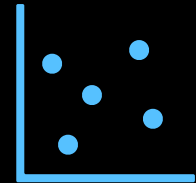




ISTx1003: Statistisk læring og data science

Forelesning 2: Introduksjon og multippel lineær regresjon

Forelesningen blir tatt opp



Sara Martino, Institutt for matematiske fag, NTNU, 24. oktober 2024 14:15-16:00



ISTx1003: Statistisk læring og data science

Forelesning 2: Introduksjon og multippel lineær
regresjon

Forelesningen blir tatt opp



Takk til

Mette Langaas (pensumvideoer, kompendium, slides og inspirasjon)
Stefanie Muff (slides og inspirasjon) Ingeborg Hem Sørmoen (slides og inspirasjon)

Mine slides bygger på materiale fra både Stefanie, Mette og Ingeborg, som underviste dette kurset i 2021-2023

Praktisk om denne zoom-forelesningen

- Zoom-webinar (som i fellesdelen)
- Forelesningen blir tatt opp og lagt ut
- Dere kan bruke chatten for å sende spørsmål til meg, jeg skal prøve å følge med så godt jeg kan.
- Jeg oppfordrer dere til å sitte sammen med andre studenter og se forelesningen!
- Alle info ligger i wiki sida: <https://wiki.math.ntnu.no/istx100y/2024h/1003>

Prosjektarbeid i grupper

Oktober og november

Undervisning

Oktober

Veiledning

~~Oktober~~
November

Innlevering

18. november
12:00

Plan

Se Wiki sida for
utfyllende info

Forelesninger (små endringer kan forekomme)

- Mandag 21. oktober 15:15-16:00: Introduksjon
 - Tirsdag 22. oktober 14:15-16:00: Oppgave 1 (multippel lineær regresjon) -
 - Mandag 28. oktober 15:15-16:00 Oppgave 2 (klassifikasjon) og oppgave 3 (klyngeanalyse)
 - ??? Tirsdag 29. oktober 14:15-16:00 Oppgave 3 fors. (?), Q&A og praktisk informasjon
-
- Alle de 4 forelesningene vil være hel-digitale ✓
 - Alle forelesningene tas opp, tentative slides legges ut før forelesningen og slides med notater legges ut etter forelesningen ✓
 - Forelesningene vil i all hovedsak utdype prosjektoppgavene. Pensum dekkes av kompendier og tilhørende videoer, og blir ikke repetert her. !

Plan

Pensumvideoer

Se Wiki Sida for
utfyllende info

For å få fullt utbytte av forelesningene bør dere ha sett følgende:

- Enkel lineær regresjon (uke 9) innen mandag 21. oktober (i dag)
- Multipel lineær regresjon innen tirsdag 22. oktober (i morgen)
- Klassifikasjon innen 28. oktober
- Klyngeanalyse innen 28. oktober

Plan

Veiledning

- Digitalt:

- Tirsdag 5/11 og 12/11, 14:15-16:00

Vi øker mengden digital veiledning ved behov.

- Torsdag 24/10, 31/11, 7/11, 14/11 14:15-16:00

- Fysisk:

- Trondheim: Torsdag 24/10, 31/11, 7/11, 14/11 14:15-16:00 i S5 (SBII)

- Gjøvik: Kun digitalt

- Ålesund: Digital. I tillegg gruppene kan møte Siebe til veiledning, avtal tid med han. Se blackboardsiden ISTA: Campus Ålesund for mer informasjon.

- Forum:

- Åpent 24/7, besvares jevnlig av fagteamet

**18. november 12:00:
Leverer prosjekt i Inspira**

Plan

I dag, 21. oktober

1. Diverse informasjon ✓
2. Hvem er fagteamet? ✓
3. Statistisk læring og data science
4. Representative data
5. Prosjektmodulens temaer: Regresjon, klassifikasjon og klyngeanalyse
6. Læringsmål –
7. Pensum og læringsressurser –
8. Wiki sida –
9. Prosjektet

Hvem er vi?

