

Prediktiv analys

FÖRELÄSNING 1

Dagens fråga

- Hur taggade är ni på lära er hur man förstår data och dra insikter från??? Alltså lära er mer om machine learnig och prediktiv analys?

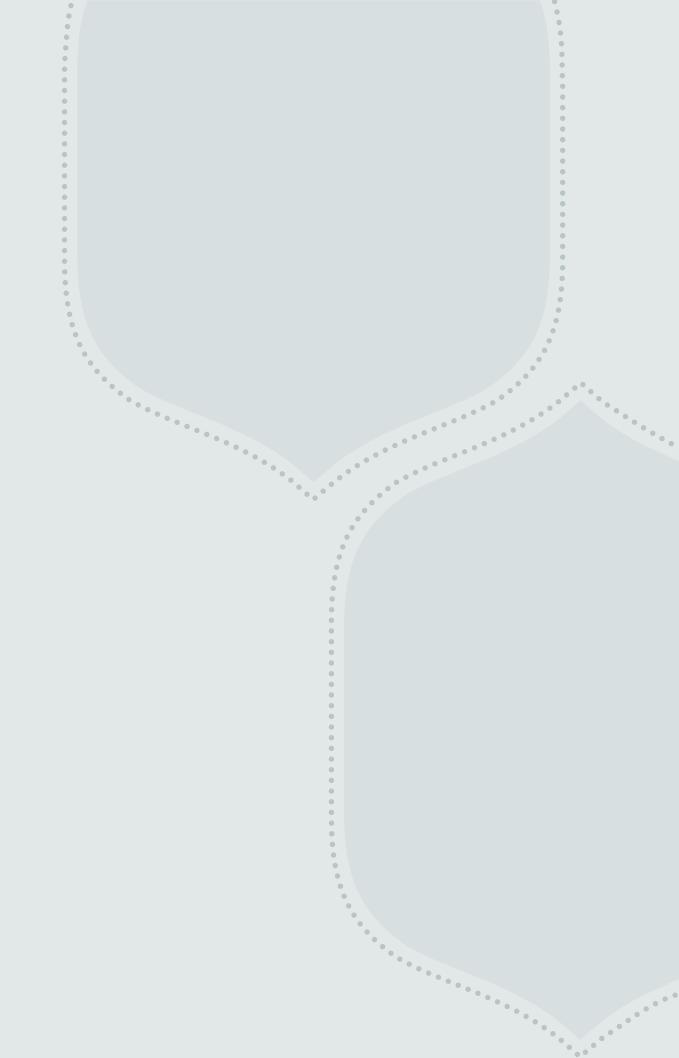
Dagens agenda

- Intro till kursen och kursplanering
- Etablering (repetition?) av koncept: prediktiv analys, algoritm, statistik, machine learning, neuralt närvverk, data mining, business intelligence, regression, klassificering, datarensning
- Repetition installera Conda, VSC, Jupyter Notebook, Virtual Environment, GitHub Desktop och kursens repository
- Inför nästa lektion: läsa och på om grundläggande koncept vi kommer prata om i kursen

Vad är prediktiv analys och varför behöver vi det som Data scientist?

ANVÄND HISTORISK DATA FÖR ATT FÖRSÖKA
FÖRUTSE VAD SOM KOMMER HÄNDA I FRAMTIDEN. PÅ
DETTA SÄTTET KAN FATTA SMARTARE BESLUT

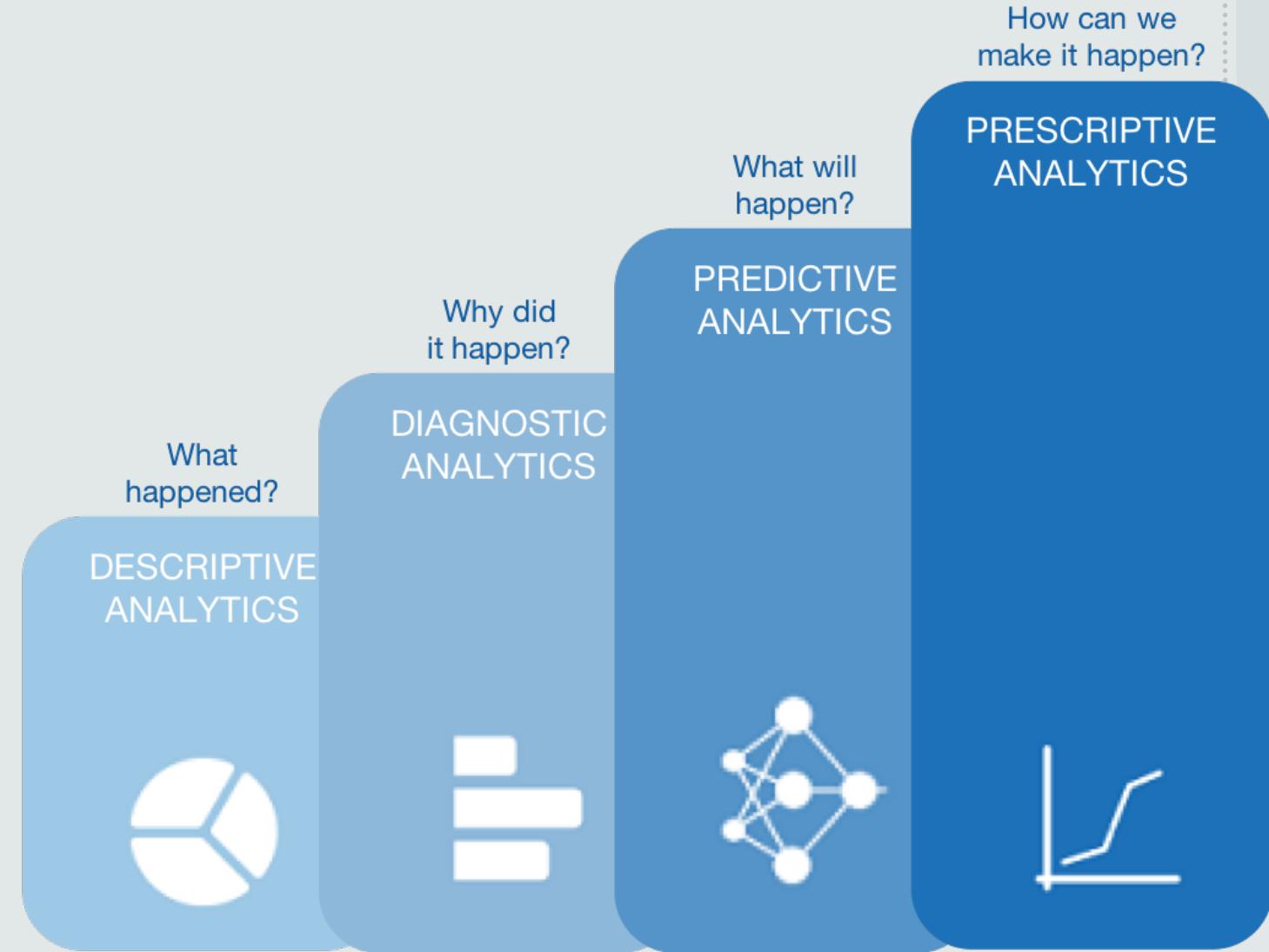
Studieplanering





Terminologi

ETABLERA VIKTIGA KONCEPT

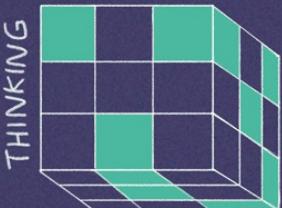
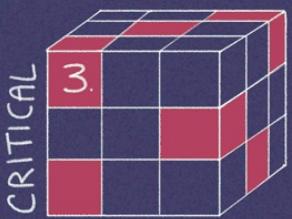


Vad innebär det att ha en bra analytisk förmåga och varför behöver man det som Data scientist?

