Machine Learning

Digits Classification(Mnist data)



Gayathree Basarahalli

EC Utbildning

202503

# Abstract

Project fokuserar på klassificering av handskrivna data med hjälp av MNIST- datasetet. Support Vector machine(SVM), Random Forest model and Ensemble modell tränades med hjälp av hyperparameter-tuning. Den bästa modellen,SVM,valdes baserat på valideringsdata från de tränade modellerna. Slutligen utvärderades modellen på testdata med hjälp av accuracy score and confusion matrix.

**Nyckleord** MNIST-dataset, Exploratory Data Analysis(EDA), Principal Component Analysis(PCA),Machine Learning(ML), Support Vector Machine(SVM), Random Forest Modell, Ensemblemodell, hyperparameter tuning,Voting Classifier, Accuracy Score, Confusionmatris, Streamlit.

Innehållsförteckning

[Abstract 2](#_Toc193713647)

[1 Inledning 4](#_Toc193713648)

[2 Teori 5](#_Toc193713649)

[2.1 Support Vector Machine(Non linear classifier) 5](#_Toc193713650)

[2.2 Random Forest Model 5](#_Toc193713651)

[2.2.1 Nyckle hyperparametrar 6](#_Toc193713652)

[2.3 Ensemble Model 6](#_Toc193713653)

[2.3.1 Defination 6](#_Toc193713654)

[2.3.2 Typer av ensemblemodel 6](#_Toc193713655)

[2.3.3 Fördelar med Ensemble method 6](#_Toc193713656)

[3 Metod 7](#_Toc193713657)

[3.1 Import necessary Libraries: 7](#_Toc193713658)

[3.2 Data Förbehandling : 7](#_Toc193713659)

[3.3 Modellval 8](#_Toc193713660)

[3.3.1 SVM modell 8](#_Toc193713661)

[3.3.2 Random Forest modell 8](#_Toc193713662)

[3.3.3 Ensemble modell 8](#_Toc193713663)

[3.4 Utvärdering med test data 8](#_Toc193713664)

[4 Resultat och Diskussion 9](#_Toc193713665)

[4.1 Accuracy Score 9](#_Toc193713666)

[4.2 Utvärdering med test data 9](#_Toc193713667)

[4.2.1 Bästa modell Accuracy : 9](#_Toc193713668)

[4.2.2 Confusion Matrix : 10](#_Toc193713669)

[4.2.3 Klassificeringrapport 11](#_Toc193713670)

[5 Slutsatser 12](#_Toc193713671)

[5.1 Modeller Val 12](#_Toc193713672)

[5.2 Hyperparameter 12](#_Toc193713673)

[5.3 Data 12](#_Toc193713674)

[5.4 Träning precision och test precision 12](#_Toc193713675)

[Källförteckning 13](#_Toc193713676)

# Inledning

Maskininlärning är en process där data används för att hjälpa en dator att lära sig utan direkt mänsklig interaktion. Det anses vara en delmängd av artificiell intelligens (AI). Maskininlärning använder algoritmer för att identifiera mönster inom data, och dessa mönster används sedan för att skapa en datamodell som kan göra förutsägelser. Med mer träningsdata förbättras noggrannheten hos maskininlärningsmodeller.

Genom att använda ML med Mnist dataset som består av 70,000 bilder av handskrivna siffror from 0 till 9, kan olika maskininlärningmodeller tränas för att klassificera dessa siffror.

Syftet med denna rapport är att utveckla och testa olika machinelearning modeller och välja bästa modell för bildklassificering, särskilt för att identifiera hand skrivna siffror.

För att uppfylla syftet så kommer följande frågeställningar att besvaras:

1. Frågan baserat på vilka kriterier, valdes dessa modeller för träna Mnist data kommer att studeras?
2. Varför används vi hyperparameter för de välda modellen ?
3. Varför använde vi delmängd av data, och vad var fördelarna och nackdelarna med att använda mindre data?
4. Väd är skillnaden mellan träningsprecision och testprecision,och varför är de viktigt?

# Teori

## Support Vector Machine(Non linear classifier)

SVM (Support Vector Machine) är en övervakad maskininlärningsalgoritm som kan användas för klassificering eller regressionsuppgifter. Den fungerar genom att hitta ett hyperplan som bäst separerar data i olika klasser. SVM är särskilt effektiv i högdimensionella utrymmen och är väl lämpad för komplexa, icke-linjära beslutgränser.

Kernal: När datan är inte linjärt det är viktigast att använda kernal eftersom det avgör hur modellen hanterar och sepererar data.Att välja rätt kernal är viktigt för bästa resultat,eftersom olika kernal kan ha olika effekt med data egenskapar.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 1 SVM-klassificerare med en RBF-kärna

## Random Forest Model

Random Forest är en ensemble-inlärningsalgoritm av beslutträd som oftast tränas genom ”bagging” ibland ”pasting”. Den skapar en samling beslutsträd, var och en tränad på ett slumpmässigt urval av data, och sammanfattar deras förutsägelser för att förbättra noggrannheten. Denna metod hjälper till att minska överanpassning (overfitting) jämfört med ett enda beslutsträd, vilket gör den till ett kraftfullt verktyg inom maskininlärning.