Sara



Faglig ansvarlig for
innholdet i modulen,
Trondheim

Øyvind



Trondheim

Pedro
Trondheim

Tarjei
Trondheim



Hvem er dere?

Gå til menti.com

Nummer!

Data science

Forene statistikk, dataanalyse, informatikk og tilhørende metoder for å forstå og analysere reelle fenomener

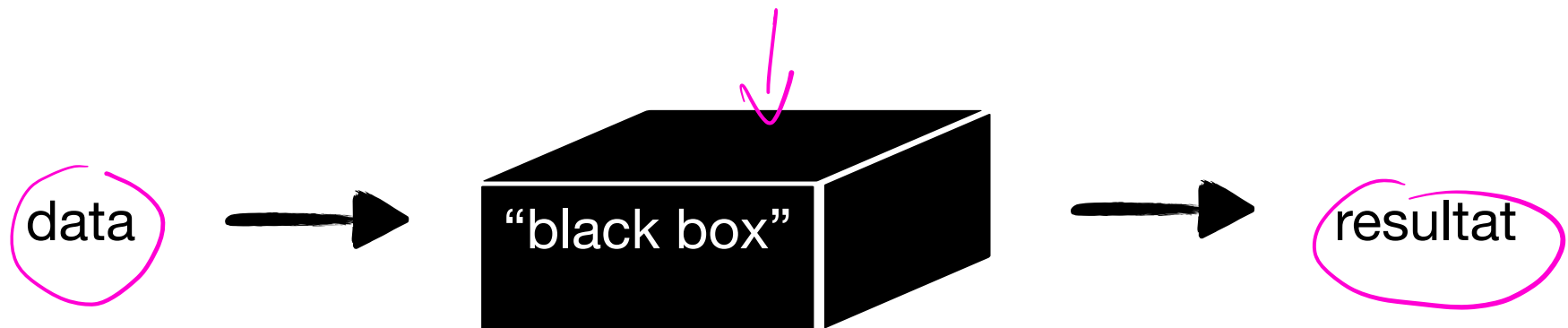


Hentet fra <https://www.sudeep.co/data-science/2018/02/09/Understanding-the-Data-Science-Lifecycle.html>

1. Forstå problemet
2. Samle inn data
3. Rydd opp i data 4
4. Utforske data
5. Restrukturere data
6. Tilpass modeller og bruk dem til prediksjon
7. Vis fram dataene, modellene og resultatene

