# Rapportera på Maskin Inlärning Modeller för MNIST Dataset

# EC Utbildning logo

Lesley Atali Etashu

EG utbildning

Maskininlärningsprojekt .

2025/03

**Tabell av Innehåll**

[Rapportera på Maskin Inlärning Modeller för MNIST Datauppsättning](#_bookmark0)  [1](#_bookmark0)

[Abstrakt](#_bookmark1)  [3](#_bookmark1)

1. [Inledning 3](#_bookmark2) [​](#_bookmark2)
2. [Teori](#_bookmark3)  [4](#_bookmark3)

2.1 [Logistisk regression](#_bookmark4)  [4](#_bookmark4)

2.2 Linjär [SVM](#_bookmark5)  [4](#_bookmark5)

2.3 [Random Forest](#_bookmark6)  [4](#_bookmark6)

2.4 [Ensemblemodell](#_bookmark7)  [5](#_bookmark7)

1. Metod 6
2. [Modell Träning](#_bookmark9)  6
3. [Resultat och diskussion](#_bookmark10)  [6](#_bookmark10)
4. [Jämförande Analys](#_bookmark15)  [6](#_bookmark15)
5. [Slutsats](#_bookmark16)  [7](#_bookmark16)

[Teoretisk Frågor](#_bookmark17)  [8](#_bookmark17)

[Självutvärdering 13​](#_bookmark27)

[Referenser](#_bookmark28)  [14](#_bookmark28)

## Abstrakt

Manuskriptet granskar utvecklingsprocessen för flera maskininlärningsmodeller som tillämpas på MNIST dataset tillsammans med deras prestanda utvärdering. A behandla av dataset förbearbetning leder till utbildning av olika klassificerare följt av en noggrannhets- och klassificeringsrapportbaserad prestationsjämförelse. Fyra metoder för maskininlärning användes i experiment, inklusive Logistic Regression tillsammans med Linear Support Vector Machine (SVM) medan Random Forest samt en Ensemble-klassificerare som består av dessa tre metoder. RandomForest -modellen visar mest exakta prestanda när den tillämpas på denna uppsättning modeller.

## Introduktion

Maskininlärningsexperter överväger MNIST dataset som ett standardriktmärke för bildklassificeringsforskning. Denna datauppsättning fungerar som ryggraden för att testa många maskininlärningsalgoritmer på grund av dess enkla struktur och medelstora tillsammans med lättbegripliga format (Goodfellow et al., 2016).

Huvudmålet med detta projekt fokuserar på att tillämpa flera maskininlärningsmodeller för att känna igen siffror i MNIST-datauppsättningen medan de utför deras jämförande analys. Denna studie presenterar fyra modeller för övervägande som inkluderar Logistic Regression kombinerat med Linear Support Vector Machine (SVM) tillämpad genom Stokastisk Gradient Descent (SGD) och Random Forest och en ensemblemodell som aggregerar förutsägelser med hjälp av tekniker för majoritetsröstning. Enligt Hastie et al. (2009) både Logistic Regression och SVM utmärker sig på att arbeta i högdimensionella utrymmen eftersom de är etablerade linjära modeller. Samtidigt använder Random Forest ( Breiman , 2001) beslutsträd i ensemblestrukturer för att upptäcka icke-linjära mönster. klassificerares prediktiva kraft förbättras när ensemblemetoder inklusive majoritetsomröstning förenar olika klassificerare enligt Dietterich (2000). utvärdering fokuserar på modell noggrannhet tillsammans med beräkningsmässigt effektivitet och allmän effektivitet för att känna igen handskrivna siffror.

Detta dokument genomför en omfattande studie av MNIST dataset tillsammans med dess förbearbetningssteg och modellutbildningstekniker och utvärderingsåtgärder samt presentera resultat genom jämförande analyser. Detta projekt undersöker hur klassificeringsmetoder utförs i sifferdetekteringsoperationer samtidigt som man undersöker hur samarbete med flera modeller kan uppnå bättre precision när de isoleras från varandra. resultaten från denna forskning blir viktiga för att förklara olika metoder som används för att välja modeller tillsammans med optimering av bildklassificeringsprocedurer .

