Sifferigenkänning



Jacob Andersson

EC Utbildning

Machine Learning - Kunskapskontroll

2024/03

# 

# Abstract

This study presents a machine learning-based method for handwritten digit recognition, focusing on the MNIST dataset. Several classification algorithms, including Support Vector Machines, K-Nearest Neighbor, and Decision Trees, are evaluated and compared for their ability to correctly identify handwritten digits. The models are trained and evaluated on a validation set to identify the most effective one. The goal is to create a model that achieves an 85% accuracy rate. Additionally, a web application is developed using Streamlit to allow users to upload handwritten images and receive predictions based on the best model.

Innehållsförteckning

[Abstract 2](#_Toc162021965)

[1 Inledning 1](#_Toc162021966)

[1.1 Frågeställningar 1](#_Toc162021967)

[2 Teori 2](#_Toc162021968)

[2.1 Maskininlärningsmodeller 2](#_Toc162021969)

[2.2 SVC 2](#_Toc162021970)

[2.3 K-Nearest Neighbor 2](#_Toc162021971)

[2.4 Decision Trees 3](#_Toc162021972)

[2.5 Välja hyperparametrar 3](#_Toc162021973)

[3 Metod 4](#_Toc162021974)

[3.1 Datainsamling MNIST-Dataset 4](#_Toc162021975)

[3.2 Modellutveckling: 4](#_Toc162021976)

[3.3 Test och utvärdering av den bästa modellen: 5](#_Toc162021977)

[3.4 förbearbetning och prediktion av nya bilder: 5](#_Toc162021978)

[3.5 Streamlit applikation 5](#_Toc162021979)

[4 Resultat och Diskussion 7](#_Toc162021980)

[4.1 Prediktion av handskrivna siffror 7](#_Toc162021981)

[4.2 Begränsningar och framtida förbättringar 7](#_Toc162021982)

[5 Slutsatser 8](#_Toc162021983)

[6 Teoretiska frågor 9](#_Toc162021984)

[7 Självutvärdering 11](#_Toc162021985)

[Källförteckning 12](#_Toc162021986)

# Inledning

Maskininlärningsmodeller är otroligt viktiga inom en rad olika områden, inklusive datavetenskap, marknadsföring, finans och detaljhandel, bland andra. Idag är det knappast någon bransch som inte påverkas av den revolution som maskininlärning har medfört. (2)

Syftet med denna rapport är att prediktera handskrivna siffror med hjälp av maskininlärningsmodeller. Att känna igen handskrivna siffror är en grundläggande uppgift inom mönsterigenkänning och maskininlärning. I arbetet används MNIST-datasetet som består av 28x28 pixlar gråskalebilder av handskrivna siffror. MNIST-datasetet har tjänat som ett benchmark för att utvärdera olika maskininlärningsalgoritmers prestanda inom detta område.

I denna studie utforskas effektiviteten hos olika klassificeringsmodeller för att känna igen handskrivna siffror. Specifikt undersöks prestandan hos Support Vector Classifier (SVC), K-Nearest Neighbor (KNN) och Decision Trees.

Datasetet delas upp i tränings-, validerings- och testdata, med lämpliga förbehandlingssteg såsom standardisering. Grid Search cross-validation används för att finjustera hyperparametrarna hos modellerna vilket optimerar deras prestanda.

Vidare Skapas en webbapplikation med Streamlit, som gör det möjligt för användare att ladda upp bilder av handskrivna siffror för förutsägelse med den bäst presterande modellen som identifierats under vårt experiment.

## Frågeställningar

Frågor som kommer besvaras i denna rapport är:

1. Kan den bästa modellen prediktera med en pricksäkerhet på 85 procent?
2. Vilken maskininlärningsmodell presterar bäst för handskriven sifferigenkänning på MNIST-datasetet?
3. Hur kan modellprestandan förbättras genom optimering av hyperparametrar och förbehandlingstekniker?

# Teori

I det här kapitlet beskrives teorin, som behövs för att förstå de olika maskininlärningsmodellerna och hur dom tränas.

## Maskininlärningsmodeller

Det finns en mängd olika modeller att välja mellan när det kommer till maskininlärning. I denna uppgift används tre olika modeller som tränas och utvärderas för att hitta den bäst presterande.

## SVC

SVC är en Support Vector Machine modell, som kan användas för både klassificeringsproblem och regressionsproblem.

En bild som visar text, skärmbild

Automatiskt genererad beskrivning  
*Fig. 1 Klassificerar datapunkter med SV*

## K-Nearest Neighbor

KNN är en metod för dataklassificering som försöker uppskatta sannolikheten för att en given datapunkt tillhör en viss klass baserat på vilken klass de närliggande datapunkterna tillhör.

En bild som visar text, skärmbild, diagram, Teckensnitt

Automatiskt genererad beskrivning  
*Fig. 2 Visar hur datapunkter placeras i klasser*

## Decision Trees

Decision Trees (beslutsträd) är användbar modell som också kan hantera klassificerings- och regressionsprobelm. Beslutsträd kan hantera komplexa dataset och fungerar genom att sortera eller göra förutsägelser baserat på hur tidigare frågor har besvarats.