Statistisk læring

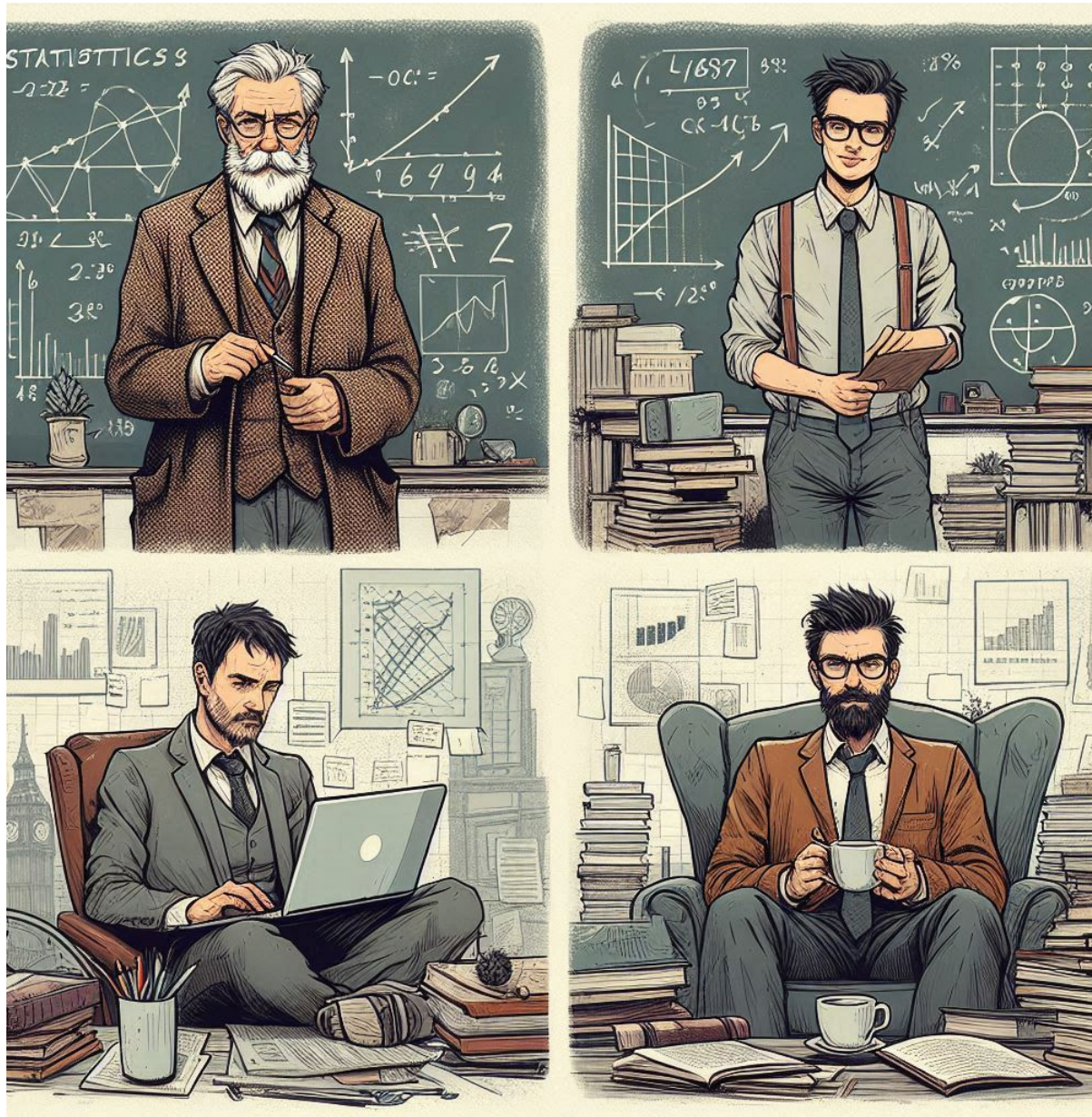
Metoder som hjælper oss å lære av data



Representative data

Professors in statistics

Bins



Representative data

Italians



Representative data

Bilde-generatoren favoriserer mannlige forelesere, fordi i mesteparten av bildene den er trent på er foreleserne menn.

Algoritmen som lager bilder er jo utrolig imponerende selv om den fremstår litt diskriminerende.)

Denne diskrimineringen er kun en gjengivelse av hva algoritmen har blitt trent på, selve algoritmen er ikke diskriminerende.

