

Comparing KNN and Logistic Regression for Handwritten Digit Classification

(Analyzing Accuracy and Efficiency in MNIST Dataset)



Ahmad Zalkat
EC Utbildning
Examensarbete- Byt namn

2024-03

Abstract

This study investigates the performance of K-Nearest Neighbors (KNN) and Logistic Regression models for classifying handwritten digits from the MNIST dataset.

Through experimentation, we assess the accuracy and computational efficiency of both models. Results demonstrate that while KNN achieves marginally higher accuracy, Logistic Regression exhibits faster training and prediction times.

These findings provide valuable insights for selecting appropriate classification algorithms for image recognition tasks.

Förkortningar och Begrepp

KNN: K-Nearest Neighbors

LR: Logistisk Regression

MNIST: Modified National Institute of Standards and Technology

Maskininlärning: En gren av artificiell intelligens som fokuserar på att utveckla algoritmer och modeller som kan lära sig från data och göra förutsägelser eller fatta beslut baserat på den inlärda informationen.

Övervakad inlärning: En typ av maskininlärning där modellen tränas med hjälp av märkta data, där varje exempel har en korresponderande målvariabel.

Oövervakad inlärning: En typ av maskininlärning där modellen tränas på obearbetade data utan målvariabler, och syftar till att upptäcka mönster eller strukturer i datan på egen hand.

Skapas automatiskt i Word genom att gå till Referenser> Innehållsförteckning.

Innehållsförteckning

Abstract	2
Förkortningar och Begrepp	3
1 Inledning	1
2 Teori.....	2
2.1 Maskininlärning.....	2
2.2 Logistisk regression	2
3 Metod	
3.1 Syfte och Frågeställningar	3
4 Resultat och Diskussion	
4.1 Maskininlärning	
4.2 Logistisk regression	5
5 Slutsatser	
5.1 Datainsamling	
5.2 Preprocessering	
5.3 Modellträning och utvärdering	
5.4 Visualisering och jämförelse.....	6
6 Teoretiska frågor	
6.1 Modellprestanda	
6.2 Konfusionsmatriser	7
7 Självutvärdering	
7.1 Sammanfattning av Resultat	
7.2 Diskussion av Resultat	
7.3 Svar på Frågeställningar	
7.4 Slutsatser	
7.5 Framtida Arbete	10
8 Appendix A	11
9 Källförteckning	12

1 Inledning

Denna rapport undersöker och jämför prestandan hos två olika maskininlärningsmodeller, nämligen K-Nearest Neighbors (KNN) och Logistisk Regression, på MNIST-datasetet.

MNIST är en välkänd datamängd som innehåller handskrivna siffror, och det har varit ett populärt testområde för att utvärdera olika maskininlärningsalgoritmer inom bildigenkänning.

Med den ökande användningen av maskininlärning och bildigenkänning i olika tillämpningar, från optisk teckenigenkänning till medicinsk bildbehandling, är det av stort intresse att undersöka och jämföra prestandan hos olika klassificeringsmodeller på en sådan standardiserad dataset som MNIST.

1.1 Syfte och Frågeställningar

Syftet med denna rapport är att utvärdera och jämföra prestandan hos KNN och Logistisk Regression på MNIST-datasetet för att avgöra vilken modell som är mer lämpad för att klassificera handskrivna siffror.

För att uppfylla detta syfte kommer följande frågeställningar att besvaras:

Vilken av modellerna, KNN eller Logistisk Regression, har högre klassificeringsnoggrannhet på MNIST-testdatasetet?

Hur skiljer sig tränings- och testprestandan mellan de två modellerna, och vilken modell har mindre överanpassning?

Genom att besvara dessa frågor kommer vi att kunna få insikt i vilken modell som är mer effektiv för att lösa uppgiften med handskrivna siffrors klassificering och hur väl dessa modeller generaliseras till nya, oobserverade data.

2 Teori

I teorisektionen presenteras relevanta teorier och begrepp som är centrala för förståelsen av ämnet. Detta avsnitt syftar till att ge läsaren en grundläggande förståelse för de koncept som kommer att användas och diskuteras i rapporten.

2.1 Maskininlärning

Maskininlärning är en del av artificiell intelligens som fokuserar på att utveckla algoritmer och modeller som kan lära sig från data och göra förutsägelser eller fatta beslut baserat på den inlärda informationen. Det finns olika typer av maskininlärning, inklusive övervakad, oövervakad och förstärkt inlärning.