## Data Förvärv och Förbearbetning

### Logistisk regression:

### Detta är Machine Learning-modellen som har använts för att lösa klassificeringsproblem och detta görs genom att förutsäga kategoriska utfall. Det har oftast två utfall som kallas binomial.

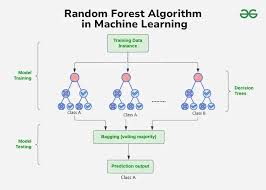
### C:\Users\lesle\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E7CE68A8.tmp

### Linjär SVM:

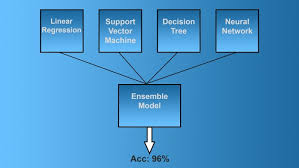
### En övervakad maskininlärningsmodell som används för att lösa klassificerings- och regressionsproblem. Den är bra på att lösa binära klassificeringsproblem som kräver att data klassificeras i två grupper. Huvudsyftet är att ta reda på den bästa möjliga linjen som separerar datapunkten för olika dataklasser.

### C:\Users\lesle\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\31B8DCF6.tmp

**2.3 Random Forest:** En maskininlärningsalgoritm som används för att lösa både klassificerings- och regressionsproblem. Jag bygger en samling beslutsträd under träning och kombinerar senare deras resultat för att göra en mer exakt förutsägelse.

****

**2.4 Ensemblemodeller:** Kombinerar olika modeller för att förbättra den allmänna förutsägelsen. De huvudsakliga teknikerna i ensemblemodellen är packning, boosting och stapling.

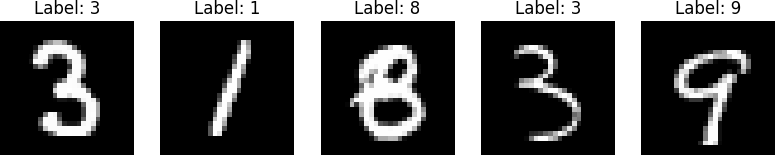
****

## Metod

Data som användes i detta projekt var MNIST datauppsättning som erhölls var erhållits i CSV formatera från [pjreddie.com](https://pjreddie.com/media/files/) . Det består av:

* + - mnist\_train.csv: 60 000 prover för utbildning.
    - mnist\_test.csv: 10 000 prover för testning.

Figur nedan visa dataset 5 etiketter.



## Modell Utbildning

Varje modell var utbildad använder förbearbetade utbildning uppsättning ( X\_train\_scaled , y\_train ). utbildning involverad process :

* Logistisk Regression och SVM använder gradientbaserad optimering.
* Slumpmässig Skog använder multipel beslut träd för majoritet röstning.
* Ensemble Röstning kombinerande förutsägelser av tre klassificerare.

## Resultat och diskussion

Utvärdering var genomfördes på testa uppsättning ( X\_test\_scaled , y\_test ) använder noggrannhet och klassificeringsrapporter .

### Logistisk Regression

* + - **Noggrannhet:** 89,27 %
    - **Observationer:** Uppträder väl men kämpar med överlappande siffror.

### Linjär SVM

* + - **Noggrannhet:** 90,22 %
    - **Observationer:** Ger bättre marginal separation än Logistisk Regression.

### Slumpmässig Skog

* + - **Noggrannhet:** 95,45 %
    - **Observationer:** Hög noggrannhet men beräkningsmässigt dyr.