Analytiska färdigheter är förmågan att kunna förstå olika aspekter av ett problem
och hitta idealalösningar på ett effektivt sätt

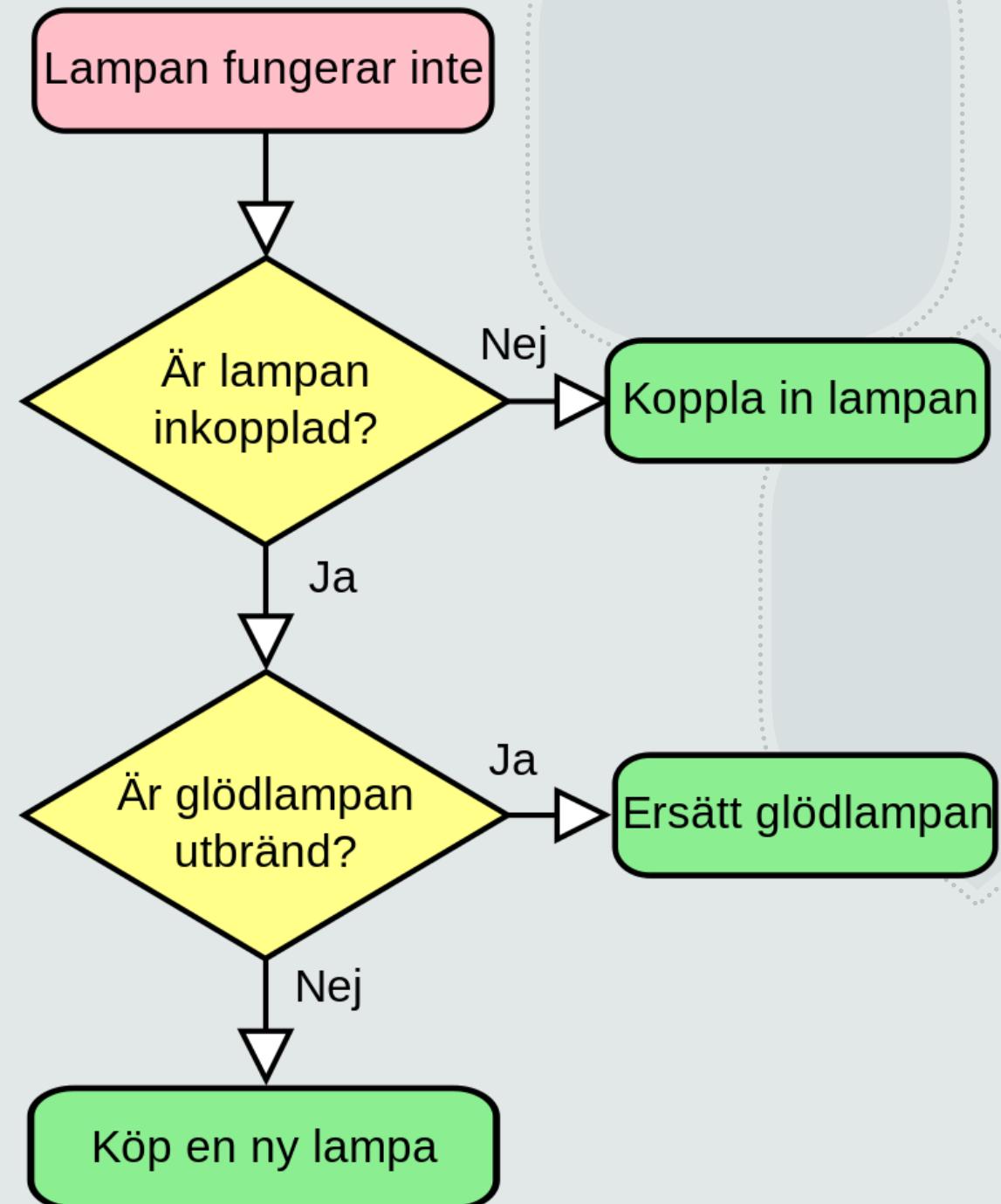


The 5 Types of Analytical Skills



Algoritm

- En algoritm är en funktion eller modell som ett datorprogram kan följa och köra
- Algoritm används inom matematik och datorvetenskap
- Algoritm är en lista regler som löser ett problem
- Algoritmen behöver ha stegen i rätt ordning för att lösa ett problem
- Tänk på algoritmen som ett recept!

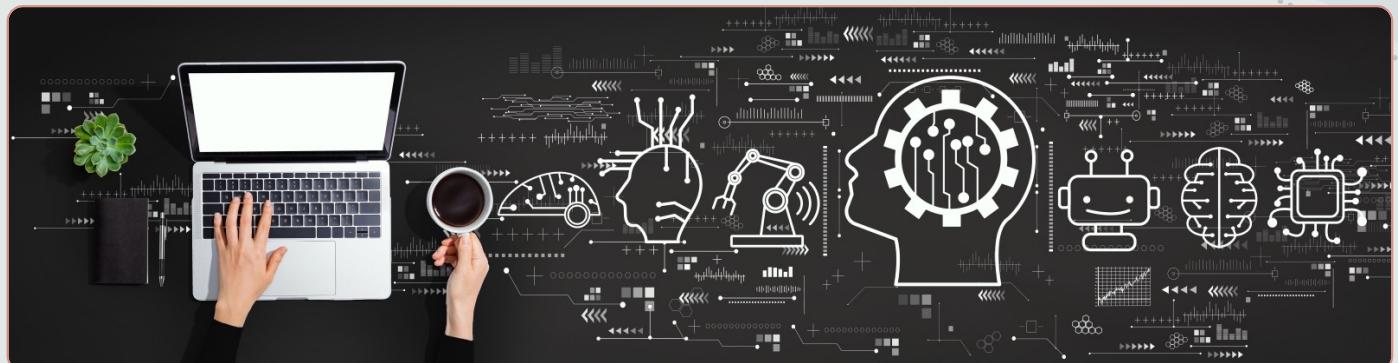


Statistik

- Statistik är en gren inom tillämpad matematik som sysslar med insamling, utvärdering, analys och presentation av data eller information
- Uppgifter som beskriver en sak eller en händelse med siffror
Exempel hur många barn föddes i Göteborg 2019
- Samla in uppgifter och arbeta med dem
- Resultatet används för att visa hur något är för tillfället eller för att förutsäga framtidiga händelser
- Resultatet presenteras ofta i procentsatser, tabeller eller diagram

Machine Learning

- Metoder för att träna datorer med hjälp av data till att upptäcka och lära sig regler för att lösa en uppgift.
- Detta UTAN att datorn har blivit programmerat med regler för den uppgiften!
- Dator blev skapat för att vi villa vara lata - de skulle lösa problem åt oss. Första exempel är kalkylatorn
- Nu programmerar vi datorn och kan skapa ännu mer avancerade program än kalkylatorn - så vi kan vara ännu latare
- När vi programmerar specificerar vi betingelser om exakt vad datorn ska göra
- Vad om datorn kan göra saker själv? Lära datorn att identifiera objekt och ändra vad den gör av sig själv när förutsättningarna ändras?
- Om vi visar datorn massor med olika data för att den ska lära sig identifiera får vi en automatiserad metod.