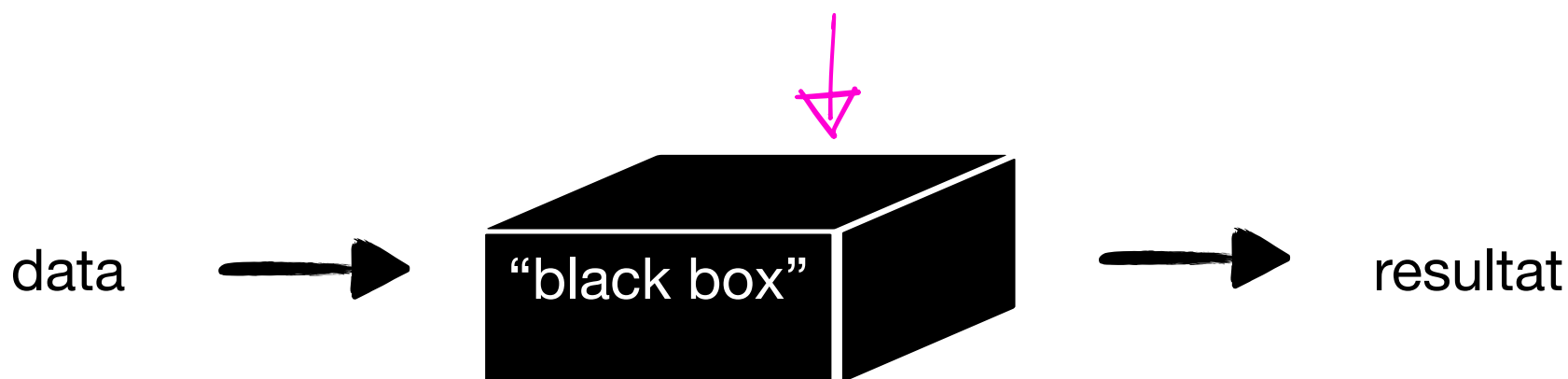
Representative data

Vi bør alltid være litt skeptiske til resultater fra algoritmer vi ikke vet mye om.

Vi må også huske på at algoritmene er laget for å gjøre spesifikke ting, som kan være noe annet enn det vi bruker dem til.

Statistisk læring

Metoder som hjelper oss å lære av data



Enkel lineær regresjon

Multippel lineær regresjon

Klassifikasjon

Klyngeanalyse

Python

Regresjon

Antagelser

$$Y \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

• konstant varians

$$\epsilon_i \sim N(0, \sigma^2)$$

Mål: Forstå sammenheng mellom respons og forklaringsvariabler, og predikere en respons for nye forklaringsvariabler.

Respons variabel

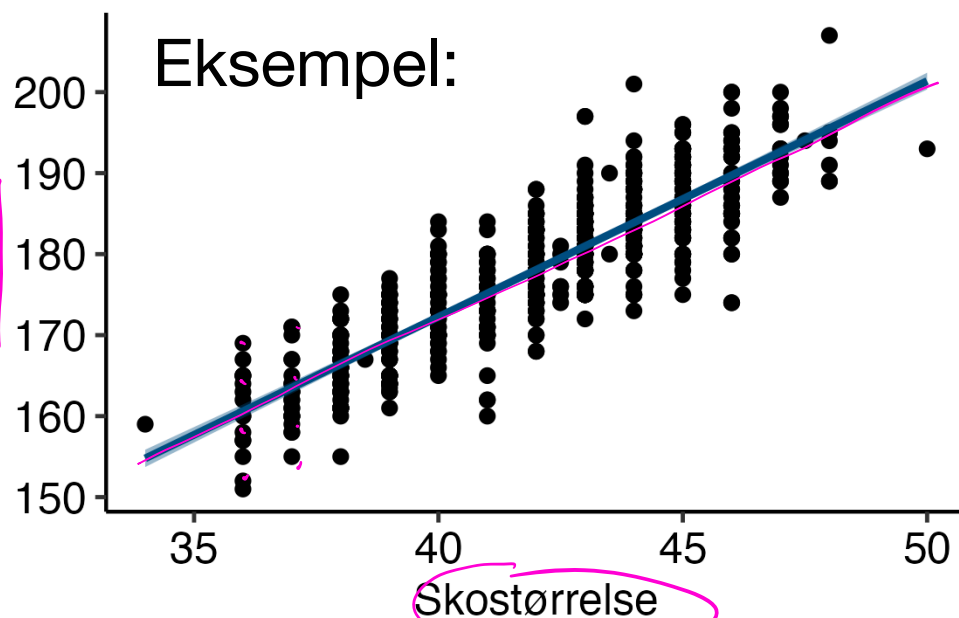
$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

forklars variabel

Enkel lineær regresjon

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} + e_i$$

Multippel lineær regresjon



I prosjektet skal vi analysere priser på LEGO.

