Bakgrund

MNIST (Modified National Institute of Standards and Technology) är en klassisk dataset som innehåller handskrivna siffror (0-9) som är skannade och digitaliserade för maskininlärning. Datan har använts i forskning och för att utvärdera prestanda på olika maskininlärningsalgoritmer.

Syfte och Frågeställning

Syftet med detta projekt är att utforska och analysera MNIST-datasetet med hjälp av två olika maskininlärningsalgoritmer, nämligen Random Forest och Logistisk Regression. Målet är att jämföra prestandan hos dessa två modeller och att identifiera vilken av dem som ger de bästa resultaten på MNIST-datasetet.

- Vilken av de två modellerna, Random Forest eller Logistisk Regression, ger bäst resultat på MNIST-datasetet?
- Vilka faktorer påverkar modellernas prestanda på MNIST-datasetet?

Databeskrivning / EDA (Exploratory Data Analysis)

MNIST-datasetet innehåller 70 000 bilder på handskrivna siffror (0-9). Varje bild är 28 x 28 pixlar stor och varje pixel representeras av en gråskala från 0 till 255. Datan är redan delad i en träningsuppsättning med 60 000 bilder och en testuppsättning med 10 000 bilder. För att utforska datasetet gör vi en EDA-analys där vi undersöker datatypen för varje variabel, nollvärden och statistik över variablerna.

Metod och Modeller

I denna rapport har två modeller använts för att modellera MNIST-datasetet: Random Forest och Logistisk Regression. Båda modellerna är klassificeringsmodeller som används för att förutsäga vilken siffra som representeras av en bild.

Random Forest är en ensemble-metod som består av flera beslutsträd. För att skapa en förutsägelse använder algoritmen flera träd för att skapa en konsensusprognos. Varje beslutsträd byggs med hjälp av slumpmässigt valda undermängder av data och funktioner, vilket gör modellen robust mot överanpassning.

Logistisk Regression är en linjär klassificeringsmodell som använder en logistisk funktion för att modellera sannolikheten för att ett datapunkt tillhör en viss klass. Modellen använder viktade linjär regression för att förutsäga sannolikheten för att en viss siffra representeras av en bild.

Projekt resultat och analys

För att utvärdera modellerna har träningssatsen delats upp i en träningsuppsättning och en valideringsuppsättning. Modellerna har tränats på träningsuppsättningen och sedan utvärderats på valideringsuppsättningen. Resultaten av modellerna visas nedan:

Modell	Tränings-accuracy	Validerings-accuracy
Random Forest	1.000	0.968
Logistisk Regression	0.928	0.910

Slutsats och förslag på potentiell vidareutveckling

Baserat på tabellen ovan kan vi se att Random Forest-modellen presterade bättre än Logistisk Regression-modellen på valideringsuppsättningen. Random Forest-modellen hade en valideringsaccuracy på 0,968 jämfört med Logistisk Regression-modellens 0,91. Detta tyder på att Random Forestmodellen kan vara en bättre modell för att förutsäga siffrorna i MNIST-datasetet.

En potentiell vidareutveckling av detta projekt skulle vara att använda en mer avancerad modell, till exempel en djup inlärningsmodell som Convolutional Neural Networks (CNN) eller en ensemble av flera modeller. En annan möjlig förbättring skulle vara att använda dataaugmenteringstekniker för att öka mängden tillgängligt träningsdata.