

# Handwritten Digits Recognition



Garima Choudhary

EC Utbildning

2024 March

## Abstract

The primary objective of this project is to develop and evaluate a machine learning model for digit recognition using the classical MNIST dataset. Digit recognition entails training a machine learning system to identify handwritten digits from various sources such as emails, bank cheques, images, and more. This project aims to implement a model capable of recognizing digits in real-world scenarios, including online handwriting recognition on computer tablets, identifying number plates of vehicles, processing bank cheque amounts, and extracting numeric entries from hand-filled forms.

**Keywords:** Digit recognition, Image Processing, MNIST Dataset, Model Evaluation, Model Training, Model Evaluation and Performance Metrics.

## **Acknowledgement**

I would like to express my gratitude towards Antonio, my Mentor, for his continuous support and efforts. And I want to say thank you to everybody in my class for being so kind and generous.

## Table Of Contents

Abstract.....	2
Acknowledgement.....	3
1 Inledning.....	6
1.1 Syfte.....	6
1.2 Frågeställning.....	6
1.3 TextensDisposition.....	6
2 Theori.....	8
2.1 Data Overview.....	8
2.2 Problem Statement.....	8
2.3 KNN.....	8
2.4 Logistic Regression.....	8
2.5 SVM.....	9
2.5 Performance Measure .....	9
2.5.1 Confusion Matrix.....	9
3 Methodology.....	10
3.1 Data Collection.....	10
3.2 Data Preprocessing.....	10
3.3 Feature Extraction.....	10
3.4 Model Training.....	10
3.5 Model Evaluation.....	10
4 Result & Performance.....	11

4.1 Support Vector Machine (SVM).....	11
4.2 K-Nearest Neighbors (KNN).....	11
4.3 Logistic Regression (LR).....	12
5 Final Conclusion.....	13
Theoretical questions.....	14
Streamlit web App.....	20
Självutvärdering.....	23
Appendix A.....	24
Table of Figures.....	24
Källförteckning.....	25

# 1. Inledning

Genom denna analytiska forskning kommer jag att försöka besvara dessa frågor, vilket motiverade mig att göra. Eftersom jag vet av egen erfarenhet att jag tenderar att bli skeptisk till nästan varje säljsamtal och marknadsföringsbudskap, vilket förmodligen inte gör mig till en enkel eller idealisk konsument. Handskriven sifferigenkänning är ett betydande problem inom maskininlärning och datorseende. Denna rapport presenterar en studie fokuserad på att använda K-Nearest Neighbors (KNN) och logistisk regression klassificerare för handskrivna siffror igenkänning.

## 1.1 Syfte

Detta projekt utforskar effektiviteten hos modeller för maskininlärning och rangordnar deras prestanda. KNN och logistisk regression klassificerare för att korrekt känna igen handskrivna siffror. Genom att jämföra prestandan för dessa två klassificerare strävar vi efter att förstå deras styrkor och begränsningar i denna uppgift.

## 1.2 Frågeställning

Den här studiens centrala fråga är vad som är motivationen bakom analytisk forskning, är det värt det?

Varför måste vi känna igen de handskrivna siffrorna?

Kan en sådan typ av teknik vara en game changer i framtiden?

Kan klassificerare för KNN och Logistic Regression effektivt klassificera handskrivna siffror?

## 1.3 Textens disposition

Denna rapport inleds med en introduktion till problemet med handskrivna siffror igenkänning och motivationen bakom studien. Den fördjupar sig sedan i den teoretiska bakgrunden för klassificerare för KNN och Logistic Regression. Metodavsnittet beskriver datainsamling, förbearbetning, modellutbildning och utvärderingsprocesser. Efter det presenterar resultat- och diskussionssektionen resultaten av studien, inklusive klassificerarnas prestationsmått och deras implikationer. Slutligen avslutas rapporten med en sammanfattning av de viktigaste resultaten och insikterna från studien.

## **2.Theory**

### **2.1 Data overview**

MNIST-datauppsättningen är en mycket använd riktmärkesdatauppsättning inom området maskininläring, särskilt för uppgifter relaterade till bildklassificering. Den består av 28x28

pixlar gråskalebilder av handskrivna siffror (0-9) tillsammans med motsvarande etiketter. Datauppsättningen innehåller 60 000 träningsbilder och 10 000 testbilder, vilket gör den lämplig för utbildning och utvärdering av maskininlärningsmodeller.

