Stratifisert splitting

E10: TRIPOD-retningslinjer

Hva er TRIPOD?

Transparent **R**eporting of a multivariable prediction model for **I**ndividual **P**rognosis **O**r **D**iagnosis

Hovedpunkter for rapportering av ML i medisin:

1. Tydelig beskrivelse av studiepopulasjon
2. Definer utfall og prediktorer klart
3. Beskriv missing data håndtering
4. Rapporter modellutvikling detaljert
5. Intern validering (kryssvalidering)
6. Ekstern validering hvis mulig
7. Rapporter kalibrering og diskriminering
8. Diskuter begrensninger

Hvorfor viktig?

TRIPOD sikrer **reproduserbarhet** og **kvalitetskontroll** av prediksjonsmodeller i helseforskning.

Oppsummering: E01–E10

Confusion Matrix

TP	FN
FP	TN

Metrikker:

Recall $= \frac{TP}{TP+FN}$

Precision $= \frac{TP}{TP+FP}$

Specificity $= \frac{TN}{TN+FP}$

Avansert:

F1-score:
Kombinerer Precision + Recall

ROC/AUC:
Helhetsbilde uavhengig av terskel

Nøkkelpunkter

- Velg metrikk basert på medisinsk kontekst (FN vs. FP konsekvenser)
- Accuracy er ofte utilstrekkelig – bruk F1, AUC, precision, recall
- ROC/AUC gir helhetsbilde uavhengig av terskelverdi
- Følg TRIPOD for transparent rapportering