Kunskapskontroll 2

Isaac Högfeldt

EC Utbildning

Machine learning kunskapskontroll 2

202503

Abstract

In this report a study was concluded on the MNIST data set for handwritten numbers. Three machine learning models were used, Random Forest, Extra Trees and a voting classifier of the two previously mentioned. The entire training dataset was 50 000 images, validation and test dataset was 10 000 images each. Extra Trees showed best accuracy on the test set, with an accuracy of 96,82%.

Innehållsförteckning

1 Inledning

Handskriven text är fortfarande en utmaning för populära appar som sägs kunna läsa av bilder för att sedan konverteras till digital text. Denna rapport undersöker och jämför prestandan hos olika maskininlärningsmodeller för denna klassificerings uppgift.

1.1 Syfte och Mål

- Utvärdera och jämföra olika maskininlärningsmodeller effektivitet för sifferigenkänning
- Undersöka om kombinerade modeller (Voting Classifier) kan förbättra resultaten

2 Teori

2.1 Maskininlärning

- **Supervised learning:** Är en maskininlärningsmetod där modellen tränas på en dataset som innehåller både indata (features) och tillhörande korrekta svar (labels). Målet är att modellen ska kunna lära sig sambandet mellan indata och labels för att sedan kunna göra korrekta förutsägelser på ny data. (Wikipedia (a), 2025)
- **Klassificeringsproblem och regrissionsproblem:** Inom maskininlärning delas problem ofta in i två huvudsakliga kategorier: klassificeringsproblem och regressionsproblem. Skillnaden mellan dessa typer av problem är avgörande för att välja rätt maskininlärningsmodell och utvärderingsmetodik.
    - Klassificering: Klassificering innebär att förutsäga en diskret kategori eller klass. Exempel kan vara att klassificera en bild på en handskriven siffra utifrån vilken siffra det är.
    - Regression: Regression handlar istället om att förutsäga ett kontinuerligt numeriskt värde. Exempel kan inkludera att förutsäga huspriser baserat på data som storlek, läge och byggnadsår.

  (Wikipedia (a), 2025)

- **Tränings-, validerings- och test- dataset:** När en dataset används för maskininlärning delas den ofta upp i tre delar:
    - Träningsset**:** Används för att träna modellen och justera dess parametrar.
    - Valideringsset**:** Används för att utvärdera modellen och finjustera hyperparametrar utan att påverka träningsdata. Här finns valet att kombinera tränings- och valideringsdatan för en slutgiltig träning.
    - Testset**:** Hålls separat och används enbart för att slutligt bedöma modellens prestanda, vilket ger en objektiv mätning av dess generaliseringsförmåga på ny data.

  (Wikipedia (a), 2025)

- **Cross-validation:** Är en metod som används för att utvärdera en modells prestanda och säkerställa att den generaliserar bra till ny data. En vanlig variant är k-fold cross-validation, där datasetet delas upp i *k* lika stora delar (folds). Modellen tränas på *k-1* folds och utvärderas på den återstående folden. Denna process upprepas *k* gånger, så att varje fold används som valideringsdata en gång. Detta minskar risken för bias från en enda tränings- eller valideringsuppdelning och säkerställer att modellen utnyttjar datan på ett effektivt sätt. (Wikipedia (a), 2025)
- **Confusion matrix:** Är ett verktyg som används för att utvärdera prestandan hos en klassificeringsmodell. Matrisen är en tabell som visar hur väl modellen förutspår de faktiska klasserna i datan. (Wikipedia (b), 2025)
- **Accuracy:** Accuracy score är ett mått på hur många förutsägelser som är korrekta i förhållande till det totala antalet förutsägelser. (Wikipedia (c), 2025)

2.2 Maskinlärningsmodeller

- **Random Forest:** Ett beslutsträd där varje träd tränas på ett slumpmässigt urval av data och features. Slutligen ges prediktioner baserat på majoritetsröstning av de olika träden. (Wikipedia (d), 2025)
- **Extra Trees:** Liknar Random Forest men med mer slumpmässig noddelning. (Wikipedia (d), 2025)
- **Voting Classifier:** Kombinerar flera modellers prediktioner genom röstning. (Wikipedia(e), 2025)

2.3 MNIST-dataset

Ett dataset som innehåller 70 000 handskrivna normaliserade svart-vita 28x28 pixel bilder på siffror. Datasetet används används ofta för att träna bildigenkänningsmodeller. Datan är tagen från amerikanska high school studenter. (Wikipedia (f), 2025)

3 Metod

Studien genomfördes enligt följande:

1. Datainläsning och förbehandling av MNIST-datasetet (70 000 bilder)
2. Uppdelning av data:
    a. 50 000 bilder för träning
    b. 10 000 bilder för validering
    c. 10 000 bilder för slutgiltig testning
3. Träning av tre modeller med standardparametrar
4. Utvärdering:
    a. Accuracy-mätning
    b. Confusion matrix
    c. Visuell inspektion av felprediktioner

Besvara nedanstående teoretiska frågor koncist.

1. Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?

Tränings setet är till för att träna modellen (modell.fit), validering är till för att välja bäst presterande modell samt för att sedan träna om den bästa modellen på träning + validering. Test setet är till för att testa den slutgiltiga modellen.

2. Julia delar upp sin data i träning och test. På träningsdatan så tränar hon tre modeller; ”Linjär Regression”, ”Lasso regression” och en ”Random Forest modell”. Hur skall hon välja vilken av de tre modellerna hon skall fortsätta använda när hon inte skapat ett explicit ”validerings-dataset”?

Den modell som presterar bäst på test setet. Detta funkar fortfarande men ger ofta mindre korrekt modell än med ett explicit validerings-dataset.

3. Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?

När man vill att modellen ska gissa ett numerisk värde, kontra t.ex en etikett / string. Logarithmic regression > gissa värde på ett hus (linjärt samband), random forest > bild-igenkänning (icke linkärt samband)

4. Hur kan du tolka RMSE och vad används det till:

$$RMSE = \sqrt{\frac{1}{n}\sum(yi - \hat{y}i)^2}$$

Root mean squared error är lite som det låter, roten ur medel-felet i kvadrat från en modell. Används för att mäta prestationen av en regressionsmodell. Går inte att använda på kategorisk data utan att på något vis klassa hur stort fel det är mellan olika etiketter.

5. Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?

Motsatsen till regressionsproblem, modellen ska istället gissa på en kategori / string. Logarithmic regression > gissa sjukdom (linjärt samband), random forest > spam filter (spam kan vara mycket olika typer av mail, därav passar randomforest med olika modeller som tillsammans röstar)

En confusion matrix visar hur modellen gissade på datan den predikade där x-raden är vad den gissade och y-raden vad det korrekta svaret var. Med andra ord kan man se hur ofta modellen gissade fel och i så fall vad den “förvirrade” sig att svaret var. T.ex.

Bok	0.5	0.25	0.25
Träd	0.33	0.62	0.01
Sko	0.25	0.20	0.55
	Bok	Träd	Sko

6. Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.

K-means använder klustring vilket är ett sätt att gruppera liknande data tillsammans. Ett exempel är att en affär vill kategorisera sina kunder efter kundens motivation till att köpa produkter, t.ex små köp hela tiden, nästan enbart rabatterade produkter, eller större köp då och då.

7. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding. Se mappen ”l8” på GitHub om du behöver repetition.

Används för att omvandla kategorisk data, t.ex bostad som “nära till havet” eller “nära till förskola”, till numeriska värden som modellen kan använda. Ordinal encoding ger 1,2,3 i rangordning, one-hot encoding ger binär kod t.ex färg RGB och dummy variable ger likt one-hot encoding men bara 2 kolumner.

8. Göran påstår att datan antingen är ”ordinal” eller ”nominal”. Julia säger att detta måste tolkas. Hon ger ett exempel med att färger såsom {röd, grön, blå} generellt sett inte har någon inbördes ordning (nominal) men om du har en röd skjorta så är du vackrast på festen (ordinal) – vem har rätt?

Färgen saknar ordning och är därmed nomial. Vacker är ordinal eftersom man kan klassa mest vackrast till minst vackrast. Båda har rätt var för sig.

9. Kolla följande video om Streamlit: https://www.youtube.com/watch?v=ggDa-RzPP7A&list=PLgzaMbMPEHEx9Als3F3sKKXexWnyEKH45&index=12

Och besvara följande fråga:

- Vad är Streamlit för något och vad kan det användas till?