## **2.2 Problem statement**

Här har vi att göra med klassificeringsproblemet. Klassificering kan vara binär eller multiklass, binär klassificering består av två värden antingen sant eller falskt, dvs. den binära klassificeraren skiljer bara mellan två klasser. Medan binära klassificerare fungerar mellan två klasser, kan multiklassklassificerare särskilja mer än två klasser

Random forest, Logistic Regression Classifier, K-Nearest Neighbors (KNN) och Logistic Regression är vanliga klassificerare inom maskininläring. I detta forskningsarbete kommer jag att använda dessa nedan nämnda klassificerare.

### **2.3.1 KNN**

K \_ Nearest Neighbor är en instansbaserad övervakad maskininlärningsalgoritm som klassificerar objekt baserat på majoritetsklassen av deras k-närmaste grannar. Denna kan användas för att lösa både klassificerings- och regressionsproblem.

### **2.3.2 Logistic Regression**

Logistisk regression är en probabilistisk klassificeringsalgoritm som modellerar sannolikheten för ett binärt utfall med hjälp av en logistisk funktion.

### **2.3.3 SVM**

Support Vector Machine (SVM) är en kraftfull maskininlärningsalgoritm som används för klassificerings- och regressionsuppgifter. Det fungerar genom att hitta det optimala hyperplanet som bäst separerar olika klasser i datautrymmet. SVM är särskilt effektivt i högdimensionella utrymmen och används flitigt i olika applikationer, inklusive handskriven sifferigenkänning.

## **2.4 Performance Measures**



As we are dealing with a classification problem here so we can use ROC Curve, AUC-ROC, Specificity, Accuracy, Recall, F1-Score and Precision. I am going to use confusion matrix and F-score for evaluation of used classifiers, in my research work.

### 2.4.1 Confusion Matrix

En förvirringsmatris är ett prestationsutvärderingsverktyg som används i maskininlärning för att bedöma prestandan hos en klassificeringsmodell. Den tillhandahåller en tabellrepresentation av faktiska kontra förutspådda klassetiketter, vilket möjliggör analys av modellens noggrannhet, precision, återkallelse och andra mätvärden. Matrisen är organiserad i rader och kolumner, där varje rad motsvarar de faktiska klassetiketterna och varje kolumn motsvarar de förutsagda klassetiketterna. Genom att undersöka ingångarna i förvirringsmatrisen kan man få insikter i modellens prestanda, inklusive dess förmåga att korrekt klassificera instanser som tillhör olika klasser och dess tendens att göra specifika typer av fel.

## 3.Method

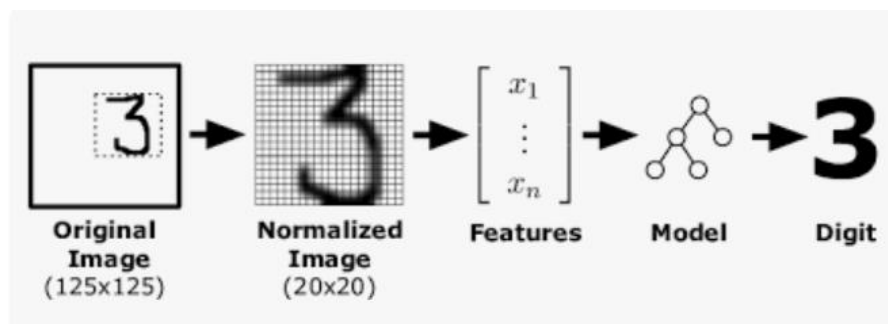


Figure: 1

### 3.1 Data Collection:

Skaffa en datauppsättning med handskrivna siffror, till exempel MNIST-datauppsättningen. importera alla viktiga bibliotek.

### 3.2 Preprocessing

Normalisera pixelvärden och utför alla nödvändiga bildförbättringar

### 3.3 Feature Extraction

Extrahera funktioner från bilderna, såsom pixelintensiteter eller funktioner på högre nivå erhållna genom tekniker som Principal Component Analysis (PCA).

### 3.4 Model Training

Träna klassificerare av KNN, SVM och Logistic Regression på de extraherade funktionerna med hjälp av märkta data.