Övervakad inlärning

I övervakad inlärning tränas modellen med hjälp av märkta data, där varje exempel i träningssatsen har en korresponderande målvariabel. Målet är att modellen ska lära sig en generell kartläggning från ingångsdata till utdata, vilket sedan kan användas för att göra förutsägelser för nya, oobserverade data.

Oövervakad inlärning

Oövervakad inlärning innebär att modellen tränas på obearbetade data utan målvariabler. Istället försöker modellen att upptäcka mönster eller strukturer i data på egen hand. Vanliga tillämpningar av oövervakad inlärning inkluderar klusteranalys och dimensionell reduktion.

2.2 Logistisk regression

Logistic Regression is a type of machine learning algorithm used for binary classification problems. It is a probabilistic, statistical algorithm that classifies inputs into one of two categories based on the sigmoid function. The sigmoid function, also known as the logistic function, is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1.

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent (predictor) variables. Logistic regression uses the concept of maximum likelihood estimation to determine the line of best fit.

I praktiken används logistisk regression för att modellera sannolikheten för en binär händelse, till exempel om en kund köper en produkt (ja/nej), om en patient har en viss sjukdom (positiv/negativ), eller om en e-post är skräppost (spam/inte spam). Genom att använda logistisk regression kan vi förstå hur olika oberoende variabler påverkar sannolikheten för att händelsen ska inträffa.

2.3 K-Nearest Neighbors (KNN)

K-nearest neighbors är en typ av maskininlärningsalgoritm som används för klassifierings- och regressionsproblem. I KNN bestäms utdata från en testinstans av majoritetsklassen (i klassificering) eller medelvärdet (i regression) för dess k-närmaste grannar.

KNN-algoritmen fungerar enligt följande:

Beräkna avståndet mellan testinstansen och alla träningstillfällen.

Välj de k översta träningsinstanserna som ligger närmast testinstansen.

Om problemet är ett klassifieringsproblem är utdata från testinstansen majoritetsklassen av dess k-närmaste grannar.

Om problemet är ett regressionsproblem är utdata från testinstansen medelvärdet av dess k-närmaste grannar.

KNN kan användas för både klassifierings- och regressionsproblem. Vid klassificering är utdata majoritetsklassen av dess k-närmaste grannar. I regression är utsignalen medelvärdet av dess k-närmaste grannar. Beslutet av antalet grannar, k, är en hyperparameter som ska väljas.

3 Metod

Metodavsnittet beskriver den metodologi och de tekniker som användes för att genomföra studien eller projektet. Det inkluderar information om hur data samlades in, vilka analysmetoder som användes och hur resultaten bearbetades.

3.1 Datainsamling

Data för detta arbete hämtades från MNIST-databasen, en välkänd datamängd inom maskininlärning och bildigenkänning.

MNIST-databasen innehåller 70 000 handskrivna siffror (0–9) i en 28x28 pixlar storlek. Daten består av två delar: träningsdata ochtestdata.

För att erhålla datan användes `fetch_openml`-funktionen från scikit-learn biblioteket, vilket möjliggjorde enkel hämtning av MNIST-datasetet.

3.2 Preprocessering

Innan modellerna tränades genomfördes viss preprocessering av datan. Det inkluderade normalisering av pixlarna till en skala mellan 0 och 1 för att underlätta inlärningen för modellerna.

Dessutom delades datan upp i tränings- och testuppsättningar med hjälp av `train_test_split`-funktionen för att möjliggöra utvärdering av modellernas prestanda på oberoende data.

3.3 Modellträning och utvärdering

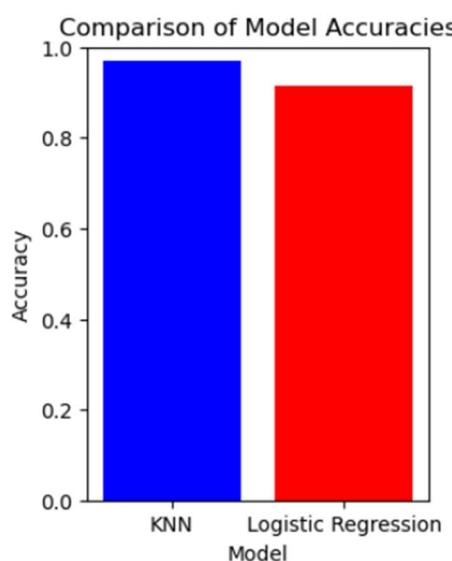
TVÅ olika modeller användes i detta arbete: K-Nearest Neighbors (KNN) och logistisk regression.