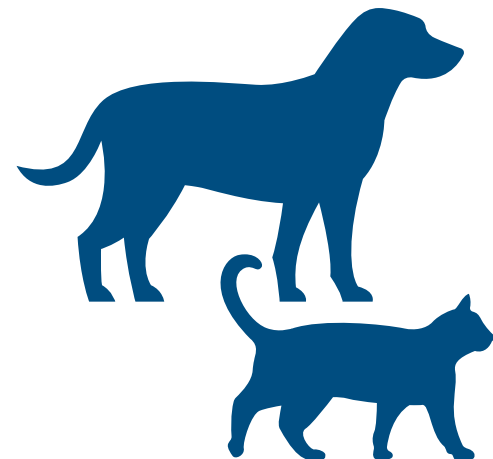


Klassifikasjon

Y_i ategorisk

Mål: Tilordne nye observasjoner til én eller flere kjente klasser. Dette er 'supervised learning' (veiledet læring).

For eksempel: Lære en datamaskin å se forskjell på bilder av hunder og katter.



I prosjektet skal vi se på fotballresultater fra den engelske herreligaen fra 22-23-sesongen.

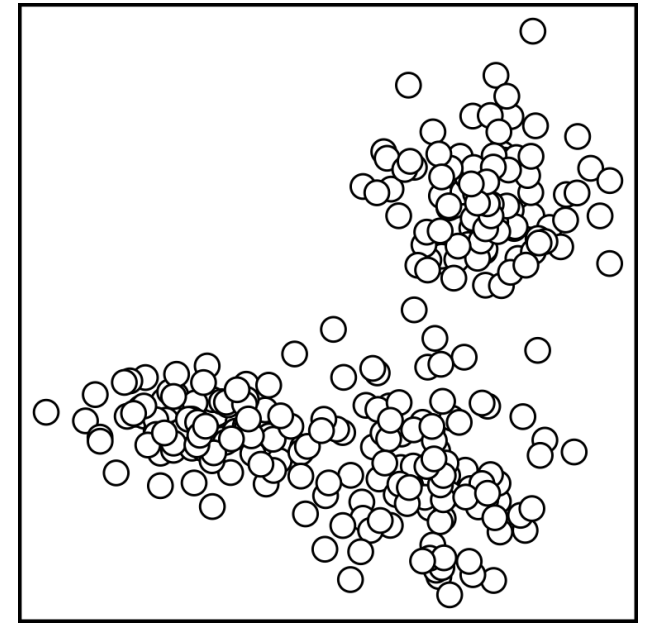


Klyngeanalyse

Ingen respons

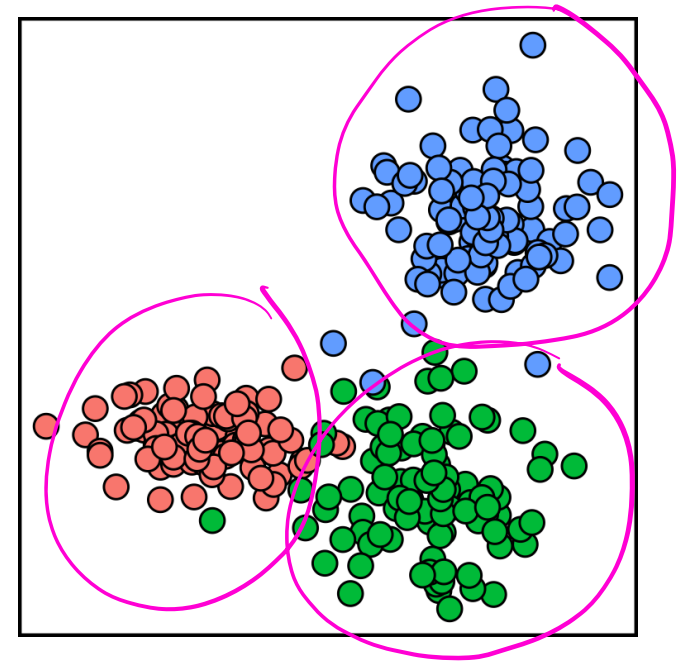
– Unsupervised learning

Mål: Finne struktur i dataene når vi ikke har noen respons eller fasit. Dette er 'unsupervised learning' (u-veiledet læring).