### 3.5 Model Evaluation

Utvärdera klassificerarnas prestanda med hjälp av mätvärden som noggrannhet, precision, återkallelse och F1-poäng på en separat testuppsättning.

## 4. Result and performances

Based on the F1 scores and confusion matrices obtained from the evaluation, we observed the following:

Model	F- Score
SVM	.975
KNN	.954
LR	.883

Table:1

### 4.1 Support Vector Machine (SVM)

Uppnådde en F1-poäng på 0,975. Förvirringsmatrisen avslöjade hög precision och återkallningsvärden över alla klasser, vilket indikerar stark prestanda i korrekt

klassificering av siffror. SVM:s förmåga att hantera högdimensionell data och fånga komplexa samband mellan funktioner bidrog till dess utmärkta prestanda.

## **4.2 K-Nearest Neighbors (KNN)**

Uppnådde en F1-poäng på 0,954. Medan KNN visade konkurrenskraftiga prestanda, visade dess förvirringsmatris något lägre precision och återkallelse jämfört med SVM, särskilt när det gällde att skilja mellan liknande siffror. KNN:s beroende av lokal information och avståndsmått kan ha lett till vissa felklassificeringar, särskilt i fall med bullriga eller överlappande datapunkter.

## **4.3 Logistic Regression (LR)**

Uppnådde en F1-poäng på 0,883. LR uppvisade lägst F1-poäng bland de tre modellerna. Förvirringsmatrisen framhävde anmärkningsvärda avvikelser, särskilt när det gäller att skilja mellan vissa sifferpar. LR:s linjära beslutsgräns kan ha begränsat dess förmåga att fånga komplexa mönster som finns i data, vilket resulterar i suboptimal prestanda jämfört med SVM och KNN.

## 5. Conclusion

Sammanfattningsvis visar alla utvalda klassificerare, nämligen SVM, K-Nearest Neighbors och Logistic Regression, lovande prestanda i handskriven sifferigenkänning. Support Vector Machine (SVM) överträffade både K-Nearest Neighbors (KNN) och Logistic Regression (LR) modeller när det gällde att klassificera handskrivna siffror. Med en F1-poäng på 0,975 uppvisade SVM överlägsen prestanda, stödd av hög precision och återkallningsvärden över alla klasser i förvirringsmatrisen. KNN följde tätt med en F1-poäng på 0,954, medan LR släpade efter med en poäng på 0,883. Dessa resultat indikerar SVMs robusthet när det gäller att fånga intrikata mönster i högdimensionella data, vilket gör det till det föredragna valet för handskrivna sifferigenkänningsuppgifter. Ytterligare forskning kan utforska ytterligare tekniker för extraktion av funktioner och modellarkitekturer för att förbättra klassificeringsprestanda.

## Part-2

### Teoretiska frågor

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

**Svar:** Kalle delar upp sin data i "Training", "Validation" och "Test" eftersom detta är den viktigaste delen av val av maskininlärningsmodell. eftersom vi genom detta kontrollerar prestandan för den valda modellen.

**Träning:** Exakta träningsdata hjälper modellen att lära sig de rätta mönstren.

**Validering:** validering av data hjälper utvecklare att finjustera modellen korrekt

**Test:** testdata ger tillförlitliga mätvärden så att de med säkerhet kan distribuera sin AI-lösning.

**Fråga 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 ”validerings-dataset”?**

**Svar:** I avsaknad av en dedikerad valideringsdatauppsättning (En valideringsdatauppsättning är väsentlig för att utvärdera modellprestanda, ställa in hyperparametrar, förhindra överanpassning), kommer Julia att försöka välja den lämpligaste modellen bland linjär regression, lassoregression och slumpmässig skog. Genom att analysera deras prestanda på träningsdata. Detta innebär att utvärdera mätvärden som Mean Squared Error (MSE), Root Mean Squared Error (RMSE) eller R-squad för att mäta hur väl varje modell passar träningsdatan. Dessutom överväganden som komplexiteten i problem, tolkbarheten av modellerna och potentialen för över- eller underanpassning spelar avgörande roller för att fatta beslutet. Genom att noggrant väga dessa faktorer kan Julia välja den modell som har den optimala balansen mellan prestanda och enkelhet

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

**Svar:** Regression är en metod för att förstå sambandet mellan oberoende variabler eller egenskaper och en beroende variabel eller utfall. Utfall kan sedan förutsägas när sambandet mellan oberoende och beroende variabler har uppskattats.