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning  
*Fig. 3 Visar hur man men hjälp av beslutsträd delar in data i olika klasser*

## Välja hyperparametrar

För att skapa en bra modell används hyperparametrar som styr inlärningen hos modellen.  
För att hitta dom som fungerar bäst andvänds Grid Search funktionen. Grid Search testar alla olika Hyperparametrar för varje modell och tar fram dom som har presterat bäst.

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning  
*Fig. 4 Visar användning av Grid Search som väljer bästa värdena för varje hyperparameter*

En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning

  
*Fig. 5 Printar ut dom bästa hyperparametrarna för SVC modellen*

# Metod

## Datainsamling MNIST-Dataset

Datainsamling och förberedelse:

MNIST-datasetet är ett välkänt dataset som innehåller 28x28 pixlar stora bilder på handskrivna siffror från noll till nio och innehåller totalt 70 000 bilder.

För att möjliggöra modellträning och utvärdering delas datasetet upp i tre delar, träningsuppsättning (för att träna modellen), valideringsuppsättning (för att optimera hyperparametrar och undvika överanpassning) och testuppsättning (för att utvärdera den slutliga modellens prestanda).

## Modellutveckling:

En enkel utforskning av datan görs för att visa några exempelbilder på dom handskrivna siffrorna. Sedan skapas och tränas tre olika klassificeringsmodeller: Support Vector Classification modell (SVC), K-Nearest Neighbors modell (KNN) och en Decision Trees Modell.

För varje modell utförs hyperparameter-optimering med hjälp av grid search för att hitta de bästa parametrarna. Modellerna utvärderas sedan med valideringsdata och confusion matrices genereras för att visualisera deras prestanda.

Efter utvärderingen testas den bästa modellen med testdata och dess prestanda utvärderas. Slutligen sparas den bästa modellen för framtida användning.

En funktion definieras också för att förbearbeta en ny bild innan den används för att göra förutsägelser med den bästa modellen. En huvudfunktion används för att applicera denna process på en given bild och gör en förutsägelse.

Den modell som visade högst accuracy\_score på valideringsuppsättningen valdes Support Vector Classification modellen (SVC) som fick en accuracy score på 94 procent

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning  
*Fig. 4 Visar användning av Grid Search som väljer bästa värdena för varje hyperparameter*

En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning *Fig. 5 Printar ut som bästa hyperparametrarna*

## Test och utvärdering av den bästa modellen:

Slutligen beräknas och rapporteras noggrannheten för den bästa modellen på testuppsättningen. Denna noggrannhet representerade andelen korrekta förutsägelser jämfört med det totala antalet exempel i testuppsättningen och ger en klar bedömning av modellens övergripande prestanda på nya och oberoende data.

Resultaten av test- och utvärderingsprocessen bekräftar den tidigare observationen av den valda modellens höga prestanda och bekräftar dess användbarhet för korrekt klassificering av handskrivna siffror från MNIST-datasetet.

Bästa modellen sparas för framtida användning, vilket möjliggör att den kan användas för att göra förutsägelser på nya, osedda data.

## förbearbetning och prediktion av nya bilder:

För att förutsäga siffror från nya bilder, börjar processen med att läsa in och konvertera bilderna till gråskala för enklare bearbetning. Därefter anpassas bildstorleken till 28x28 pixlar och eventuell bildförbättring kan appliceras för att förbättra kvaliteten. Genom thresholding och binarisering separeras siffrorna tydligt från bakgrunden. Sedan omvandlas bilden till en vektor och pixelvärden normaliseras för att matcha träningsdata. Den bästa modellen används sedan för att göra en förutsägelse baserat på den förbearbetning bilden. Resultatet tolkas slutligen för att identifiera den förutsagda siffran. Denna process möjliggör automatiserad sifferigenkänning för en mängd olika tillämpningar.

## Streamlit applikation

För att kunna ladda upp egna handskrivna bilder och låta modellen prediktera dom skapas en Streamlit applikation. Först laddas den bästa modellen in, sedan gjordes samma förbearbetning som tidigare.

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning  
*Fig. 6 har kan du ladda upp egna bilder på Streamlit applikationen*

En bild som visar skärmbild, svart och vit

Automatiskt genererad beskrivning En bild som visar skärmbild, svart, text, design

Automatiskt genererad beskrivning  
*Fig. 7 uppladdad Handskriven siffra Fig. 8 Bilden transformeras so modellen kan predikta*

# Resultat och Diskussion

Efter att ha genomfört träningen och utvärderingen av flera maskininlärningsmodeller för handskriven sifferigenkänning på MNIST-datasetet, kunde följande resultat och observationer dras

|  |  |
| --- | --- |
| **Accuracy score** | |
| Support Vector Classifier | 0.94 |
| K-Nearest Neighbor | 0.897 |
| Decision Trees | 0.779 |

Support Vector Classification modellen presterade bäst med en accuracy score på 94 procent.