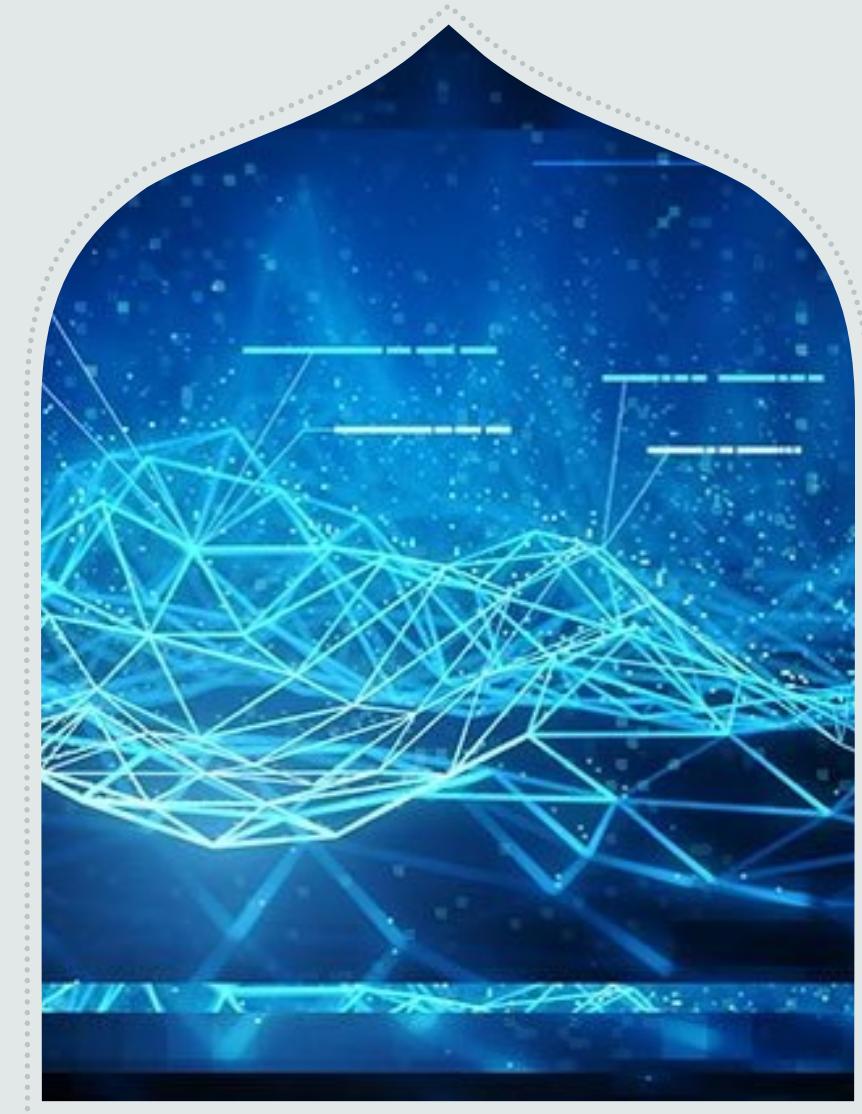
Neuralt nätverk

- Ett neuralt nätverk är en machine learning metod
- Ett neuralt nätverk är löst baserad på strukturen av hjärnan som består av neuroner och kopplingar mellan de
- Ett neuralt nätverk kallas också **deep learning**
- Metoden är bra på att lära komplexa mönster i stora dataset och är populära för att de kan lära komplicerade icke-linjära funktioner



Data mining

- Data mining kallas också prediktiv analys eller machine learning
- Hur kan en organisation gå från att vara informationsrik till att bli insiktsfull och hur tar man bäst beslut utifrån denna insikt?
- Det finns ofta stora mängder data inom ett företag men de vet ofta inte hur denna informationen ska användas
- Data mining är de verktyg man kan använda för att ta reda på information från data genom att upptäcka mönster och relationer i stora mängder data
- På detta sättet kan Data mining användas till problemlösning och optimering
- Data mining är uppbyggd av matematiska och statistiska metoder som beslutsträd, regressionsanalys, algoritmer och neurala nätverk



Business Intelligence (BI)

KNOWLEDGE MANAGEMENT

BENCHMARKING

- Ge rätt person rätt information vid rätt tillfälle
- Business intelligence (BI) är ett samlingsbegrepp för de metoder, programvara, färdigheter och tekniker för organisationer att bättre förstå sin verksamhet.
- BI är **beslutsstöd**:
 - 1.Hämta in data
 - 2.Omvandla data till information
 - 3.Få ut information till organisationen som beslutsstödet avser att stödja
 - 4.Generera ny kunskap hos informationsmottagaren
- BI är metoden/processen för att analysera data och presentera informationen från datan så att styrande inom ett företag kan ta smartare beslut
- Målet med BI är ofta att omvandla passiv information sparad inom ett företag till aktiv data som dagligen kan användas som underlag i beslut
- Visualiseringsverktyg som automatisk samlar data i diagram som kan analyseras
- BI är mer mot företag för att öka omsättning

DATA MINING

DATA VISUALIZATION

MEASUREMENT ANALYSIS

REPORTING

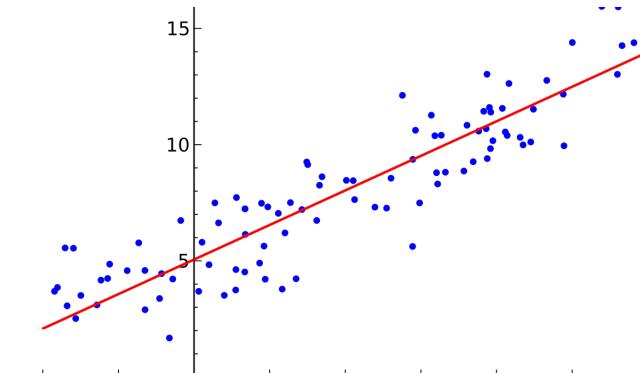
COLLABORATION PLATFORM

Regression

- Används inom statistik där målet är att skapa en funktion som bäst passar observerad data
- Enkel linjär regression ska man anpassa en rät linje data och regressionsekvationen blir då