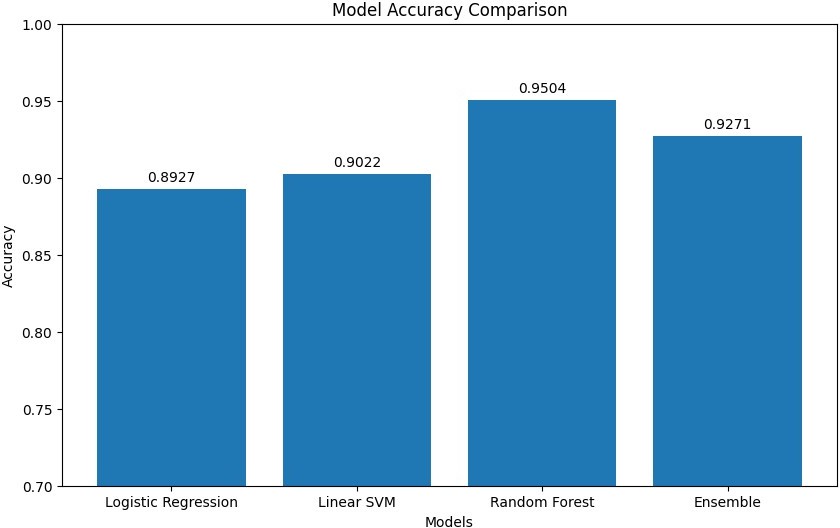
### Ensemble Modell

* + - **Noggrannhet:** 92,71 %
    - **Observationer:** Överträffar enskild modeller, utnyttjande deras styrkor.

## Jämförande Analys

|  |  |
| --- | --- |
| **Modell** | **Noggrannhet** |
| Logistisk Regression | 89,27 % |
| Linjär SVM | 90,22 % |
| Slumpmässig Skog | 95,45 % |
| Ensemble Modell | 92,71 % |

I figur nedan är visuell representation av modeller jämförelse.



De Slumpmässig Skog försedd högsta noggrannhet av kombinerande förutsägelser från olik modeller.

## Slutsats

De forskning experimentera bevisar Slumpmässig Skog inlärning tjänar effektivt för siffra identifieringsändamål. bästa prestanda kom från Ensemble-metoder men deras beräkningsbelastning är fortfarande hög. Framtida förbättringar inkluderar:

* De modellering behandla skall omfatta djup inlärning närmar sig genom Convolutional Neural Networks (CNN).
* Hyperparameter justering möjliggör överlägsen prestanda nivåer i a system.
* Ytterligare data generation tekniker hjälp slumpmässig skog modeller få mer universella förmågor.

De MNIST klassificering ML rörledning tar emot grundlig dokumentation angående dess noggrannhet och skalbarhetspotential samtidigt som effektivitetsparametrar demonstreras.

# Teoretisk Frågor

### Karl delade dessa data eftersom när man tränar en maskininlärningsmodell delas data in i tre huvudundergrupper för att säkerställa korrekt modellutvärdering och förhindra överanpassning:

* + **Utbildningsuppsättning** : Detta är den största delen av datasetet, som används för att träna maskininlärningsmodellen. Modellen lär sig mönster och samband från dessa data.
  + **Valideringsuppsättning** : Detta används för att stämma hyperparametrar och jämför olik modeller. Det hjälper till att förhindra överanpassning, eftersom modellens prestanda utvärderas på osynliga data före slutlig testning.
  + **Testuppsättning** : Denna uppsättning används endast efter slutgiltigt modellval för att ge en opartisk utvärdering av modellens prestanda på helt nya data.

### Hur burk Julia välja bäst modell om hon gjorde inte skapa a godkännande dataset?

Julia utbildad tre olik modeller: **Linjär Regression, Lasso Regression och Slumpmässig Forest** , men hon delade bara upp sina data i tränings- och testset. Utan en valideringsuppsättning kan hon:

* + Använd **korsvalidering** : Detta innebär att dela upp träningsdata i flera veck, träning på några veck och validerar på andra i a roterande sätt. K-veck korsvalidering (till exempel, k=5 eller 10) är ett vanligt tillvägagångssätt.
  + Jämföra prestanda använder **metrik** sådan som RMSE (för regression), RZ göra, eller MSE.
  + Använda **Testa data för slutlig utvärdering** , men vara försiktig av övermontering. Om hon låtar hennes modeller som använder testdata, kanske det inte längre fungerar som en opartisk utvärdering.

### Vad är a regression problem?

A **regression problem** är a typ av övervakas inlärning uppgift där mål är till förutspå kontinuerliga numeriska värden. Exempel inkluderar:

#### Modeller använd :

* 1. Linjär Regression
  2. Lasso regression
  3. Ås Regression
  4. Beslut Träd (för regression)
  5. Slumpmässig Skog Regressor
  6. Neural Nätverk

#### Applikationer :

* 1. Förutsäga hus priser
  2. Uppskattning stock marknadsföra trender
  3. Prognoser försäljning inkomst
  4. Förutsäga väder temperaturer

### Tolkning och använda av RMSE

RMSE (Rot Betyda Kvadrat Fel) åtgärder hur väl a regression modellens förutsägelser matcha faktiska värden. Det beräknas som:

RMSE = square root of ( (1/n) \* sum from i = 1 to n of (yi - ŷi)Z )

#### Tolkning :

* + - A lägre RMSE indikerar bättre modell noggrannhet.
    - RMSE har samma enheter som beroende variabel, tillverkning det tolkbar.