For eksempel: Gruppere punktene.

I prosjektet skal vi lære en algoritme å se forskjell på ulike håndskrevne tall.



Læringsmål

Etter du har gjennomført denne modulen skal du kunne:

- forstå når du kan bruke regresjon, klassifikasjon og klyngeanalyse til å løse et ingeniørproblem
- kunne gjennomføre multippel lineær regresjon på et datasett
- bruke logistisk regresjon og nærmeste nabo for utføre en klassifikasjon
- bruke hierarkisk og k-means klyngeanalyse på et datasett, forstå begrepet avstandsmål klyngeanalyse
- og kunne kommunisere resultatene fra regresjon/klassifikasjon/klyngeanalyse til medstudenter og ingeniører
- bli en kritisk leser av resultater fra statistikk/maskinlæring/statistisk læring/data science/kunstig intelligens når disse rapporteres i media, og forstå om resultatene er realistiske ut fra informasjonen som gis
- kunne besvare prosjektoppgaven på en god måte!

Pensum og læringsressurser

Pensum

Pensum er definert som “**svarene på det du blir spurt om på prosjektoppgaven**”, og de kan du finne ved å bruke læringsressursene.

Det blir ikke spørsmål fra prosjektmodulen på den skriftlige eksamenen, **men det du lærer i modulen** vil gi deg økt forståelse av temaene i fellesmodulen, og spesielt inferensdelen (estimering, hypotesetesting, enkel lineær regresjon).

Pensum og læringsressurser

Læringsressurser

- Kompendier (laget av Mette Langaas)
- Tilhørende videoer (laget av Mette Langaas)
- Zoom-forelesningene (med slides)
- Notat om interaksjoner (laget av Ingeborg Hem Sørmoen)

Kompendier og tilhørende videoer ligger allerede på [Wiki Sida](#).

Zoom-forelesningene og slides legges også ut.

Mange andre tilgjengelige kilder på nett også, men jeg garanterer ikke at de dekker de riktige delene av temaene.

Wiki Sida

Vi ser hvor informasjonen ligger på [Wiki sida](#).

Prosjektet

Dere skal bruke mye Python, og vi har laget notatbøker til dere.
Jeg anbefaler alle å gjøre prosjektet lokalt på egen maskin.
Hvordan? Se Blackboard.

The screenshot shows a Blackboard course interface. On the left, a sidebar lists course content, with 'Python/Jupyter' circled in pink. The main content area displays two documents: 'Python og jupyterhub' and 'Hvordan kjøre Jupyter-notatbøker på egen PC?'. The first document provides instructions on using Python and Jupyter notebooks, including a link to <https://s.ntnu.no/isthub>. The second document explains how to run Jupyter notebooks locally, with a list of links and a terminal command: `pip install -r requirements.txt`.

Sammenslått - Statistikk ISTX1001 ISTX1002 ISTX1003 VB6044 VB6200 (2024 HØST) Python/Jupyter

Python/Jupyter

Build Content Assessments Tools Partner Content

Python og jupyterhub

I dette emnet får du *erfaring med* å bruke python til å behandle datasett og regne på sannsynligheter.

Vi har kodet det du trenger i jupyter notatbøker som du kan redigere selv for å finne svar på øvingsoppgaver.

Notatbøkene finner du her (feide-innlogging):

<https://s.ntnu.no/isthub>

Dersom huben ikke fungerer kan alle notatbøker lastes ned herfra: <https://github.com/imfdrift/istx100y2024>

Hvordan kjøre Jupyter-notatbøker på egen PC?

[Her er en video](#) som går igjennom hvordan få tilgang til Jupyter-notatbøker uten å bruke Jupyter-huben. Dette kan for eksempel være oppgavene uten nett.