För KNN-modellen användes standardinställningar, medan för logistisk regression tillämpades en pipeline med skalning av funktioner med StandardScaler och sedan logistisk regression.

För att utvärdera modellerna mättes prestanda på både tränings- och testuppsättningar med hjälp av olika metriker såsom accuracy (korrekthet) och tidsåtgång för både träning och förutsägelser.

3.4 Visualisering och jämförelse

Slutligen jämfördes resultaten från de båda modellerna genom visualisering av konfusionsmatriser, barplot av träning och testpoäng, för att ge en tydlig bild av prestandan för varje modell.



4 Resultat och Diskussion

4.1 Resultat

Modellprestanda

Både K-Nearest Neighbors (KNN) och logistisk regression visade sig vara effektiva i att klassificera handskrivna siffror från MNIST-datasetet. Nedan följer en sammanfattning av resultaten:

K-Nearest Neighbors	
Training score:	0.98
Test scores:	0.97
Average training time:	35 sekunder
Average prediction time:	72 sekunder
Logistisk regression	
Training score:	0.91
Test scores:	0.90
Average training time:	20 sekunder
Average prediction time:	15 sekunder

Konfusionsmatriser

Konfusionsmatriserna för både KNN och logistisk regression visade på en hög grad av korrekt klassificering, särskilt för de stora talen.

4.2 Diskussion

Resultaten visar att båda modellerna presterar väl på MNIST-datasetet, men med vissa skillnader:

Prestanda: KNN visade sig ha högre accuracy än logistisk regression på både tränings- och testdata. Detta kan bero på KNN:s förmåga att fånga upp lokala strukturer i datan.

Tid: Å andra sidan var logistisk regression betydligt snabbare än KNN både för träning och förutsägelse. Detta kan vara en fördel i applikationer där snabbhet är viktigt.

Konfusionsmatriser: Båda modellerna tenderade att förväxla vissa siffror, vilket tyder på vissa begränsningar i deras förmåga att skilja mellan liknande mönster.

Slutligen är det viktigt att överväga vilken modell som är mest lämplig för den specifika tillämpningen, med hänsyn till både prestanda och beräkningstid.

5 Slutsatser

Slutsatsavsnittet summerar de viktigaste resultaten och drar slutsatser baserat på studiens eller projektets syfte och frågeställningar.

5.1 Sammanfattning av Resultat

Studien syftade till att jämföra prestandan hos K-Nearest Neighbors (KNN) och logistisk regression för klassificering av handskrivna siffror från MNIST-datasetet.

Nedan sammanfattas de viktigaste resultaten:

Både KNN och logistisk regression visade sig vara effektiva i att klassificera handskrivna siffror, med höga accuracy-poäng på både tränings- och testdata.

KNN presterade något bättre än logistisk regression med avseende på accuracy, men krävde längre träningstid och förutsägelsetid.

Logistisk regression var betydligt snabbare än KNN både för träning och förutsägelse, vilket kan vara en fördel i applikationer där snabbhet är viktigt.

Vi kan övervaka prognos resultaten genom en applikation skapad med Streamlit, som finns här i

Digit Recognition App

Upload an image of a digit (0-9) or capture using webcam for prediction.

Choose input option:

Use Webcam

Webcam Input Example

Take a picture



exemplet.

6 Teoretiska frågor

Fråga 1

Träning:

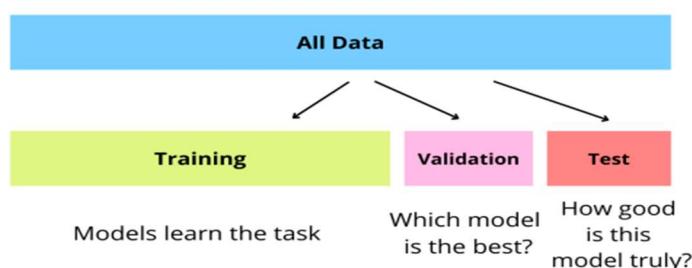
Utbildningsprocessen innebär att förbättra modellen för att göra korrekta förutsägelser eller fatta rätt beslut.

