Tellende prosjekt i ISTx1003 høst 2024

Statistisk læring og data-science

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.
- 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 Inspera.
- 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 Inspera.

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

Oppgave 1 - Multippel lineær regresjon

Superkort oppsummering

- Rapporten skal være på **maksimalt 5 sider** (referanser og kode kommer i tillegg).
- Det legges vekt på at rapporten skal være sammenhengende.
- `'oppg_1.ipynb'` inneholder kode for å renske dataene.
- Problemstilling bestemmes **først**.
- Det anbefales å sjekke med faglærer (Sara) at problemstillingen er god før dere starter.

Oppgavebeskrivelse

I denne oppgaven skal vi analysere datasettet brukt i Peterson and Ziegler (2021) ved hjelp av multippel lineær regresjon (artikkelen finnes her). Datasettet består av 1304 observasjoner fra 1. januar 2018 til 11. september 2020. Hver observasjon beskriver et LEGO-sett, og består av følgende variabler:

`Set_Name` Navn på settet
`Theme` Tema settet hører til
`Pieces` Antall brikker i settet
`Price` Pris (i $)
`Pages` Antall sider i instruksjonsmanualen
`Unique_Pieces` Antall unike brikker

i tillegg til 8 andre variabler som vi ikke skal fokusere på i denne oppgaven.

Disse dataene henter du ved å laste ned "Supplementary materials" fra artikkelen og åpne mappen "Data.zip". Så laster du datasettet opp til Jupyterhuben.

Datasett hentet på nettet må gjerne reformateres og renskes litt opp i. I Jupyter notatboka som heter `'oppg_1.ipynb'` har vi laget kode som gjør den jobben. Der ligger det også noen relevante kodesnutter dere kan dra nytte av når dere skriver rapporten. Merk at koden ikke kjører før dere har lagt til datasettet.

I denne oppgaven skal dere selv definere en problemstilling (innenfor visse rammer, se punkt 1 under) som dere skal prøve å besvare ved å bruke LEGO-datasettet og multippel lineær regresjon. Oppgavens pipeline er som følger:

1. Definere problemstilling. Problemstillingen skal formuleres på en måte slik at dere må lage en ny kategorisk variabel som grupperer de ulike temaene. Eksempler på problemstillinger er:
   (a) Er LEGO for gutter dyrere enn LEGO for jenter?
   (b) Betaler vi ekstra for LEGO som assosieres med et varemerke?
   (c) Varierer prisen på LEGO i ulike aldersgrupper?
   (d) Har/hadde medlemmene på vår gruppe dyr LEGO (som barn)?

   Merk at alle disse fire problemstillingene har pris som respons, og antall brikker vil være forklaringsvariabel. Eksempler på tilhørende grupperinger kan da være henholdsvis:

(a) kjønn (lego for gutter, jenter, kjønnsnøytralt)
(b) varemerke (ja, nei, ingen informasjon) (Disney er for eksempel et eget varemerke)
(c) alder (baby/småbarn, skolebarn, eldre, uvisst) (presisering: DUPLO er for eksempel for baby/småbarn, Star Wars er for eldre)
(d) hva gruppemedlemene eier/eide som barn (ja, nei, usikker/fantes ikke)

Den nye kategoriske variablen skal ha maksimalt 4 ulike kategorier (helst akkurat 3), og kun fantasien setter grenser.

2. Finne (det har vi gjort for dere) og laste ned data. Og deretter rense dataene (det har vi også gjort for dere).

3. Pre-prosessere dataene så dere kan svare på problemstillingen (altså: gruppere temaene). Husk å beskrive hvilke temaer som havner i hvilken kategori **og hvorfor**! Hele datasettet skal brukes.

4. Formulere og begrunne modell(er) og hypotese(r) for å svare på problemstillingen. Husk å forklare hva som er respons, hva som er/kan være forklaringsvariabler, **og hvorfor**!

5. Tilpasse modell(er) og evaluere den/dem.

6. Diskutere om problemstillingen kunne besvares med datasettet dere endte med (hvorfor/hvorfor ikke?). Hva ble konklusjonen? (Husk at det ikke er lov til å endre problemstilling etter dere har undersøkt dataene).

Rapporten dere skal levere skal struktureres i samme rekkefølge som pipelinen for oppgaven. I tillegg til at rapporten inneholder relevante analyser og faglig innhold, legges det også vekt på at den er sammenhengende. Husk å forklare teori, metoder og begreper, referere til kilder dere bruker, og rapportere, tolke og diskutere relevante modeller, tall og resultater. Lærebok, kompendier og videoer er helt greit å referere til. Rapporten skal inneholde figurer (husk relevante navn på akser). Koden dere har brukt skal IKKE inkluderes underveis i rapporten, men til slutt som et appendix (eller la koden være tilgjengelig på nett).

Rapporten skal være på **maksimalt 5 sider** (ikke inkludert referanser og kode), skrevet med Times New Roman skriftstørrelse 12, Calibri (body) skriftstørrelse 11, eller lignende. **Kun de første 5 sidene rettes!!!**

Merk: Problemstillingen skal defineres **først**, og ikke endres i ettertid når dere ser resultatene. En god rapport er ikke ekvivalent med spennende resultater – å svare at "vi har ikke grunnlag for å svare på problemstillingen" (med analyser og begrunnelser) er også et godt svar. Vi anbefaler at dere sjekker med faglærer (Sara) at problemstillingen deres er god før dere begynner.

Referanser

Peterson, A. D. and Ziegler, L. (2021). Building a Multiple Linear Regression Model With LEGO Brick Data. *Journal of Statistics and Data Science Education*, 29(3):297–303.

Vi har laget et svarark i Word dere **skal** bruke. Der står alle spørsmålene i oppgave 2 og 3, og dere skal skrive svarene rett inn der. Når det i oppgaveteksten står at dere skal legge ved et bilde, kan dere ta et skjermbilde av jupyter notatboka (eller bruke den oppgitte kodesnutten for lagring av plott som oppgis i `oppg_1.ipynb`) og lime inn i Word. Der dere skal levere kode kan dere lime inn koden, eller ta skjermbilde. Her skal kun det vi spør etter være med i svarene, og merk at det ofte står en begrensning på antall setninger dere skal svare med.

Oppgave 2 - Klassifisering

Se Jupyter notatboka som heter `'oppg_2.ipynb'`, og svarark for utfylling her.

Oppgave 3 - Klyngeanalyse

Se Jupyter notatboka som heter `'oppg_3.ipynb'`, og svarark for utfylling her.