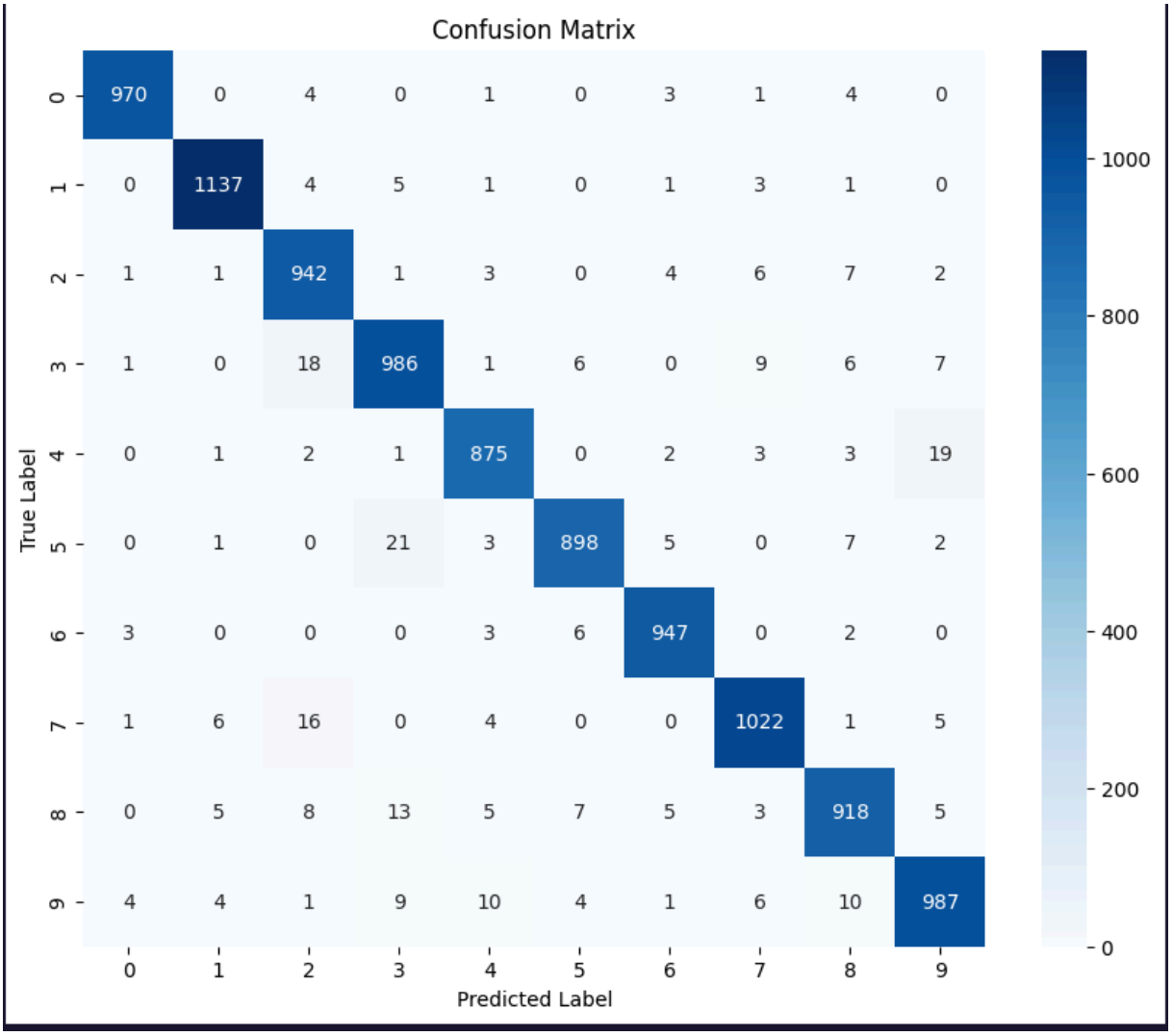
Streamlit är ett python baserat bibliotek för att deploya web apps specifikt för machine learning och data science. Helt enkelt kan du köra dina python script och web app i samma fil.

4 Resultat och Diskussion

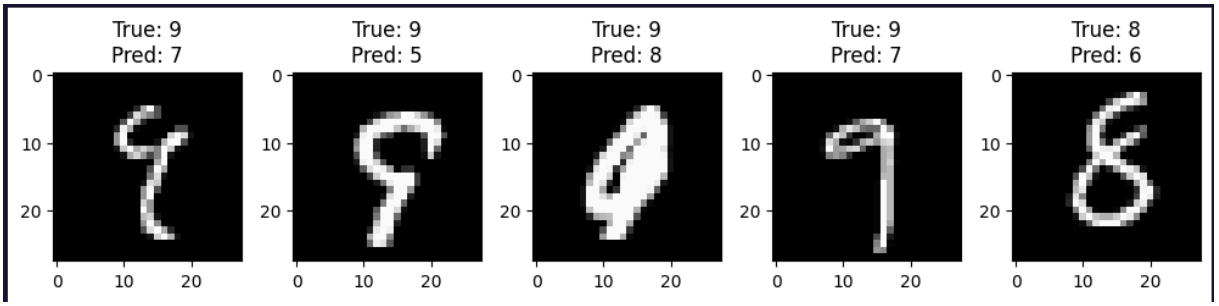
Accuracy för olika modeller på validerings set	
Random Forest	96,92%
Extra Trees	97,15%
Voting	97,13%

Tabell 1: Accuracy score för de valda modellerna

Resultaten visar att alla tre modellerna är mycket effektiva för att prediktera MNIST-data, vilket tyder på att maskinlärningsmodellernas är kraftfulla verktyg för detta typ av klassificeringsproblem. Extra Trees visade sig prestera bäst. Även om Voting Classifier teoretiskt sett borde dra nytta av att kombinera styrkorna hos flera modeller, presterade den marginellt sämre än Extra Trees. Detta kan indikera att de använda grundmodellerna redan har mycket hög individuell prestanda, vilket begränsar de ytterligare vinsterna från röstningen.