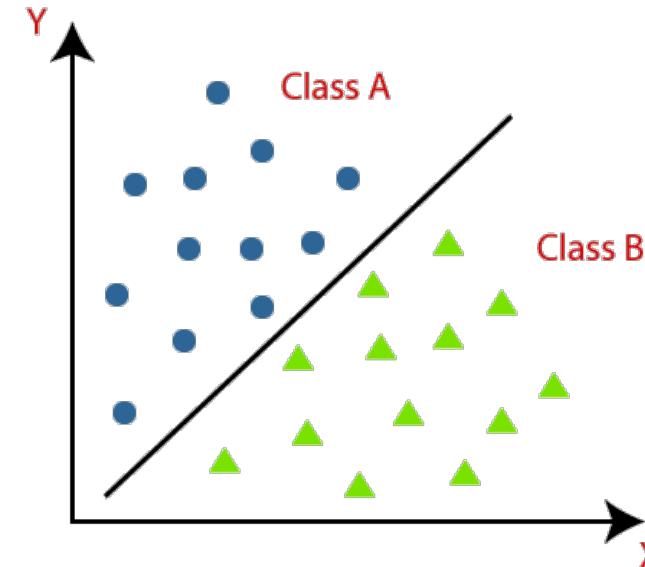
$$y = a + bx$$

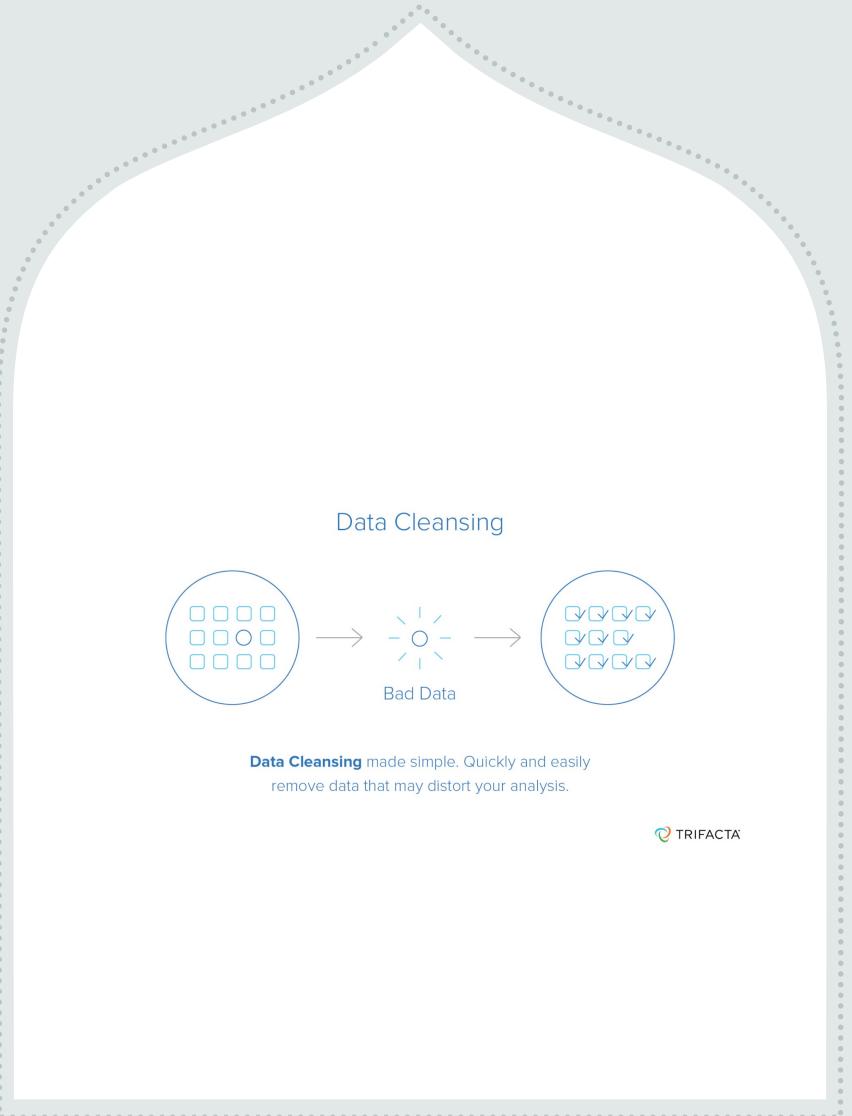
- Hittar **sambandet** mellan beroende variabel y och andra oberoende variabler x



Klassificering

- Klassificering är processen att prediktera klassen till ett givet datapunkt
- I motsättning till regression, där vi önskar prediktera ett kontinuerligt tal, är det i klassificering ett diskret värde vi ska prediktera
- Till exempel om en mail är “spam” or “not spam” eller om bilden är av en ”hund” eller en ”katt”





Datarensning

- Datarensning är processen där man upptäcker och korrigrar felaktig data
- Felaktig data kan också tas bort
- Identifiera inkomplett, okorrekt, felaktig, irrelevant delar av datan.
- Den felaktiga datan kan ersättas, ändras eller raderas
- Efter rensning borde data vara enhetlig med samma typ data
- Felen i datan kan komma från inmatningsfel, korruption i överföring eller lagring

Installera Conda

- För att få Python installerar man Miniconda
<https://conda.io/projects/conda/en/latest/user-guide/install/index.html#regular-installation>
- Aktivera så att Powershell kan använda Conda genom att öppna Anaconda Prompt:
`conda init powershell`
`conda activate`
- Starta om Powershell
- Kan hända Powershell ger felmeddelandet: "cannot be loaded because running script is disabled"
- Ändra execution policy: `Set-ExecutionPolicy -Scope CurrentUser -ExecutionPolicy Unrestricted`

Installera VSC

- Installera VSC <https://code.visualstudio.com/Download>
- Installera Python Extension i VSC. Följ instruktionerna:
<https://code.visualstudio.com/docs/python/python-tutorial>

Jupyter Notebook och Conda virtual environment

- För att registrera ett nytt environment för att använda med notebooks installerar man ett Jupyter Notebook i sitt base environment. Detta behöver du bara göra en gång! I powershell:

```
conda activate base
```

```
conda install -c conda-forge notebook nb_conda_kernels
```

- Skapa ett nytt conda environment och installera **python** och **ipykernel**. Välj själv namn istället för **NEW_ENV**

```
conda create --name NEW_ENV python=3.9 ipykernel
```

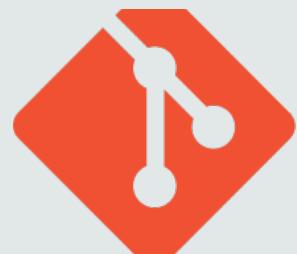
Conda virtual environment

- Aktivera ditt environment:

```
conda activate NEW_ENV
```

- Installera alla nödvändiga python paket. Se till att du står i samma mapp som `requirements.txt`. Använd `cd` för att ändra mapp du står i och `ls` för att se innehållet i mappen. Kör

```
pip install -r requirements.txt
```