Lenker og kode fra video:

- [Lenke for installasjon.](#)
- [Lenke for nedlasting av notatbøker.](#)
- Fil for oppsett av pakker: [requirements.txt](#)

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

Prosjektet

Oppgave 1: Regresjon (50%, prosjektrapport)

Oppgave 2: Klassifikasjon (30%, 4 deloppgaver, 15 spørsmål)

Oppgave 3: Klyngeanalyse (20%, 4 deloppgaver, 10 spørsmål)

→ svare på spørsmål

→ Rapport

Hvordan evalueres prosjektet?

Prosjektrapporten i oppgave 1 får en poengscore fra 0-50.

Oppgave 2 og 3 har hhv. 15 og 10 spørsmål, som altså teller 2% hver.

Dette summeres, og prosentvurderingsmetoden brukes for å bestemme karakter.

For eksempel gir 40 poeng i oppgave 1, 13/15 i oppgave 2 og 9/10 i oppgave 3:

$$40 + 13 \cdot 2 + 9 \cdot 2 = 40 + 26 + 18 = 84, \text{ altså en } \underline{\underline{B}}.$$

Karakterskala for prosentvurderingsmetoden

A: 89–100 poeng

B: 77–88 poeng

C: 65–76 poeng

D: 53–64 poeng

E: 41–52 poeng

F: 0–40 poeng

Husk: Prosjektet teller 30% av endelig karakter, eksamen teller 70%.

	E eksamen	D eksamen	C eksamen	<u>B eksamen</u>	A eksamen
<u>E prosjekt</u>	E	D	D	<u>C</u>	B
D prosjekt	E	D	C	<u>C</u>	B
C prosjekt	D	D	C	<u>B</u>	B
B prosjekt	D	C	C	<u>B</u>	A
A prosjekt	D	C	B	<u>B</u>	A

Generell informasjon

I prosjektdelen av ISTx1003 Statistikk, Statistisk læring og data science, har vi fokus på tre hovedtemaer: *regresjon, klassifikasjon og klyngeanalyse*.

- Dette er oppgaveteksten til den tellende prosjektoppgaven, der besvarelsen teller 30% av karakteren i emnet.
- Dere er 3-6 studenter i hver gruppe. Gruppene dere skal levere prosjektet i er de samme gruppene som ble etablert i forprosjektet. *wiki!*
- Informasjon om prosjektmodulen finnes i *Blackboard*, sammen med alt kursmateriellet.
- Oppgave 1 skal skrives som en rapport på maksimalt 5 sider.
- Oppgave 2 og 3 skal utføres i notatbøker som er lastet opp på Jupyterhyben. De kan også finnes på Blackboard under fanen for 'ISTx1003'. Svarene skrives i svar-arket.
- De ulike oppgavene (1, 2, 3) er vektet ulikt (50%, 30% og 20%). Oppgave 1 rettes som en helhetlig besvarelse. Hvert spørsmål i oppgave 2 og 3 teller likt. Karakteren settes med prosentvurderingsmetoden hvor poeng blir konvertert i en prosentandel (ikke-heltall prosent blir avrundet).
- Prosjektet leveres i Inspira.
- Alle studenter i samme gruppe får samme bokstavkarakter. Merk NTNUs regelverk for klage på karakter: Ved klage på karakterfastsettingen av gruppearbeid, der det gis en felles karakter, klager du individuelt. En eventuell endring etter klage vil kun gjelde for den/de gruppemedlemmene som har klaget.

Dere skal levere **to pdf-filer**. Én for oppgave 1 og én for oppgave 2 og 3.

Begge filene **skal** være pdf-filer, og dere er selv ansvarlige for at alle svar er med i de to filene. Filene skal være ryddige og oversiktelige.

Frist for innlevering er 18. november klokka 12:00 i Inspira.

Kontakt eksamenskontoret ved problemer. Det kan ikke gis utsettelse på innleveringsfristen.

Hva skjer nå? (mandag)

Ny forelesning i morgen klokka 14:15 (digitalt, samme link). Da forsetter vi å snakke om oppgave 1.

For størst utbytte av forelesningen, se pensumvideoene om MLR før i morgen 14:15!

Start gjerne å diskutere problemstilling til oppgave 1.

Vi sees i morgen!