### Nyckle hyperparametrar

1. N-estimators : Definierar antalet träd i skogen. Fler träd förbättrar vanligtvis modellens prestanda men ökar beräkningskostnaderna.
2. max\_features: Begränsar antalet funktioner som ska beaktas när en nod delas. Detta hjälper till att kontrollera övermontering.
3. max\_depth: Styr det maximala djupet för varje träd. Ett grunt träd kan underpassas medan ett djupt träd kan passa över. Så att välja rätt värde på det är viktigt.
4. max\_leaf\_nodes: Begränsar antalet lövnoder i trädet och kontrollerar därmed dess storlek och komplexitet.
5. max\_sample: Förutom funktionerna har vi en stor uppsättning träningsdatauppsättningar. max\_sample bestämmer hur mycket av datamängden som ges till varje enskilt träd.
6. min\_sample\_split: Anger det minsta antal sampel som krävs för att dela en intern nod.

## Ensemble Model

### Defination

Ensemble learning kombinerar förutsägelser från flera modeller (kallade "svaga inlärare" eller "basmodeller") för att göra en starkare, mer tillförlitlig förutsägelse. Målet är att minska felen och förbättra prestandan.

### Typer av ensemblemodel

1. Bagging (Bootstrap Aggregating): Modeller tränas oberoende av olika delmängder av data, och deras resultat beräknas i genomsnitt eller röstas om.
2. Boosting: Modeller tränas sekventiellt, var och en lär sig av misstagen i den tidigare modellen.

### Fördelar med Ensemble method

1. Minskad överanpassning: Genom att aggregera förutsägelser av flera modeller kan ensembler minska överanpassningen som enskilda komplexa modeller kan uppvisa.
2. Förbättrad generalisering: Det generaliserar bättre till osynliga data genom att minimera varians och bias.
3. Ökad noggrannhet: Att kombinera flera modeller ger högre prediktiv precision.
4. Robusthet mot buller: Det mildrar effekten av bullriga eller felaktiga datapunkter genom att utjämna förutsägelser från olika modeller.
5. Flexibilitet: Den kan fungera med olika modeller inklusive beslutsträd, neurala nätverk och stödvektormaskiner vilket gör dem mycket anpassningsbara.
6. Bias-Variance Tradeoff: Tekniker som packning minskar variansen, medan boosting minskar bias som leder till bättre övergripande prestanda.

# Metod

## Import necessary Libraries:

To work with machine learning model, we need to import all necessary libraries.Following are libraries we considered for Mnist data.

import numpy as np

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.datasets import fetch\_openml

from sklearn.metrics import accuracy\_score

from sklearn.model\_selection import GridSearchCV

from sklearn.decomposition import PCA

from sklearn.svm import SVC

from sklearn.model\_selection import StratifiedKFold

from sklearn.ensemble import RandomForestClassifier

import seaborn as sns

import joblib

from sklearn.ensemble import VotingClassifier

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import classification\_report

## Data Förbehandling :

MNIST-datat hämtades från Scikit-Learn datasets och importerades med funktionen fetch\_openml. Datasetet består av 70 000 handskrivna bilder med siffrorna 0 till 9, där varje bild har en upplösning på 28×28 pixlar. Indata lagrades i variabeln x, medan targetdatan lagrades i y. För att förstå databasen i en bra sätt genomfördes Exploratory Data Analysis.Detta inkluderade att undersöka datans struktur och visualisera bilden.

Det tar mycket tid att bearbeta hela datasetet, valdes en delmängd av 25 000 bilder för att minska beräkningstiden och ändå uppnå goda resultat.

x\_train = x[ :15000]

y\_train = y[ :15000]

x\_val = x[15000:20000]

y\_val = y[15000:20000]

x\_test = x[20000:25000]

y\_test = y[20000:25000]

Normalisera datan för att alla värderan ligger mellan 0 och 1.

x\_train = x\_train / 255.0

x\_val = x\_val / 255.0

x\_test = x\_test / 255.0

## Modellval

### SVM modell

I den modell vi använde RBF-kernel och Stratified K-Fold. Vi används Grid Search method för att testa olika kombinationer av hyperparametrar för att identifiera de bästa parametrarna för modellen. Med de bästa parametrarna tränades SVM-modellen med träning data.

Med valideringdata fick vi accuracy score 97.28 for SVM modell.

### Random Forest modell

Med GridSearchCV hittade vi de bästa hyperparametrarna från de alternativ vi angav , tränade modellen med träning data.

Med valideringdata fick vi accuracy score 95.8 for Random Forest modell.

### Ensemble modell

Vi kombinerade både SVM och Random Forest och använde en Voting Classifier för att välja de bästa parametrarna från båda modellerna. Den tränade modellen med träning data.

Med valideringdata fick vi accuracy score 96.66 for Ensemble modell.

