Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Drughunters 2025 – Matematikopgave

Introduktion

Drughunters Medicines Corp. er en (fiktiv) medicinalvirksomhed, der forsøger at udvikle nye typer medicin, der er mere effektive end de eksisterende. For at hjælpe med dette er forskere ved Drughunters Medicines Corp. interesserede i at finde ud af, hvor mange mennesker, der lider af en bestemt sygdom, hvor udbredt denne sygdom er, og om der er forskel på, hvor i landet, de syge befinder sig. Det gør de for at finde ud af, om de skal forske i denne sygdom.

Nu hvor vi lever i en stadig mere digital verden, bliver det nemmere og nemmere at indsamle og gemme data fra den virkelige verden, for eksempel data indsamlet fra områder i Danmark fordelt på forskellige faktorer: alder, køn etc. Ud fra disse data har vi en ny måde at forsøge at forstå mennesker, der lider af en bestemt sygdom.

I denne opgave kommer I til at arbejde som forskere hos Drughunters Medicines Corp., og I kommer til at anvende matematik og kunstig intelligens på data fra den virkelige verden for at få viden om tusindvis af mennesker diagnosticeret med skizofreni i alle aldre og fra alle dele af Danmark.

Efter selve opgaveformuleringen er der links til steder, hvor I kan få hjælp til at finde ud af, hvordan I udfører analyserne i f.eks. EXCEL. Der er ikke noget specifikt krav om brug af bestemt software.

Data (se hjemmesiden Drughunters) er præsenteret i en excel-fil. I bestemmer selv, om I vil omorganisere data på en helt anden måde. Data er hentet fra Udvalgte kroniske sygdomme og svære psykiske lidelser (esundhed.dk).

Hjernesygdommen

1) I datasættet er data fra mennesker diagnosticeret med skizofreni i Danmark. Beskriv kort sygdommen i nogle få sætninger (symptomer, behandlinger, prognoser osv.) ved at søge oplysninger om sygdommen.

Statistisk analyse af data ved hjælp af deskriptiv statistik

Når vi som forskere taler om, at vi skal analysere data, betyder det, at vi ønsker at få et overblik over data og at kunne drage nogle konklusioner vedrørende sammenhænge i data (hvis der overhovedet er nogle). Med sammenhænge menes, om flere variable påvirker hinanden, f.eks.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

køn og sygdom. Før vi begynder de mere avancerede matematiske analyser, er det meget vigtigt at bygge noget basisviden op omkring vores data.

2) Kig datasættet igennem og overvej (kort!) hvorfor antal mennesker diagnosticeret med skizofreni i datasættet er deleligt med 5, og hvorfor det mindste antal (udover 0) er 25

I kan bagerst i dette dokument se videoer om, hvordan I filtrerer (dvs. vælger) blandt de variable, I gerne vil bruge til analysen. Her er også videoer, der beskriver, hvordan I laver beskrivende/deskriptiv statistik i EXCEL, og hvordan I konstruerer diverse figurer, tendenslinjer etc.

3) Lav tabeller med beskrivende/deskriptiv statistik. I kan f.eks. starte med at se på antal voksne kvinder diagnosticeret med skizofreni i 2023 på tværs af alle regioner. De data er lagt i en separat fane i excel-arket (det kan være nødvendigt at lægge filtreringer separat for at få EXCEL til at udføre de analyser, I ønsker).

   Prøv selv at lave andre filtreringer og læg eventuelt i en separat fane for at kunne analysere (dvs. lave optællinger, figurer etc.).

4) Diskutér om der allerede nu kan etableres nogle konklusioner baseret på de tabeller og grafiske repræsentationer, I har lavet. Er antallet af mennesker diagnosticeret med skizofreni afhængig af f.eks. regioner? Køn? Alder?

I bestemmer selv, hvor meget beskrivende statistik og hvor mange figurer I vil præsentere for at give et overblik over (sammenhængene mellem) de vigtige variable i jeres datasæt.

I skal nu være omhyggelige med, hvordan I filtrerer for at kunne svare på de næste spørgsmål (og husk at lægge filtreringen for sig).

5) Lav et plot af årstal (tid) mod antal af mennesker for alle aldre diagnosticeret med skizofreni for hver af de 5 regioner. Indsæt tendenslinjer (regressionslinjer) i plottet på formen $y = ax + b$ og forklar for hver af de 5 linjer hvorvidt betydningen af a og b giver mening. Tilføj også R^2 og forklar, hvad dette tal viser.

6) Forklar med jeres egne ord, hvad I observerer i disse plots. Kan man på denne baggrund sige noget om udbredelsen af skizofreni?

7) Med disse 5 modeller kan I nu lave en forudsigelse (prædiktion) af, hvor mange mennesker diagnosticeret med skizofreni, der er i Danmark år 2030 i de forskellige regioner. Kommentér dette resultat.