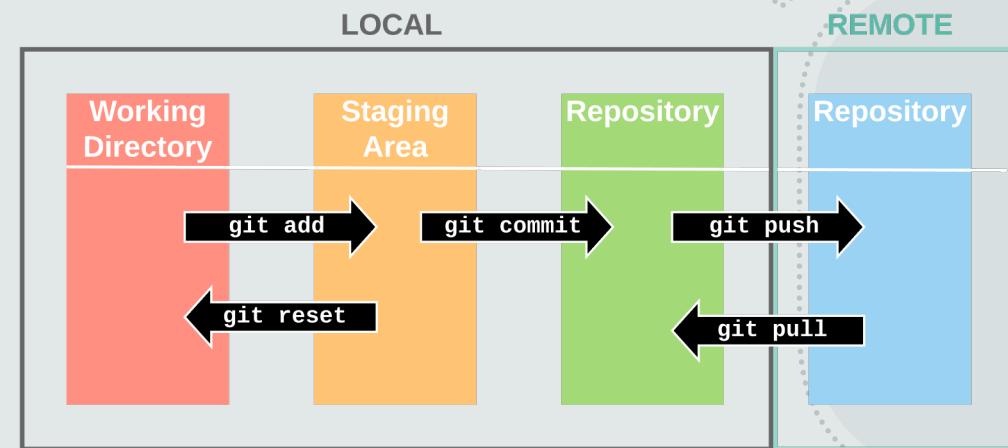
git

Versionshantering

- Spara data över hur filer/data har sett ut och förändrats över tid.
- Möjlighet att "spola tillbaka".
- "Loggbok" över *vem* som ändrat något, *när* det ändrades, *vad* som ändrades.
- Verktyg när man felsöker kod.

Git struktur

- Git är ofta strukturerat i två delar: **local** och **remote**
- **Local** = det som finns på din dator, det vill säga: ett lokalt repo, en staging area och din working directory (din kopia som du jobbar i).
- **Remote** = ett repos som finns på en delad server och fungerar som master/main för projektet. Ofta används t.ex. Github för att hantera master/main-repot.



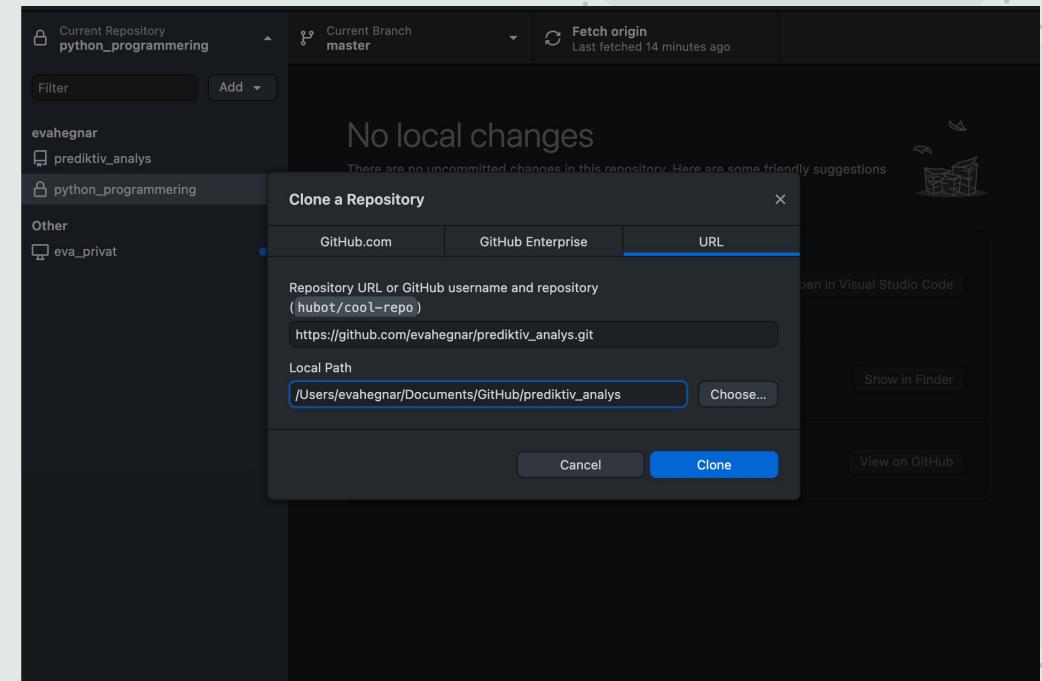
Github Desktop

- <https://desktop.github.com>
- Mac användare behöver flytta den nerladdade filen till “Applications foldern“ på datorn



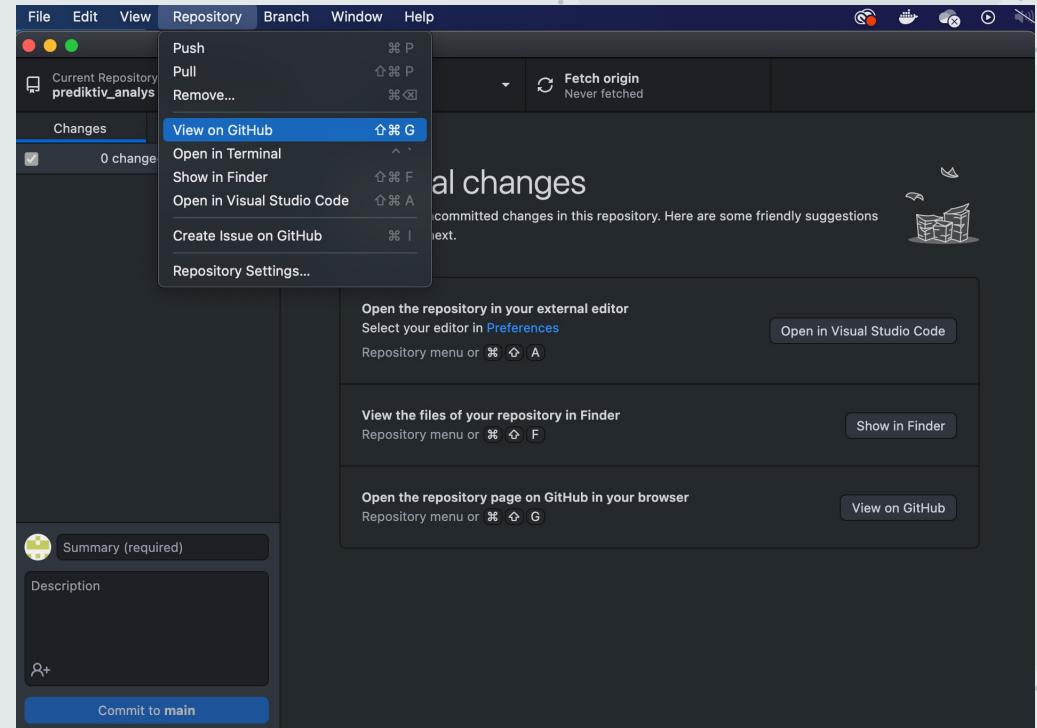
Clone repository

- Starta GitHub Desktop och logga in på ditt konto
- Välj Add.. Clone Repository.. URL och ange
https://github.com/evahegnar/prediktiv_analys.git
- Ändra Local Path om du vill bestämma var repot ska sparas lokalt på din dator