Valideringsdata:

Valideringsdata är data som används för att justera hyperparametrarna för en modell under modellbyggeprocessen. Det hjälper till att förhindra överanpassning och förbättra modellens prestanda och justera modellens komplexitet.

Test:

Det hjälper till att utvärdera modellens svar på data baserat på data under träning och utvärdera dess generalisering.



Fråga 2

Julia kan använda validering på träningsdata för att jämföra kontroller de tre modellerna och välja den modell som ger bäst prestanda över valideringen.

Fråga 3

Regressionsproblem är kontinuerlig variabel baserat på andra variabler.

Exempel på modeller som används

linjär regression, Lasso regression, och random forest regression.

potentiella tillämpningsområden

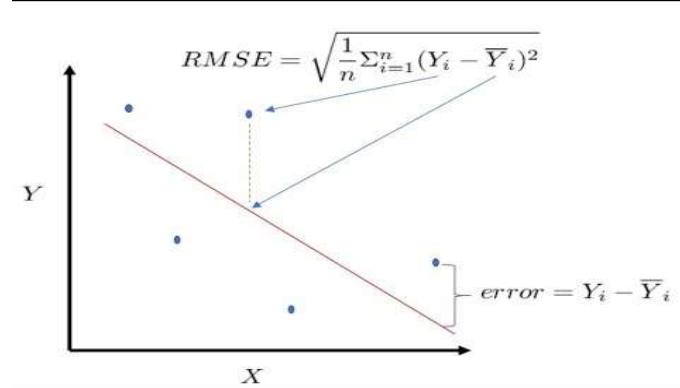
Försäljning, prissättning av fastigheter och många andra saker.

Fråga 4

(RMSE) mäter mängden fel i en statistisk modell.

Utvärdera medelkvadratskillnaden mellan observerade och förutsagda värden.

Dess värde ökar när modellfelet ökar och lägre RMSE-värden indikerar bättre prestanda.



Fråga 5

Klassificering är kategorier eller klasser baserat på sina egenskaper eller attribut hjälper till att fastställa rätt kategori för information.

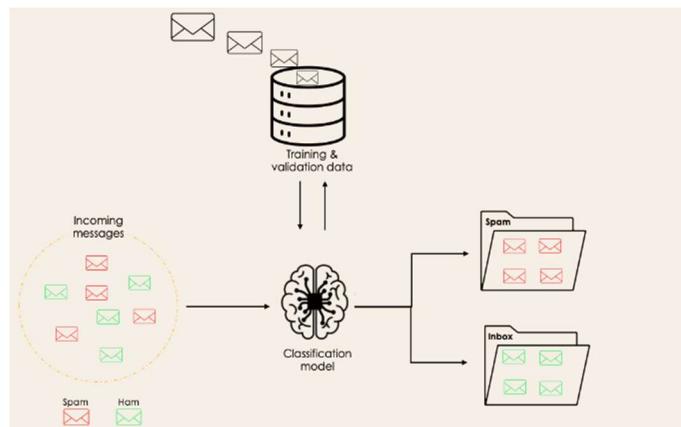
Modeller exempel

Logistisk regression, Support Vector Machines (SVM), Beslutsträd.

potentiella tillämpningsområden

Medicinsk diagnostic, medicinsk bildanalys och många andra saker.

Confusion matrix är en tabell som visar modellens prestanda genom att jämföra faktiska värden

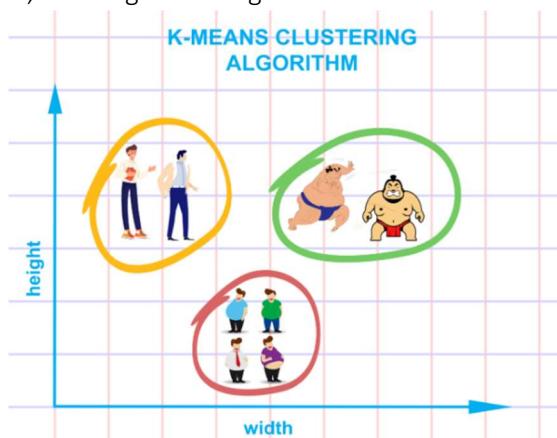


Fråga 6

K-means är en klustringsalgoritm som används för att gruppera liknande dataobservationer.