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Hvordan kan studerende anvende avancerede dataanalysemetoder til at forstå og forudsige sygdomsudbredelse?

I denne del af opgaven skal I arbejde videre med de data, I har analyseret i statistikdelen, men nu med mere fokus på avancerede metoder inden for Data Science. I skal anvende såkaldt polynomisk regression til at modellere data og forudsige fremtidige tendenser.

Polynomisk regression er en form for regression, hvor sammenhængen mellem den uafhængige variabel (x) og den afhængige variabel (y) modelleres som et (n)te-grads polynomium, hvor n>1 (dvs. ikke en lineær regression). Dette kan være nyttigt, hvis data ikke følger en lineær tendens.

8) Forklar kort, hvad polynomisk regression er, og hvordan det adskiller sig fra den regression, der spørges til i 5). Diskutér i hvilke tilfælde en regression som i 5) ikke er tilstrækkelig til at modellere data.

9) Kig på jeres deskriptive arbejde og jeres tendenslinjer fra 7). Udvælg en region og en aldersgruppe (det kan også være ”Alle”) der har en udvikling, hvor polynomisk regression vil være brugbart.

Polynomisk regression kan give en bedre tilpasning til data men kan også føre til såkaldt overfitting, hvor modellen tilpasser sig godt til kendt data, men har en dårligere prædikteringsevne. Når man klassisk tilpasser en regressionskurve til data, bruger man ofte metoden ”Mean Squared Error” (MSE) til at evaluere nøjagtigheden af ens model. MSE beregner den gennemsnitlige kvadrerede forskel mellem de forudsagte værdier fra modellen og de rigtige værdier.

10) Hvorfor beregnes kvadratet af afvigelserne?

I dette sidste afsnit skal I bruge denne vedlagte Colab-Notebook, hvor der er lavet et interaktivt værktøj, hvor I kan lege med modelvalg, og antal variable.

11) Forklar med jeres egne ord, hvad der sker i det første eksempel i notesbogen, når antallet af variable øges. Hvad er årsagen til dette, og hvad skal man være opmærksom på?

12) Hvad sker der, når I bruger regressionen til at fremskrive udviklingen? Kan I bruge dette resultat til prædiktere den fremtidige udvikling af skizofreni tilfælde i Danmark?

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Vi er som forskere interesseret i at finde frem til den sande underliggende tendens. Vi bruger værktøjer og metoder til at komme så tæt på som muligt og accepterer, at virkeligheden er kompleks. Vi introducerer nu her til sidst en machine learning metode kaldet ”krydsvalidering”. I stedet for at tilpasse vores model på hele datasættet, opdeler vi det nu i et trænings- og testdatasæt. Vi tilpasser vores model på træningssættet og evaluerer med MSE på testsættet. Dette gøres nu på ”kryds og tværs”, hvor opdelingen af test- og træningsdata varieres, så modellen har været trænet på alle dele af data og evalueret på alle dele af data.

Der er mange forskellige måder at opdele trænings- og testdatasættet på. I andet afsnit af Colab-Notesbogen er der givet ét eksempel på, hvordan krydsvalidering kan bruges til at finde den bedste model ved hjælp af k-fold krydsvalidering.

13) Diskuter hvordan opdelingen af data i test- og træningsdata kan påvirke jeres endelige model.

I skal nu hjælpe Drughunters Medicines Corp. med at finde et område, hvor de skal fokusere deres forskning. Det skal gerne være en population, hvor der er en stigende tendens i antallet af skizofrenitilfælde. I kan bruge de interaktive widgets i notesbogen til hjælp.

14) Lav en forudsigelse for en udvalgt region og aldersgruppe (kan også være ”Alle aldersgrupper”) baseret på jeres bedste model, fremskriv til et selvvalgt årstal og argumenter for jeres valg. Diskuter om jeres forudsigelse er realistisk og om der er belæg for at bruge den polynomiske regression frem for den I fandt i 7).

Vi har i denne opgave kun brugt årstal som uafhængig variabel. Det er ikke en del af denne opgave at udforske dette område, men med mere avancerede machine learning modeller kan man også inddrage flere variable, også selv om de ikke er kontinuerte. Det kan være kategoriske variable som regionen eller ordinale variable som aldersgruppe.

15) I har nu lavet både statistik og data science i denne opgave.

a) Hvilke anbefalinger vil I give Drughunters Medicine Corp for deres fremtidige forskning indenfor Skizofreni?
b) Hvad er jeres forventninger til udviklingen af antallet af patienter, som diagnosticeres med skizofreni?