#### Användning :

* + - Till jämföra olik regression modeller.
    - Till mäta hur långt förutsägelser avvika från faktisk värden.
    - Till ställa in hyperparametrar och förbättra modell noggrannhet.

### Vad är a klassificeringsproblem ?

A **klassificering problem** innebär förutsäga kategorisk etiketter snarare än kontinuerlig värden.

#### Exempel av modeller :

* + - Logistisk Regression
    - Beslut Träd
    - Slumpmässig Skog
    - Stöd Vektor Maskiner (SVM)
    - Neural Nätverk

#### Applikationer :

* + - Spam upptäckt (Spam/Inte skräppost)
    - Sjukdom diagnos (COVID 19 positiv/negativ)
    - Känsla analys (positiv/negativ)

**Förvirring Matris** : A förvirring matris är a tabell begagnad till utvärdera prestanda av a klassificeringsmodell. Den består av:

* + **Sann Positivt (TP)** : Rätta positiv förutsägelser.
  + **Falsk Positivt (FP)** : Felaktigt förutspått som positiv.
  + **Sann Negativt (TN)** : Rätta negativ förutsägelser.
  + **Falsk Negativt (FN)** : Felaktigt förutspått som negativ.

Det hjälper i datoranvändning noggrannhet, precision, återkallande och F1- poäng.

### Vad är K-medel klustring och där är gällde det ?

K-medel är en **utan tillsyn inlärning algoritm** begagnad för klustring data till K grupper.

#### Arbetssätt Princip :

* + - Välja **K** -kluster.
    - Tilldela data poäng slumpvis till kluster.
    - Beräkna klunga **tyngdpunkt** .
    - Tilldela om poäng baserad på närmast tyngdpunkter.
    - Upprepa tills konvergens.

#### Exempel Applikationer :

* + - Kund segmentering i marknadsföring
    - Bild kompression
    - Anomali upptäckt i cybersäkerhet

K-medel är allmänt begagnad förfallen till dess enkelhet, effektivitet och effektivitet i hantering stor datauppsättningar. Det kräver dock att användaren specificerar K i förväg och resultaten kan variera beroende på det initiala valet av tyngdpunkter.

### Kodning tekniker (med exempel)

Kodning konverterar kategorisk data till numerisk form för maskin inlärning modeller.

#### Ordinal Kodning :

* + - Tilldelar **rankad** värden till kategorier.
    - Exempel: Skjorta storlekar → {Små: 1, Medium: 2, Stor: 3}.

#### One-Hot Kodning :

* + - Konverterar kategorier till separat binär kolumner.
    - Exempel: Färger → {Röd: [1, 0, 0], Grön: [0, 1, 0], Blå: [0, 0, 1]}.

#### Dummy Variabel Kodning :

* + - Som One-Hot Kodning men tar bort en kolumn till undvika multikollinearitet.
    - Exempel: Färger (släpper "Röd") → {Grön: [1, 0], Blå: [0, 1]}.

### Är data antingen ordinarie eller nominell?

Göran påståenden att data är antingen **ordinarie eller nominell** , medan Julia argumenterar att tolkning frågor.

* + **Nominell Data** : Inga inneboende beställa (till exempel, {röd, grön, blå}).
  + **Ordinal Data** : Har a rankad beställa (till exempel, T-shirt storlekar: S < M < L).

Julias exempel av "a röd skjorta varelse mest vacker på a part" är **subjektivt tolkning** , vilket innebär att vissa data kan vara kontextberoende. Så **Julia har rätt** .

### Vad är Strömbelyst och vad burk det vara begagnad för?

#### Vad är Strömbelyst ?

Streamlit är ett Python-ramverk med öppen källkod som förenklar processen att bygga och distribuera interaktiva webbapplikationer, särskilt för maskininlärning (ML) och datavetenskapsprojekt. Det möjliggör utvecklare och data forskare till skapa kraftfull och användarvänlig web applikationer använder bara Python-skript, utan att kräva avancerad kunskap om frontend-utveckling, som HTML, CSS eller JavaScript.