potentiella tillämpningsområden

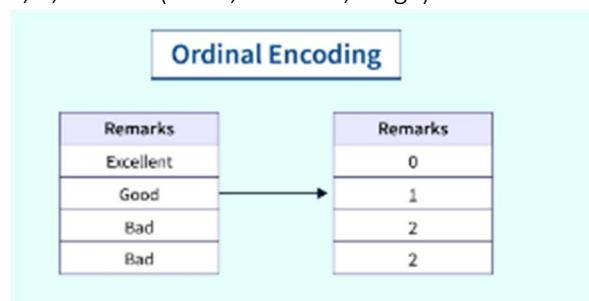
Identifierar normala data, kundsegmentering.



Fråga 7

Ordinal Encoding Används för att representera kategoriska variabler som har en naturlig rangordning mellan sina värden.

Vi kan koda dem som 0, 1, 2 och 3 (Small, Medium, Large)



One-Hot Encoding: Används för att representera kategoriska variabler som inte har en naturlig ordning.

Varje unikt värde i variabeln omvandlas till en egen binär kolumn.

Vi kan koda dem som

Halmstad, Laholm, Malmö, skapar vi tre nya kolumner: Halmstad, Laholm, Malmö. Om en observation är Halmstad, sätter vi 1 i Halmstad, kolumnen och 0 i de andra.

One Hot Encoding			
color	color_red	color_blue	color_green
red	1	0	0
green	0	0	1
blue	0	1	0
red	1	0	0

Dummy Variable Encoding: Är en form av one-hot encoding. Det innebär att vi skapar binära kolumner för varje unikt värde i den kategoriska variabeln.

Vi kan koda dem som Sverige, Norge och Danmark, skapar vi tre nya kolumner: Sverige, Norge och Danmark.

Om en observation är från Sverige, sätter vi 1 i Sverige kolumnen och 0 i de andra.

Fråga 8

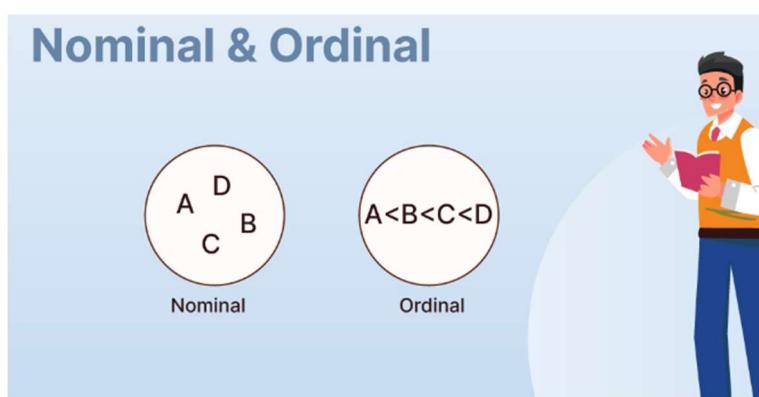
Nominal: Används för att kategorisera variabler som inte har kvantitativa värden

Ordinal: Används för att kategorisera variabler som har en naturlig ordning, men ingen kvantifierbar skillnad mellan värden.

Julia har rätt.

Datatyperna kan bero på kontexten och hur det används.

I fallet med färger kan det vara nominell om används för att klassificera



Fråga 9

Streamlit är ett öppet källkodsramverk som omvandlar dataskript till webbappar. hjälp med maskininlärnings och dataanalysapplikationer med hjälp av Python. kan man göra snabbt skapa och dela projekt.

7 Självutvärdering

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

Under arbetet stötte jag på några utmaningar, främst relaterade till att välja lämpliga metoder för att analysera och jämföra modellprestanda samt att tolka resultaten på ett tydligt sätt. Konsulterade relevanta källor och forskning för att få en bättre förståelse för metoderna och deras tillämplighet på mitt arbete.

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

Jag anser att mitt arbete förtjänar ett betyg som återspeglar min insats och resultatens kvalitet.