Exempel:

vi kan säga att ålder och längd kan beskrivas med en linjär regressionsmodell. Eftersom en persons längd ökar när åldern ökar, har de ett linjärt samband. Regressionsmodeller används ofta som statistiskt bevis på påståenden om vardagliga fakta

applikationer:

Regression är en statistisk metod som används inom finans, investeringar och andra discipliner som försöker bestämma styrkan och karaktären av sambandet mellan en beroende variabel (vanligtvis betecknad  $Y$ ) och en serie andra variabler (kända som oberoende variabler).

**Fråga 4. Hur kan du tolka RMSE och vad används det till:**

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

**Figure :2**

**Svar:** (Root mean squared error, As its name implies, the RMSE is the square root of the mean squared error (MSE)). Den mäter den genomsnittliga skillnaden mellan värden som förutspås av en modell och de faktiska värdena. Den ger en uppskattning av hur väl modellen kan förutsäga målvärdet (noggrannhet). Ju lägre värde på Root Mean Squared Error, desto bättre modell.

Nn= Num. Of observations

Yi = Actual value

Ycap= Predicted Value

**Aplikationer:**

RMSE används inom många områden som meteorologi, experimentell psykologi, ekonomi och beräkningsneurovetenskap.

**Fråga 5. Vad är "klassificeringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en "Confusion Matrix"?**

**Svar:** Klassificering är en övervakad maskininlärningsmetod där modellen försöker förutsäga rätt etikett för en given indata. I klassificeringen är modellen fullt utbildad med hjälp av träningsdata, och sedan utvärderas den på testdata innan den används för att utföra förutsägelse på ny osynlig data.

Populära modeller som används för klassificering inkluderar:

Logistisk regression, beslutsträd, Random Forest och Support Vector Machines (SVM)

**Aplikationer:**

potentiella användningsområden är sjukvård (sjukdomsdiagnos), bildigenkänning (objektidentifiering) och identifiering av skräppost och fastställande av handstilstecken och så vidare.

En förvirringsmatris presenterar en tabelllayout över de olika utfallen av förutsägelsen och resultaten av ett klassificeringsproblem och hjälper till att visualisera dess utfall. Den plottar en tabell över alla förutsagda och faktiska värden för en klassificerare

**Fråga 6. Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.**

**Svar:** K- betyder klustring, en del av det oövervakade lärandet. Den används för att gruppera liknande datapunkter i en process som kallas klustring. Clustering hjälper oss att förstå vår data på ett unikt sätt – genom att gruppera saker i – du gissade rätt – kluster.

K-Means-klustring används i en mängd olika exempel eller affärsfall i verkliga livet, som:

Akademisk prestation, Diagnostiksystem, Sökmotorer och trådlösa sensornätverk.

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

**Svar:** För dessa måste vi först tala om de kategoriska uppgifterna. Kategoriska variabler representeras vanligtvis som "strängar" eller "kategorier" och är ändliga till antalet. Här är några exempel:

Staden där en person bor: Delhi, Mumbai, Ahmedabad, Bangalore, etc.

Avdelningen en person arbetar på: Ekonomi, Personal, IT, Produktion.

Den högsta examen en person har: High school, Diploma, Bachelors, Masters, PhD.

Klasskategorisdata kan delas in i två kategorier som Nominella data och Ordinaldata (Kategorierna har en inneboende ordning) och Nominella data och (Kategorierna har ingen inneboende ordning).



Ordinalkodning I Ordinaldata bör man vid kodning behålla informationen om i vilken ordning kategorin tillhandahålls. Som i exemplet ovan ger den högsta graden en person besitter viktig information om hans kvalifikationer. En examen är en viktig egenskap för att avgöra om en person är lämplig för en tjänst eller inte.

När vi kodar nominell data måste vi ta hänsyn till en funktions närvaro eller frånvaro. I ett sådant fall föreligger inget begrepp om ordning. Till exempel staden en person bor i. För uppgifterna är det viktigt att behålla var en person bor. Här har vi ingen ordning eller ordning. Det är lika om en person bor i Delhi eller Bangalore.

