Statistisk sentralbyrå
Statistics Norway

# Kostholdsstatistikk

Metodenotat for publisering av 2018-årgangen

Magne Furuholmen Myhren og Trond Ekornrud

TALL SOM FORTELLER
NOTATER / DOCUMENTS
2024/28

I serien Notater publiseres dokumentasjon, metodebeskrivelser, modellbeskrivelser og standarder.



Publisert: 1. juli 2024
Rettet: 13. august 2024

ISBN 978-82-587-1999-8 (elektronisk)
ISSN 2535-7271 (elektronisk)

| Standardtegn i tabeller | Symbol |
|---|---|
| **Ikke mulig å oppgi tall**<br>Tall finnes ikke på dette tidspunktet fordi kategorien ikke var i bruk da tallene ble samlet inn. | . |
| **Tallgrunnlag mangler**<br>Tall er ikke kommet inn i våre databaser eller er for usikre til å publiseres. | .. |
| **Vises ikke av konfidensialitetshensyn**<br>Tall publiseres ikke for å unngå å identifisere personer eller virksomheter. | : |
| **Desimaltegn** | , |

# Forord

Helsedirektoratet har tidligere publisert kostholdsstatistikk basert på data fra forbruksundersøkelsen fra 1974 til 2012. Statistikken er publisert i ulike publikasjoner.

Arbeidet med å utvikle en ny kostholdsstatistikk ble påbegynt i 2018. Dette initiativet, som stammer fra Forbruksundersøkelsen, hadde som mål å integrere nye datakilder som tidligere ikke var benyttet i forbruksstatistikk.

Hovedformålet med statistikken er å gi en detaljert oversikt over kostholdet i den norske befolkningen i løpet av et kalenderår og per dag per person etter ulike matvaregrupper og næringsinnhold. Dette blir gjort ved å se på spiselig mengde kjøpte mat- og drikkevarer i dagligvarebutikker. Ettersom statistikken vil bli publisert årlig, vil den gi mulighet til å følge utviklingen i kostholdet i den norske befolkningen.

Statistikken skal bidra til å ivareta kunnskapsbehov slik disse er uttrykt blant annet i sentrale helsemyndigheters handlingsplaner og i en intensjonsavtale som er inngått mellom matbransjen og helsemyndighetene.

Dette notatet dokumenterer hvordan dataene har blitt behandlet for å produsere kostholdsstatistikken 2018.

Helsedirektoratet har finansiert utviklingen av denne statistikken.

Arbeidet har vært ledet av Trond Ekornrud og Magne Furuholmen Myhren. I tillegg har Eivind Kjeka Broen, Øyvind Berntsen Isachsen og Magnar Lillegård i Statistisk sentralbyrå (SSB) arbeidet med ulike deler av prosjektet. Elin Bjørge Løken og Anne Marte Wetting Johansen fra Universitet i Oslo har bidratt med gjennomgang og kvalitetssikring av predikerte data, samt å finne korrekte verdier for de mest solgte varene i utvalget. Susie Jentoft ved metodeseksjonen i SSB har bistått med faglig hjelp og støtte underveis i utviklingen av maskinlæringsmodellen og Annabelle Redelmeier i Norsk Regnesentral hadde den metodiske ideen til maskinlæringsmodellen.

Statistisk sentralbyrå, 27. juli 2024

Ann-Kristin Brændvang

# Sammendrag

Kostholdsstatistikken, som baserer seg på data brukt til å utvikle prisstatistikk (konsumpris-indeksen), viser detaljer om norsk kosthold basert på hvilke matvarer som selges i et representativt utvalg dagligvarebutikker.

Denne nye tilnærmingen til å estimere befolkningens kosthold basert på kjøpte mat- og drikkevarer, muliggjør hyppigere målinger og gir en innsikt i kostholdstrender og utvikling over tid sammenliknet med andre metoder.

Dette metodenotatet er et supplement til «Om statistikken», med mer detaljerte forklaringer av metodene vi har brukt. For å beregne tallene for 2018, har vi utviklet nye metoder som beskrives nærmere i dette notatet.

Kapittel 2 om datakilder introduserer de ulike kildene som har blitt benyttet til å utarbeide denne statistikken. Vi diskuterer noen utfordringer og valg som er tatt underveis for å tilpasse data som i utgangspunktet er samlet inn og brukt til andre formål.

En nøkkelkomponent i utarbeidelsen av statistikken er nøyaktig bestemmelse av hver vares vekt, som korresponderer med mengdevariabelen i prisdataene. I kapittel 3 beskrives de ulike metodene vi har brukt i prioritert rekkefølge.

Vi har gjort et omfattende arbeid for å identifisere korrekte energi- og næringsverdier for de mest solgte varene i datagrunnlaget. Disse verdiene brukes til å gi korrekte næringsinformasjon på de mest solgte matvarene, samt til å trene en maskinlæringsmodell som predikerer manglende verdier for de resterende varene, basert på tekstinformasjon på enkeltvare- og gruppenivå. I kapittel 4 om maskinlæring gjennomgår vi hovedpunktene i prosessen, og foreslår forbedringer for fremtidige utgivelser.

Norsk matvaregruppering er et nyutviklet kodeverk som er tilpasset norske forhold. Vi presenterer i kapittel 5 kort hvordan dette implementert. Til slutt i kapittel 6 skisseres hvordan vekting og skalering av datagrunnlaget er gjort for å få representative tall for hele befolkningen.