Detta visar att SVM modell med användning av RBF-kernal gjörde bättre separation av siffrorna. Med test data Svm modell fick accuracy score 97.18, så sparat den bästa modellen med validering och test datan med joblib.

joblib.dump(best\_model, 'best\_model\_svm.pkl')

## Utvärdering med test data

För att utvärdera SVM-modellen användes Accuracy Score, Confusion Matrix och Classification Report. Accuracy Score ger en bättre förståelse för modellens beteende med ny data. Confusion Matrix ger en mer detaljerad bild av var och hur många gånger modellen gör felaktiga förutsägelser. Slutligen förklarar Classification Report tydligt Precision, Recall och F1-score.

# Resultat och Diskussion

## Accuracy Score

Tabell 1: Vi utvärderade resultaten från tre modeller med MNIST-datasetet. Accuracy score för de tre modellerna med validering data:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Hyper parameters |
| SVM | 97.28 | Kernal : rbf,C = 10, Cross validation = Stratified Cross validation (n\_splits =5), gamma = scale |
| Random Forest | 95.8 | n\_estimators = 600,max\_features = log2,min\_sample\_leaf = 1,cross validation = 3 |
| Ensemble | 96.66 | Voting = hard |

SVM modell presterade bäst med MNIST datasetet(med test data accuracy score : 97.18) eftersom rbf-kernel gör det möjligt att skapa beslutgränser for att skilja siffrorna mer effiktivt.

RandomForest Modell är också bäst med accuracy score men mindre ä SVM modell darför att den slumpmässiga naturen hos beslutsgränserna kan leda till lägre precision.

Vi kan visa att Kombinationen av två modeller är inte alltid en bra modell, men det har också med bra score med hård rostning.

## Utvärdering med test data

### Bästa modell Accuracy :

Vi vält att SVM presterade bäst med träning data, så vi försatter SVM model med test data och vi fick resultat 96.96% (Accuracy Score).

### Confusion Matrix :

A screenshot of a graph

AI-generated content may be incorrect.

*FIgure 2 Confusion Matrix : Har i bilden vi kan se de verkliga och förutsägda värderna samt hur True och hur många gånger modellen prresterande olika siffror felaktigt.*

### Klassificeringrapport

Denna reporten med test data ger en detailjerad utvärdering av modellens prestanda baserad på Precission, Recall,och F1-Score for värje siffror.

Siffrorna 0,1,7,8 här högsta precision och recall värderna , så modellen är bra på att prediktera denna siffror.

Siffrorna 4,5,9 här lägre precision och recall värderna , så det är svärare for modllen att prediktera denna siffor.För att förbattra den, modellen kan träna med mer data och använda mer hyperparametrar.

|  |  |  |  |
| --- | --- | --- | --- |
| Siffror | Precison | Recall | F1 score |
| 0 | 0.97 | 1.00 | 0.98 |
| 1 | 0.99 | 0.98 | 0.98 |
| 2 | 0.96 | 0.96 | 0.96 |
| 3 | 0.97 | 0.96 | 0.96 |
| 4 | 0.95 | 0.97 | 0.96 |
| 5 | 0.96 | 0.95 | 0.96 |
| 6 | 0.97 | 0.97 | 0.98 |
| 7 | 0.98 | 0.98 | 0.98 |
| 8 | 0.96 | 0.98 | 0.97 |
| 9 | 0.97 | 0.95 | 0.96 |

Tabell 2: Precision, Recall och F1 score av alla siffror(0- 9)

.

# Slutsatser

## Modeller Val

De tre modellerna valdes ut på grund av deras prediktions noggrannhet i klassificeringsproblem. SVM-modellen är mycket bra på att hantera icke-linjära resultat. Slumpmässit Random Forest är mycket flexibel och genom att använda fler träd ger den bra förutsägelser och denna modell är robust på överanpassade data. Ensemblemodell, för att kombinera styrkorna hos två modeller.

## Hyperparameter

Hyperparameters help the model to predict values more accurately and help data avoiding over fitting and underfitting and works well with new data.

## Data

"MNIST är ett stort dataset, och det tar mycket tid att beräkna och träna på hela datasetet. Att använda en delmängd av data ger nästan samma resultat om vi använder cross-validation eller stratified cross-validation-teknik."

## Träning precision och test precision

Efter att ha tränat en modell med träningsdata förklarar träningsprecisionen modellens noggrannhet. Genom att testa modellen med testdata får vi en tydlig bild av hur modellen förutsäger nya data. Om träningsprecisionen är hög och testprecisionen är låg, indikerar det att modellen overfitting, och vice versa. Därför är båda värdena viktiga för att avgöra om modellen overfitting eller underfitting.

# Källförteckning

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, & TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Prgomet, A. (2023, June 3). Maskininlärning [Video]. YouTube. <https://www.youtube.com/watch?v=zORv5vxrwok&list=PLgzaMbMPEHEx9Als3F3sKKXexWnyEKH45>

GeeksforGeeks. (2025). A comprehensive guide to ensemble learning. GeeksforGeeks. <https://www.geeksforgeeks.org/a-comprehensive-guide-to-ensemble-learning/?ref=header_outind>