### One-hot-kodning


Index	Animal	One-Hot code	Index	Dog	Cat	Sheep	Lion	Horse
0	Dog		0	1	0	0	0	0
1	Cat		1	0	1	0	0	0
2	Sheep		2	0	0	1	0	0
3	Horse		3	0	0	0	0	1
4	Lion		4	0	0	0	1	0

Figure :3

Vi använder denna kategoriska datakodningsteknik när funktionerna är nominella (inte har någon ordning). I en varm kodning skapar vi en ny variabel för varje nivå av en kategorisk funktion. Varje kategori mappas med en binär variabel som innehåller antingen 0 eller 1. Här representerar 0 frånvaron och 1 representerar närvaron av den kategorin. Dessa nyskapade binära funktioner kallas dummyvariabler.

### Dummy variabel kodning

Dummy-kodningsschema är som en-hot-kodning. Denna kategoriska datakodningsmetod omvandlar den kategoriska variabeln till en uppsättning binära variabler (även känd som dummyvariabler). När det gäller one-hot-kodning, för N kategorier i en variabel, använder den N binära variabler. Dummy-kodningen är en liten förbättring jämfört med en-hot-encoding. Dummy-kodning använder N-1-funktioner för att representera N etiketter/kategorier.

För att förstå detta bättre låt oss se bilden nedan. Här kodar vi samma data med både en-hot-kodning och dummy-kodningstekniker. Medan one-hot använder 3 variabler för att representera data medan dummy-kodning använder 2 variabler för att koda tre kategorier

Column	Code	Column	Code
A	100	A	10
B	010	B	01
C	001	C	00

One- Hot Coding                      Dummy Code

**Figure:4**

**Fråga 8. Göran påstår att datan antingen är "ordinal" eller "nominal". Julia säger att detta måste tolkas. Hon ger ett exempel med att färger såsom {grön, röd, grön} 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:** Det verkar som att Göran och Julia, båda har rätt. Göran skiljer mellan "ordinala" och "nominella" datatyper, där ordningsdata har en tydlig ordning, medan nominella data inte har det. Julia betonar vikten av tolkning och illustrerar att även om färger som grönt och rött kan vara nominella, kan sammanhanget tilldela en ordinal tolkning, som i exemplet med attraktivitet på en fest.

**Fråga 9. Vad är Streamlit för något och vad kan det användas till?**

**Svar:** Streamlit är ett Python-baserat bibliotek som gör det möjligt för datavetare att skapa maskininlärningsapplikationer utan ansträngning. Det ger ett intuitivt och användarvänligt gränssnitt för att integrera sparade modeller, visa beskrivande text, visualisera data och modellprestanda och modifiera modellgångar via UI-sidofält.

Streamlit används för att snabbt utveckla webbapplikationer för prediktiv analys, som ofta bara tar några timmar att skapa. Det integreras sömlöst med populära Python-bibliotek som NumPy, Pandas, Matplotlib och mer, vilket underlättar effektiv och strömlinjeformad utveckling av datadrivna applikationer.

## **Part – 3 Streamlit Web Application**

**I denna** del av forskningen har jag försökt bygga en webbapplikation för igenkänning av handskrivna siffror. För att skapa denna applikation har jag använt KNN-modellen. I vissa fall fungerade applikationen bra men för vissa fungerade den inte på rätt sätt. Jag skulle vilja dela med mig av några av mina erfarenheter av mitt Streamlit-arbete.