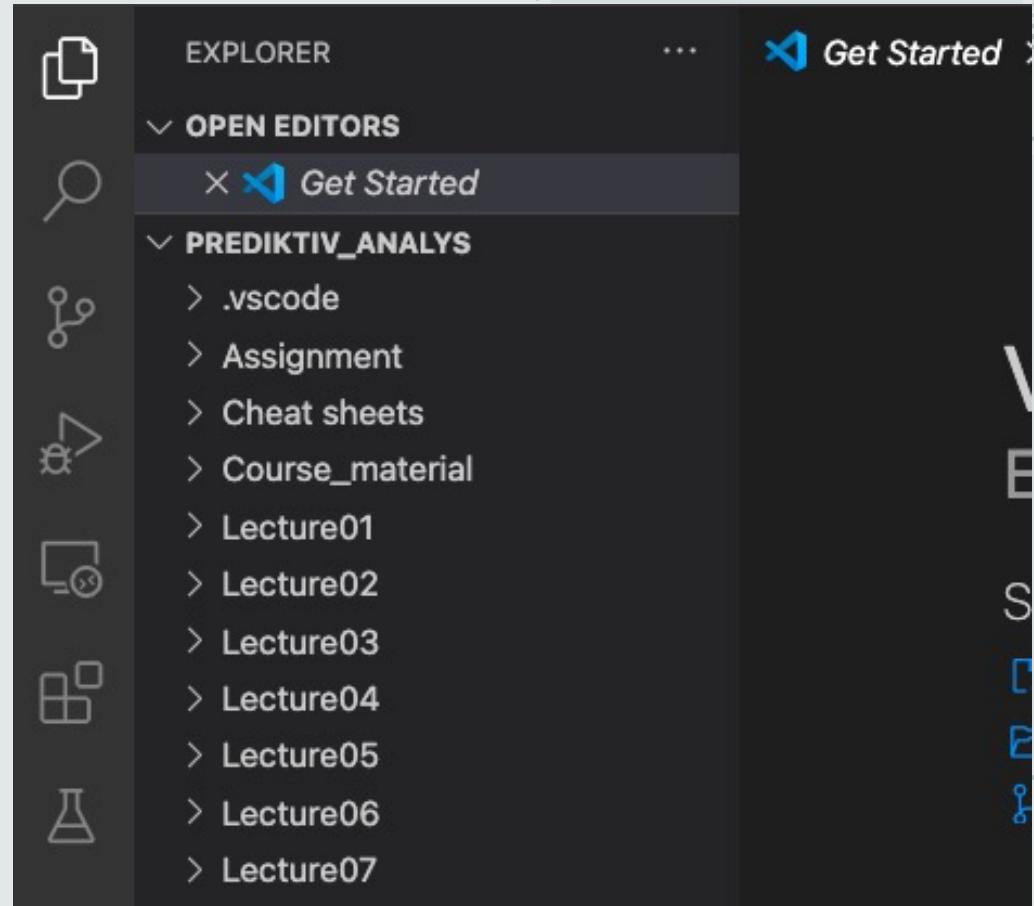
Utforska Repository

- Man kan öppna repot i VSC, Utforskaren (lokala mappen) eller på GitHub. Syns det inte nås kommandon genom Repository