Tabell 2: Confusion metrix för Extra Trees på test set



Tabell 3: Exempel på 5 felprediktioner med dess prediktions- och truevärde

Analys av felprediktioner visar att:

- Vanligaste förväxlingarna sker mellan visuellt liknande siffror (t.ex. 5/3, 4/9 och 3/2)
- Vissa felprediktioner är förståeliga även ur ett mänskligt perspektiv

5 Slutsatser

1. Undersöka om kombinerade modeller (Voting Classifier) kan förbättra resultaten:

   Ökad komplexitet genom voting gav ingen förbättring

2. Utvärdera och jämföra olika maskininlärningsmodeller effektivitet för sifferigenkänning:

   Både Random forest och Extra Trees visar sig vara effektiva på sifferigenkänning med hög accuracy.

6 Teoretiska frågor

- Varför förbättrade inte Voting Classifier resultatet?

  Modellerna kan vara för korrelerade eller så fångar Extra Trees redan upp de viktigaste mönstren.

- Hur skulle resultatet kunna förbättras?
  a. Hyperparameter optimering, större n_estimators.
  b. Fler modeller för voting classifiern.
  c. K-fold cross-validation hade kunnat minska modellernas overfitting och därmed öka modellernas accuracy.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   Rapport skrivningen är nog det svåraste för mig, jag känner att jag gjorde för lite för att verkligen framställa en komplett rapport. Jag är helt ovan vid APA systemet vilket tog mycket tid för mig.

2. Vilket betyg du anser att du skall ha och varför.
   G, jag gjorde absolut minimum för själva uppgiften då jag la mycket tid på att experimentera och lära mig i mindre experiment.

3. Något du vill lyfta fram till Antonio?

   Väldigt kul med en mer praktisk kurs på en mer utmanande nivå :)

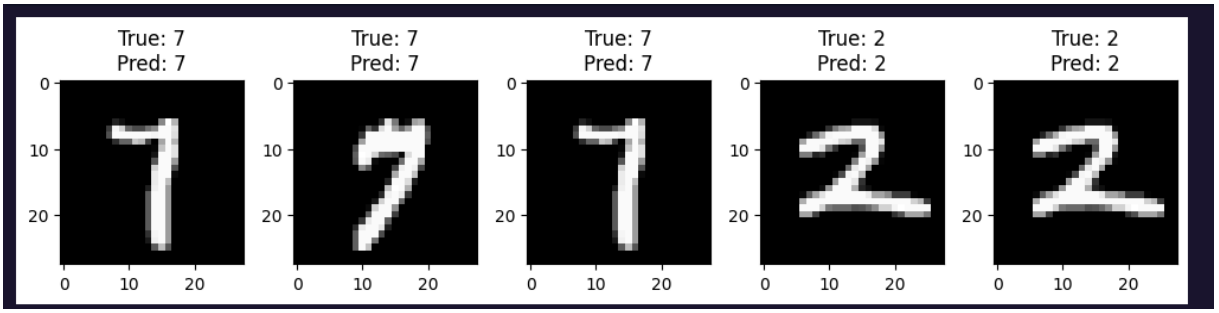
Appendix A

Modelling

```
random_forest_clf = RandomForestClassifier(n_estimators=100, random_state=42)
extra_trees_clf = ExtraTreesClassifier(n_estimators=100, random_state=42)
```

[6] ✓ 0.0s

Bilaga 1: Hyperparametrar för modeller



Tabell 4: Exempel på korrekt predikterade siffror

[Github]

Källförteckning

1. Wikipedia contributors (a) (2025). *Machine learning*. Hämtad mars 14, 2025, från Wikipedias sida https://en.wikipedia.org/wiki/Machine_learning.
2. Wikipedia contributors (b) (2025). *Confusion matrix*. Hämtad mars 18, 2025, från Wikipedias sida https://en.wikipedia.org/wiki/Confusion_matrix.
3. Wikipedia contributors (c) (2025). *Accuracy and precision*. Hämtad mars 18, 2025, från Wikipedias sida https://en.wikipedia.org/wiki/Accuracy_and_precision.
4. Wikipedia contributors (d) (2025). *Random forest*. Hämtad mars 15, 2025, från Wikipedias sida https://en.wikipedia.org/wiki/Random_forest.
5. Wikipedia contributors (e) (2025). *Ensemble learning*. Hämtad mars 15, 2025, från Wikipedias sida https://en.wikipedia.org/wiki/Ensemble_learning.
6. Wikipedia contributors (f) (2025). *MNIST database*. Hämtad mars 17, 2025, från Wikipedias sida https://en.wikipedia.org/wiki/MNIST_database.