#### Nyckel Drag av Strömbelyst :

* + **Enkel och Pythonic:** Utvecklare burk skapa interaktiv applikationer använder ren Python utan att skriva separat HTML- eller JavaScript-kod.
  + **Automatisk uppdatering:** Ansökan uppdateringar automatiskt när manus är ändrad, eliminerar behovet av manuell uppdatering.
  + **Inbyggt widgets:** Strömbelyst tillhandahåller olika interaktiv element som skjutreglage, knappar, textinmatningar och diagram, vilket gör det enkelt att bygga instrumentpaneler och visualiseringar.
  + **Sömlös integration med ML bibliotek:** Det fabrik väl med populär Pytonorm bibliotek som TensorFlow, PyTorch , scikit -learn, Matplotlib och Pandas.
  + **Lättvikt spridning:** Ansökningar burk vara utplacerade lätt använder tjänster som Streamlit Cloud, Heroku eller AWS.

#### Vad burk Strömbelyst vara begagnad för?

Strömbelyst är allmänt begagnad i data vetenskap och maskin inlärning applikationer för följande syften:

#### Byggnad Interaktiv ML Instrumentpaneler:

* + Visualisera och interagerar med maskin inlärning modeller i realtid .
  + Tillåter användare till justera parametrar och se hur modell förutsägelser förändras dynamiskt.
  + Jämförande olik ML modeller använder interaktiv widgets.

#### Utplacering Data Visualisering Verktyg:

* + Visar diagram, grafer och data tabeller använder bibliotek som Matplotlib, Seaborn och Plotly .
  + Skapande interaktiv rapporterar med filter och dynamisk uppdateringar.
  + Utforska stor datauppsättningar genom interaktiv användare gränssnitt.

#### Visa upp Maskin Inlärning Modeller med Minimal Ansträngning:

* + Demonstrerar modell förutsägelser och resultat utan kräver a komplex backend.
  + Tillåter användare till ladda upp filer och göra förutsägelser använder utbildade modeller.
  + Tillhandahåller en lätt att använda gränssnitt för icke-tekniska användare till påverka varandra med ML- modeller.

#### Prototypframställning och Experimenterande:

* + Snabbt testning idéer och algoritmer före fullskalig spridning.
  + Skapande proof-of-concept applikationer för ML och data analys projekt.
  + Samarbetar med lag av delning funktionell prototyper lätt.

#### Exempel Använda Fall:

A dataforskare bygger en **kontrollpanel för att upptäcka bedrägerier** med Streamlit . Dashboard låter användare ladda upp transaktion data, justera bedrägeri upptäckt modell parametrar och se förutsägelser i i realtid. Interaktiva visualiseringar hjälper till att analysera bedrägliga transaktioner och instrumentpanelen kan delas med intressenter för insikter.

Sammanfattningsvis Streamlit är ett utmärkt verktyg för att utveckla interaktiv och engagerande ML och dataapplikationer med minimal ansträngning, tillverkning det starkt användbar för både nybörjare och experter i fält.

# Själv Utvärdering

Machine Learning har varit ganska svårt med så många teorier att förstå och vissa modeller är svåra att träna men det är också intressant att se resultatet av var och en. Den här perioden har inte varit den bästa men glad att se att jag kunde genomföra detta. Allt jag kan säga är att varje dag vi spenderar i klassen kan jag se tillväxt och förändringar i mina studier.

# Referenser

* Breiman , L. (2001). Slumpmässiga skogar. *Machine Learning* , 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
* Dietterich , TG (2000). Ensemblemetoder i maskininlärning. I *Multiple Classifier Systems* (s. 1–15). Springer. <https://doi.org/10.1007/3-540-45014-9_1>
* <https://web.stanford.edu/~hastie/ElemStatLearn/>
* LeCun , Y., Cortes, C., & Burges, CJC (2010). MNIST Handwritten Digit Database. Hämtad från <http://yann.lecun.com/exdb/mnist/>
* <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>
* <https://scikit-learn.org/>
* Google bilder.