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Generel opgavevejledning

Overordnet set er opgaven opbygget efter følgende model:

- **Spørgsmål 1** omhandler en hjernesygdom og et datasæt. Her handler det primært om at vise, at man er i stand til at udvælge hovedtrækkene og give en så kort og præcis beskrivelse som muligt.
- **Spørgsmål 2-7** omhandler matematiske metoder til grafisk og analytisk at få overblik over og analysere data. Bemærk, at der generelt ikke er noget korrekt facit med to streger under. Alle besvarelser afhænger af de valg, der hele tiden foretages fra jeres side. Det er det, en forsker gør i sit daglige arbejde.
- **Spørgsmål 8-14** omhandler machine learning og metoder indenfor data science til analyser af data og grafisk visualisering og fortolkning af resultaterne. Dette skal ses som en videreudvikling af metoderne fra spørgsmål 2-7.
- **Spørgsmål 15** er en opsummering af, hvad I har lavet.

Vi vil rigtig gerne se jeres posterpræsentation uanset om I har svaret på alle dele af opgaven eller ej.

Til eleverne

Som forsker må man leve med, at der ikke findes endegyldige og korrekte svar. Man må opsøge viden, som andre har skabt eller ved at lave sine egne forsøg. Og så må man med åbent sind holde den viden op imod sin egen videnskabelige hypotese, som derved be- eller afkræftes – eller som oftest kræver yderligere viden for at kunne drage en konklusion. Det kan være en lang og frustrerende proces selv for garvede forskere. Derfor forventer vi selvfølgelig ikke endegyldige løsninger fra jer, men gode forslag hvor der er tænkt over usikkerheder og begrænsninger.

Vi har forsøgt at hjælpe ved at give nogle links nedenfor og på vores hjemmeside Drughunters. Men det er ikke en udtømmende liste, så I kan sikkert sagtens finde mere og anden information selv. At kunne opsøge information og have en kritisk tilgang til sine kilder er en meget vigtig kompetence som forsker.

Til finaledagen vil bedømmelseskriterierne være 1/3 formidling og 2/3 faglighed. Det betyder, at det ikke gælder om at have så meget tekst som muligt, men at der skal være et naturligt flow i fortællingen, så læseren/tilhøreren kan forstå jeres vigtigste pointer. Omvendt er det selvfølgelig heller ikke nok at have en superflot poster, hvis man ikke har svaret på spørgsmålene. Husk at til den mundtlige præsentation behøver I ikke at gennemgå posteren slavisk. Her skal I fokusere på at fremhæve de pointer, som er særligt vigtige for jeres besvarelse. Dommerne har læst posteren på forhånd, men gemmer den endelige bedømmelse til de har set jeres præsentation,

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

hvor de både vil inddrage jeres evne til at fortælle en sammenhængende historie og jeres besvarelse af opfølgende faglige spørgsmål.

Posteren skal ikke nødvendigvis være opdelt således, at I skal gennemgå alle 14 spørgsmål, men kan også være et udvalg af de figurer, tabeller og analyser, I har lavet.

Den skriftlige vurdering er selvfølgelig kun lavet på baggrund af posteren og skal ses som en kort tilbagemelding, ikke en dybtgående analyse af jeres poster.

Rent praktisk skal posteren indsendes som pdf i størrelsen 142x83 cm landskabsformat. Se kalenderen nedenfor.

Til lærerne

Brug gerne tid i klassen på at snakke om, hvordan hvert enkelt spørgsmål skal forstås, inden I kaster jer over besvarelsen. Der kan hentes inspiration til, hvordan man kan arbejde med opgaverne på vores hjemmeside Drughunters.

Hints og eksempler på referencer og links (find gerne flere selv)

I kan undervejs benytte EXCEL-funktionen 'Data Analysis'. Hvis I ikke finder den ude til højre under 'Data', kan I gøre følgende:

Tryk File -> Options -> Add-ins -> Analysis ToolPak (VBA) og installér dette

Deskriptiv statistik, optællinger og lineær regression i EXCEL:

How to Create Filter in Excel - YouTube

- Excel - statistisk bearbejdning af stort datasæt - YouTube
- Statistik i excel - YouTube
- Adding The Trendline, Equation And R2 In Excel - YouTube

Excel: Flere grafer i samme diagram (youtube.com)

Regression og krydsvalidering:

- Overfitting, modeludvælgelse og krydsvalidering – AI - Aalborg Intelligence (aalborg-intelligence.ai)

Link til Colab-Notebook med interaktive værktøjer – bruges til at løse opgave 8-13.

- Colab Notebook

Drughunters
Ottiliavej 9
2500 Valby
Tel +45 3630 1311
Fax +45 3630 1940
E-mail drughunters@lundbeck.com
www.drughunters.dk

Kalender for Drughunters 2025

Med venlig hilsen
Drughunters 2025