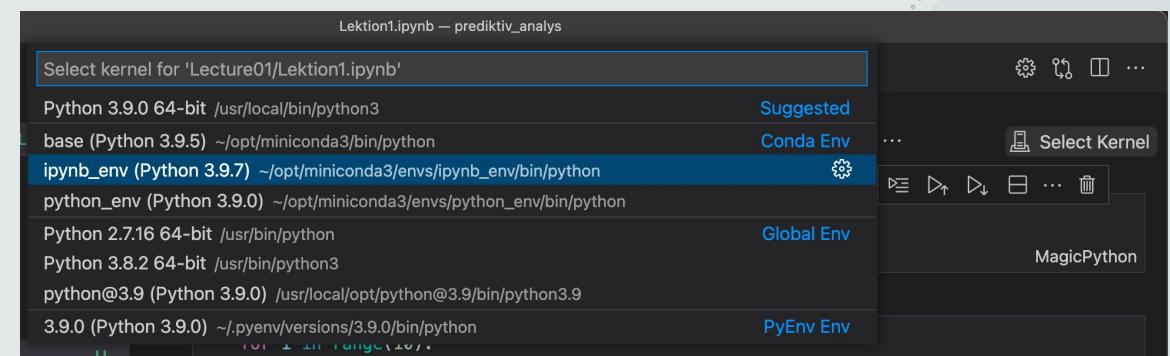
VSC

- När du har öppnat ett repo i VSC ska det se ut så här



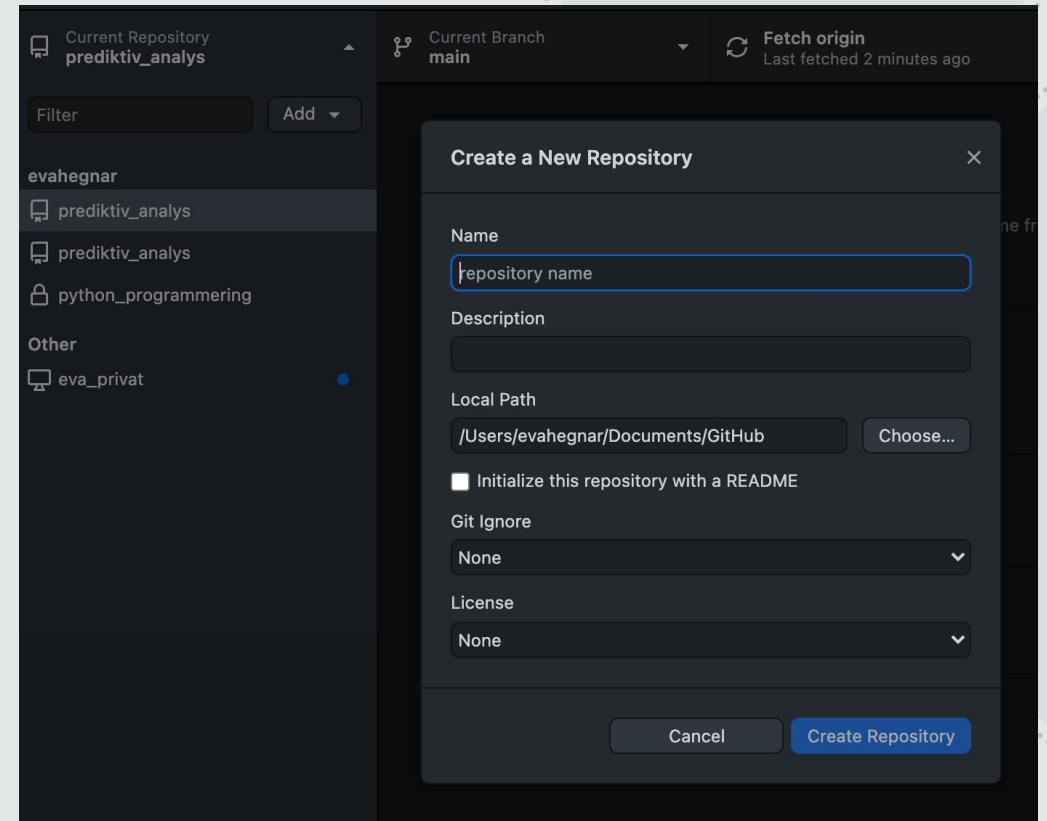
VSC och environments

- Aktivera virtual environment i VSC genom Select Kernel överst i höger hörn



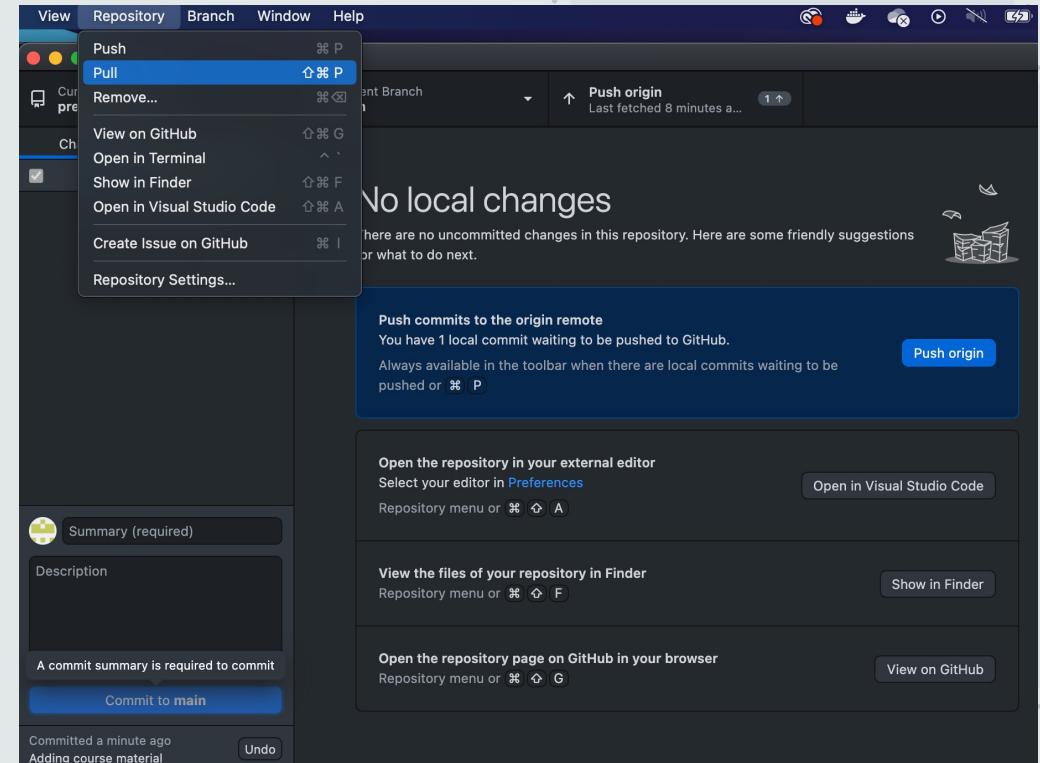
GitHub

- Du kan även skapa egna repos, öppna de i VSC och lägga till de på din personliga GitHub



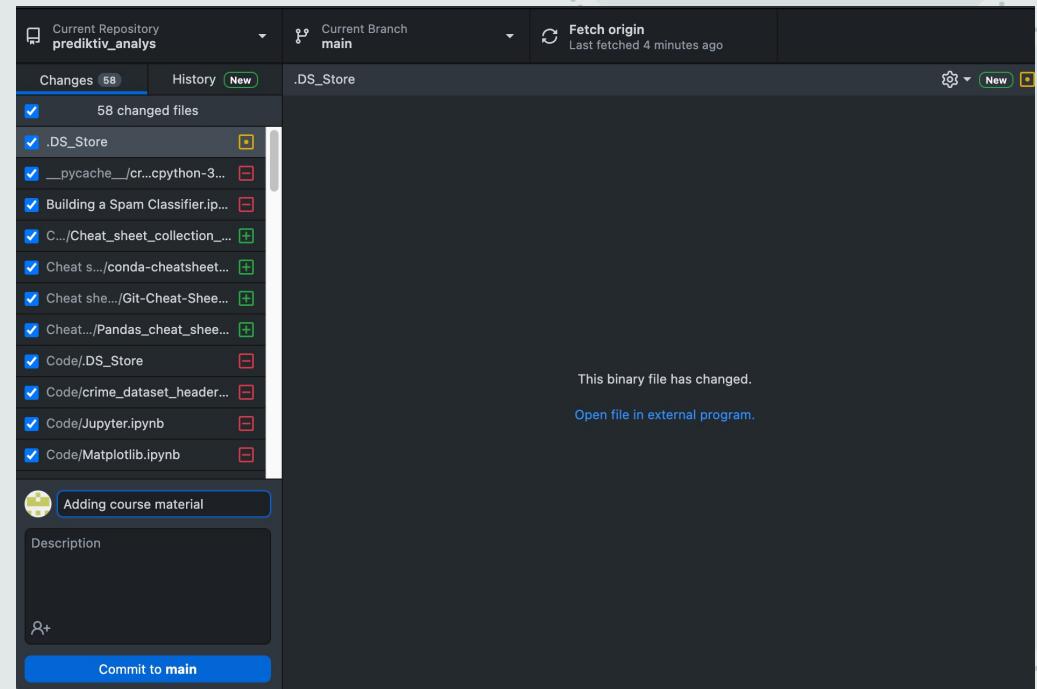
Pull

- När jag gör ändringar i repot behöver ni göra en **pull** för att få de senaste ändringarna