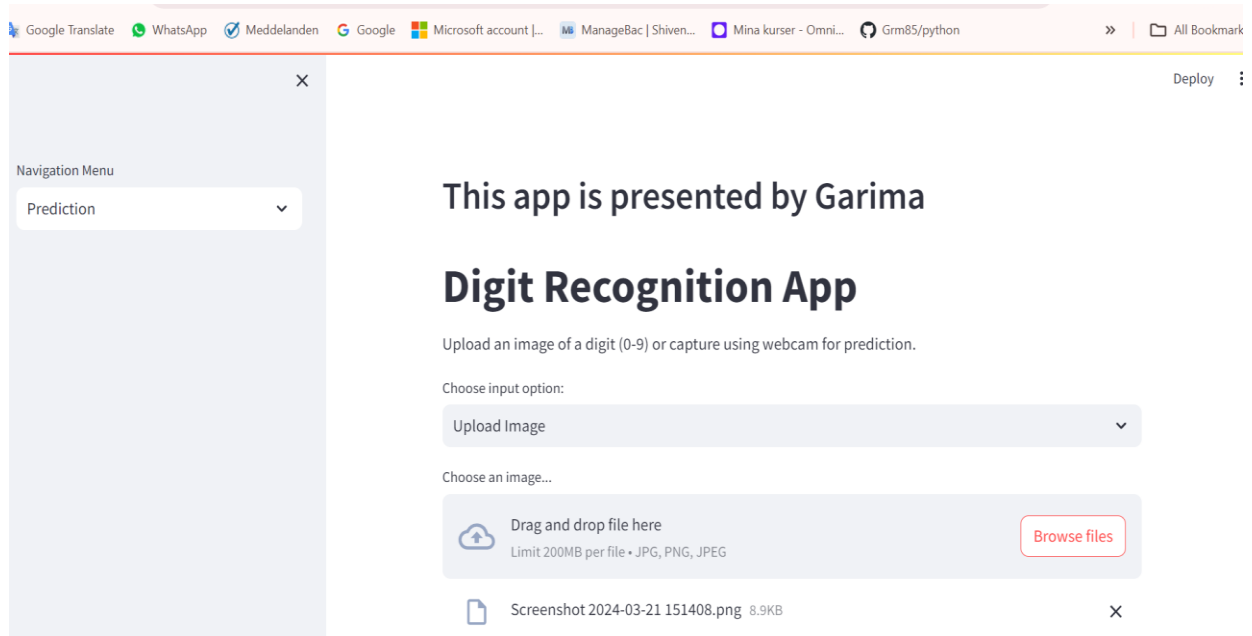


Figure:5



Uploaded Image.

Uploaded Image.

Prediction: 5.0

Figure :6



Uploaded Image.

Prediction: 8.0

Figure:7

## Självutvärdering

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

**Svar:** Challenges While MNIST is considered a standard dataset, it's not without its challenges. Achieving high accuracy on MNIST is a relatively straightforward task for modern machine learning models. As a result, researchers often use more challenging datasets to evaluate the robustness and generalization capabilities of models

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

**Svar:** Som mitt arbete säger dig!

3. Något du vill lyfta fram till Antonio?

**Svar:** Nej!

## Appendix A

Table 1 F-Score .....	
Figure 1 RMSE Formulae.....	
Figure 2 one hot coding.....	
Figure 3 Normal coding.....	
Figure 4 Streamlit demonstration.....	
Figure 5 Streamlit demonstration.....	
Figure 6 Streamlit demonstration.....	



## Källförteckning

*What is machine learning? Definition, types, and examples.* (2023, September 11).

Coursera. [https://www.coursera.org/articles/what-is-machine-learning?utm\\_medium=sem&utm\\_source=gg&utm\\_campaign=B2C\\_EMEA\\_course\\_ra\\_FTCOF\\_career-academy\\_pmax-multiple-audiences-country-multi&campaignid=20858198824&adgroupid=&device=c&keyword=&matchtype=&network=x&devicemodel=&adposition=&creativeid=&hide\\_mobile\\_promo&gad\\_source=1&gclid=Cj0KCQjw2PSvBhDjARIsAKc2cgPIbp\\_NKeFGgnRrB70A0RDPBxnnIk1bbqngbC8AtedAuT2IMguyORkaAsQUEALw\\_wcB](https://www.coursera.org/articles/what-is-machine-learning?utm_medium=sem&utm_source=gg&utm_campaign=B2C_EMEA_course_ra_FTCOF_career-academy_pmax-multiple-audiences-country-multi&campaignid=20858198824&adgroupid=&device=c&keyword=&matchtype=&network=x&devicemodel=&adposition=&creativeid=&hide_mobile_promo&gad_source=1&gclid=Cj0KCQjw2PSvBhDjARIsAKc2cgPIbp_NKeFGgnRrB70A0RDPBxnnIk1bbqngbC8AtedAuT2IMguyORkaAsQUEALw_wcB)

1.1. *Linear models.* (n.d.). Scikit-Learn. Retrieved March 22, 2024, from [https://scikit-learn.org/stable/modules/linear\\_model.html](https://scikit-learn.org/stable/modules/linear_model.html)

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