## Prediktion av handskrivna siffror

En metod för att prediktera handskrivna siffror från nya bilder utvecklads. Bilden förbehandlas och normaliseras innan den matas in i den bästa modellen för prediktion.

Ett exempel på förutsägelse visar att den utvecklade metoden kan korrekt identifiera handskrivna siffror från nya bilder.

## Begränsningar och framtida förbättringar

Eftersom hela datasetet inte användes kan det ha påverkat modellprestanda och en större del av dataset skulle kunna ge bättre generalisering.

Några av dom egna handskrivna siffrorna hade modellen svårt att prediktera rätt, framför allt urskilja nior och fyror.

Ytterligare finjustering av hyperparametrar och undersökning av andra modeller kan leda till förbättrad prestanda.

Vidare utvärdering och testning av den utvecklade metodiken på en större och mer varierad uppsättning av handskrivna siffror kan ge ytterligare insikter.

# Slutsatser

Den bästa modellen var en support Vector Classification modell som hade en precision på 94 procent. Modellen hade problem med att urskilja vissa siffror från varandra. Med finjustering och tester med flera olika modeller skulle kunna öka precisionen.

Det är viktigt att hålla isär tränings-, validerings, och testdata för att kunna analysera modellernas prestation. Det finns en risk att man under- eller överanpassar modellen vilket innebär dåliga resultat i slutändan

# Teoretiska frågor

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

*Sva*r: *Träningsdatan används för att träna sina ML modeller, Valideringsdatan används för att jämföra sina modeller och ta fram den bästa, På testdatan testar man den bästa 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 ”valideringsdataset”?**

*Svar: Kör Cross-Validation och se vilken modell som presterar bäst*

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

*Svar: regressionsproblem är att man ska prediktera en kontinuerlig variabel, några exempel på modeller som används är:*

*Linear Regression  
Lasso   
Ridge Regression*

**Modellerna kan användas tillexempel för att prediktera huspriser** **4. Hur kan du tolka RMSE och vad används det till: 𝑅𝑀𝑆𝐸 = √∑(𝑦𝑖 − 𝑦̂𝑖 ) 2 𝑖**

*Svar: RMSE är den genomsnittliga skillnaden mellan det verkliga värdet och vårt predikterade värde och kan används till att jämföra hur bra olika modeller predikterar.*

**5. Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?***Svar: Klassificieringsproblem handlar om att kunna dela upp observationerna i olika klasser, tillexempel om något kommer inträffa eller inte.  
Några modeller som använda är logistisk regression, beslutsträd och random forest.*

*Kan användas för att klassificera om ett mail är spam eller inte.*

**Confusion Matrix** *används för att visualisera hur modellen predikterar, båda rätta och fel svar.*

**6. Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.***Svar:* *K-means modellen är en unsupervised inlärningsteknik för att klassificera oetiketterade data genom att gruppera dem efter funktioner, snarare än fördefinierade kategorier.*

*Kan användas i tillexempel indelning av kunder i olika klasser*

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

*- I* ***Ordinal encoding*** *tilldelas varje unikt kategorivärde ett heltalsvärde och rankar dom.  
Till exempel är "röd" 1, "grön" är 2, och "blå" är 3*

*- Om vi har en kategori variabel för färger med tre möjliga värden: "röd", "blå" och "grön",* ***skulle one-hot encoding*** *se ut så här:*

*"röd": [1, 0, 0]  
"blå": [0, 1, 0]  
"grön": [0, 0, 1]*

*Man gör om i det här fallet förger till siffror*

*-* ***Dummy variabel encoding*** *är liknande one-hot encoding, men det innebär att endast n - 1 binära variabler skapas för n unika kategorier. Den fjärde kategorin representeras inte i modellen eftersom dess information kan extraheras från de andra.*

**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?**

*Svar: Julia har rätt*

**9. Kolla följande video om Streamlit: https://www.youtube.com/watch?v=ggDaRzPP7A&list=PLgzaMbMPEHEx9Als3F3sKKXexWnyEKH45&index=12 Och besvara följande fråga: - Vad är Streamlit för något och vad kan det användas till?**

*Svar: Streamlit är ett öppet källkodsramverk som används för att skapa användarvänliga webapplikationer för maskininlärning och dataanalys.*

# 

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Största utmaningen var att få modellen att prediktera rätt på dom handskrivna siffrorna, särskilt 4 och 9. Testade många olika pixelvärden för att urskilja siffran och bakgrunden.

1. Vilket betyg du anser att du skall ha och varför.

Jag satsar högt så jag säger VG

1. Något du vill lyfta fram till Antonio?

Trots en betydligt svårare kurs var det ett bra upplägg och utlärning.

Det finns fortfarande mycket att lära, men en bra start om man vill fortsätta med AI i framtiden.

# Källförteckning

1. <https://www.kaggle.com/code/xingewang/the-math-behind-linear-svc-classifier>
2. <https://www.coursera.org/articles/machine-learning-models>