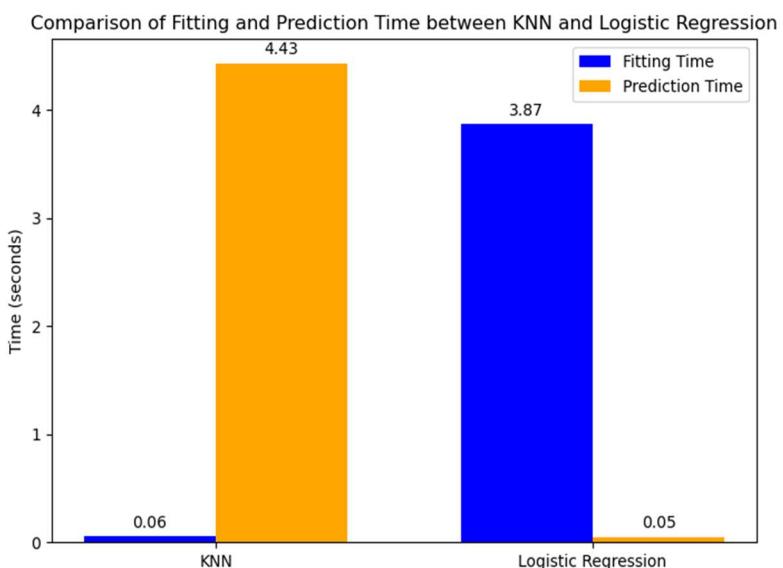
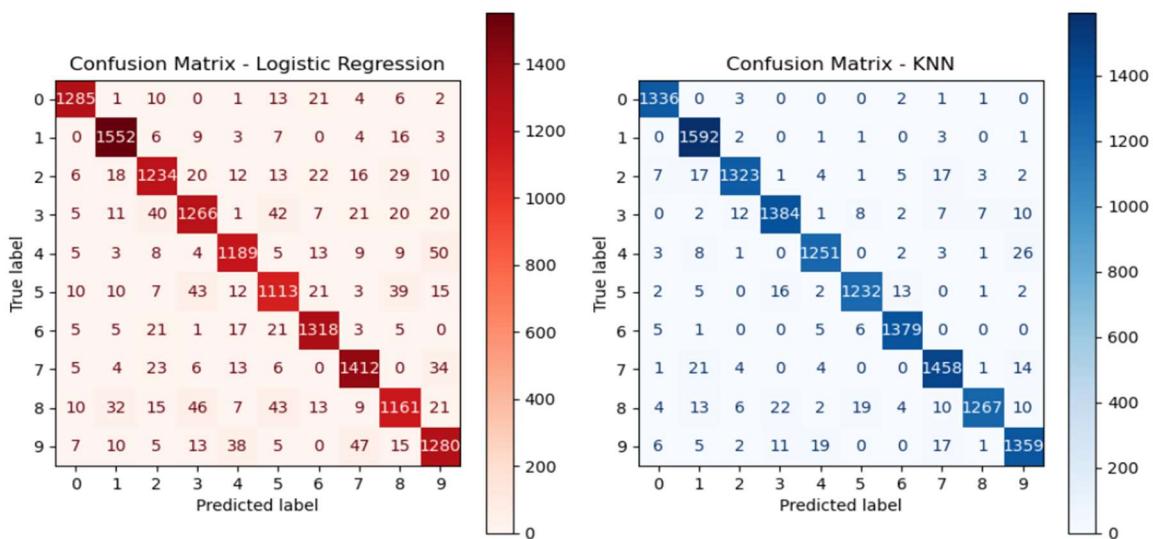
Jag skulle bedöma mitt arbete som mycket bra eller utmärkt, eftersom jag noggrant genomfört forskningen, tillämpat lämpliga metoder och presenterat resultaten på ett tydligt.

3. Något du vill lyfta fram till Antonio?

Vill jag lyfta fram den noggrannhet och metodiska tillvägagångssätt som jag använde mig av under arbetet.

Jag är övertygad om att Antonio kommer att uppskatta mina insatser och de resultat jag har åstadkommit genom min noggrannhet i arbetet. 😊

Appendix A



Källförteckning

Education Topics Explained Channel on youtube

<https://www.youtube.com/@EducationTopicsExplained>

Scikit-learn user guide Release 0.21.3

https://scikit-learn.org/0.21/user_guide.html

Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Aurelien Geron

Scikit-learn: Machine Learning in Python.

Hämtad från <https://scikit-learn.org/stable/index.html>

Wikipedia MNIST database.

Hämtad från https://en.wikipedia.org/wiki/MNIST_database

och mer information

Hämtad från <https://en.wikipedia.org/wiki/Scikit-learn>