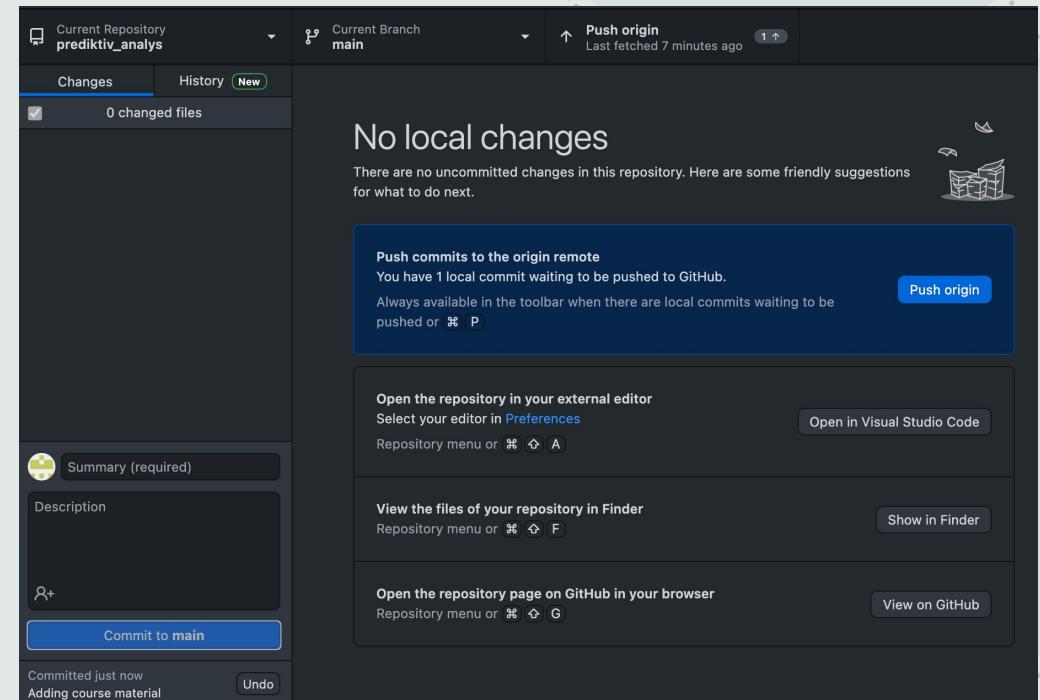
Commit

- Om ni väljer jobba i prediktiva_analys repot måste ni committa ändringarna ni har gjort före ni gör en pull
- Alla ändringar kan granskas här (detta är det samma som git add)
- Det *krävs* att ni skriver en sammanfattning
- Klicka sedan commit to main



Push

- Ni kan **inte** pusha till min repository, prediktiv_analys
- Ni kan pusha ert egna skpade repository



Jupyter notebook kommandon

Enter - redigera läge

Ecs - kommando läge

Shift+Enter - run cell

Gömma output - dubbeltklicka vänster sida av output eller använd ;

Copy-paste en linje kod - alt + shift + pil upp/ner

Kommentera bort/in kod - ctrl + shift + 7

När man är i kommando läge (Esc):

a - ny cell över

b - ny cell under

x - klipp ut vald cell

c - kopiera vald cell

v - klistra in cell

d+d - radera cell

z - ångra raderad cell

s - spara

m - ändra cell till Markdown

y - ändra cell till kod

o - göm output/vis output

Markdown .md

- Markdown kan användas ihop med Notebook för att få ”rapport”-känsla
- Cheat sheet <https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet>

The screenshot shows a Jupyter Notebook interface. At the top, there's a menu bar with File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu is a toolbar with various icons for file operations like Open, Save, and Print, along with Run, Cell, and Help buttons. The main area has two sections: 'In [6]:' containing Python code and 'Out [6]:' showing the execution results.

In [6]:

```
for i in range(10):
    print(i)
```

Out [6]:

```
0
1
2
3
4
5
6
7
8
9
```

Below the notebook interface, there's a separate section titled 'Markdown Language' with examples of headings, normal text, and lists.

Markdown Language

Heading 2

This is normal text

List

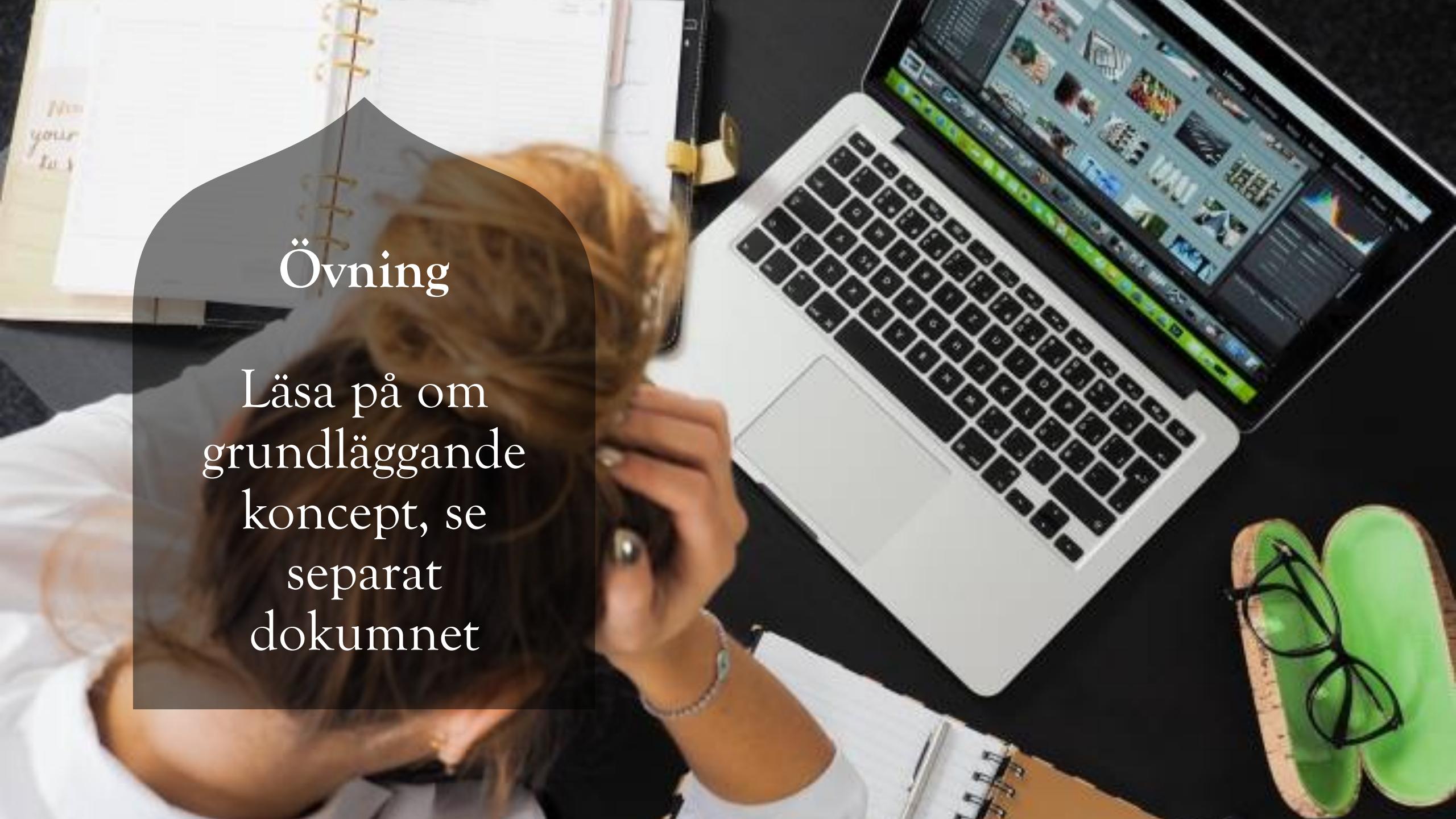
- Item 1
- Item 2

In []:

README.md

Övning

Läsa på om
grundläggande
koncept, se
separat
dokumnet



Vad har vi gjort idag?

- Intro till kursen och kursplanering
- Etablering (repetition?) av koncept: prediktiv analys, algoritm, statistik, machine learning, neuralt närvverk, data mining, business intelligence, regression, klassificering, datarensning
- Repetition installera Conda, VSC, Jupyter Notebook, Virtual Environment, GitHub Desktop och kursens repository



Nästa lektion

- Viktiga python bibliotek för prediktiv analys
- Vi går igenom grundläggande koncept i prediktiv analys
- Supervised VS unsupervised learning.
- Regression och klassificering.
- Modeller och algoritmer