# Abstract

The dietary statistics, based on data used to develop price statistics (consumer price index), provide details about the Norwegian diet based on which food items are sold in a representative selection of grocery stores.

This new approach to estimating the population's diet based on purchased food and beverages allows for more frequent measurements and provides insight into dietary trends and developments over time compared to other methods.

This methodology note is a supplement to "About the Statistics," with more detailed explanations of the methods we have used. To calculate the figures for 2018, we have developed new methods described in more detail in this note.

Chapter 2 on data sources introduces the various sources that have been used to compile these statistics. We discuss some challenges and choices made along the way to adapt data that was originally collected and used for other purposes.

A key component in the preparation of the statistics is the accurate determination of each item's weight, corresponding to the quantity variable in the price data. In Chapter 3, the various methods we have used are described in prioritized order.

We have done extensive work to identify correct energy and nutrient values for the most sold items in the data set. These values are used to provide accurate nutritional information on the most sold food items, as well as to train a machine learning model that predicts missing values for the remaining items, based on text information at the item and group level. In Chapter 4 on machine learning, we review the main points of the process and propose improvements for future releases.

Norwegian Food Grouping is a newly developed code system adapted to Norwegian conditions. We briefly present in Chapter 5 how this is implemented. Finally, in Chapter 6, we outline how the weighting and scaling of the data basis is done to obtain representative figures for the entire population.

# Innhold

# 1. Introduksjon

Arbeidet med denne statistikken er en videreføring av tidligere arbeid med kvitteringsdata (bongdata). Dette innebar at dagligvarekjedene sendte kvitteringsdata for alle kjøp i løpet av et år til SSB. Kjøpene som var gjort med betalingskort kunne så kobles til eieren av kortet for å se på forskjeller i kjøpsmønstre mellom befolkningsgrupper.

I mai 2023 vedtok Datatilsynet et forbud mot bruk av kvitteringsdata. Etter dette bestemte SSB at vi enten skulle søke etter alternative datakilder eller avslutte prosjektet. Tidlig i 2023 vurderte vi muligheten for å benytte allerede innsamlede og bearbeidede aggregerte data, som anvendes for å beregne prisstatistikk (konsumprisindeksen). Selv om disse dataene liknet, var de ikke like nok til at vi kunne overføre metodene fra kvitteringsdataene direkte, så mange av databehandlings-prosessene og metodene måtte utvikles på nytt.

En vesentlig forskjell mellom kvitteringsdata og prisdata er detaljnivået. Kvitteringsdata åpnet for detaljert statistikk på husholdningsnivå, slik at vi for eksempel kunne sammenligne innkjøpsvaner med hensyn til mat for studenter og pensjonister, eller sammenlikne husholdninger etter utdan-ningsnivå med hensyn til innkjøpsvaner av mat- og drikkevarer.

Prisdataene er imidlertid ferdig aggregerte og tillater ikke differensiering mellom ulike hushold-ningstyper. Følgelig viser de publiserte tallene kostholdet til den gjennomsnittlige nordmann av alle aldre, sosiale grupper og geografisk bosted.

Første utgave av kostholdsstatistikken[1] ble publisert 12. juni 2024. Vi publiserer tre tabeller:

1. En tabell viser gjennomsnittlig mengde energi og næringsstoffer i mat som nordmenn kjøpte per person per dag i løpet av året.
2. De to andre tabellene viser den spiselige mengden kjøpte matvarer fordelt på ulike matvare-grupperinger. Den ene tabellen viser mengden i gram per nordmann per dag, mens den andre tabellen viser dette i kilogram per nordmann per år.

[1] https://www.ssb.no/helse/helseforhold-og-levevaner/statistikk/kosthald

# 2. Datakilder

Vi skiller mellom to type datakilder i denne statistikken: den ene kildetypen beskriver hvor mye det er solgt av hver vare for hver dagligvarekjede, den andre typen er kilder som inneholder relevant matvareinformasjon.

## 2.1. Aggregerte data fra et utvalg dagligvarebutikker i Norge (KPI-data)

Disse dataene viser mengde vare solgt per butikk per uke. Det er plukket ut tre hele uker for hver måned, slik at materialet går over 36 uker spredt utover året. Dette tallmaterialet danner mengdegrunnlaget for statistikken ved å summeres, skaleres og vektes til å representere mengde vare per år på nasjonalt nivå. Varene identifiseres ved strekkode (GTIN: Global Trade Item Number) eller dagligvarekjedenes interne varekoder, PLU (Product Look-Up).

For å sikre at vårt utvalg kun omfatter matvarer, har vi benyttet flere metoder for å filtrere bort ikke-spiselige varer. Dette inkluderer bruk av dagligvarekjedenes interne varekategorier, COICOP-grupperinger, samt en manuell gjennomgang utført i samarbeid med partnere fra Avdeling for ernæringsvitenskap ved Universitetet i Oslo, heretter omtalt UiO (Standard for COICOP2018, 2024).

Vi har valgt å ekskludere tørr kaffe og te fra utvalget, da næringsinnholdet for disse varene ofte er oppgitt i ferdigblandede størrelser. Disse varene bidrar også i liten grad til det totale nasjonale forbruket.

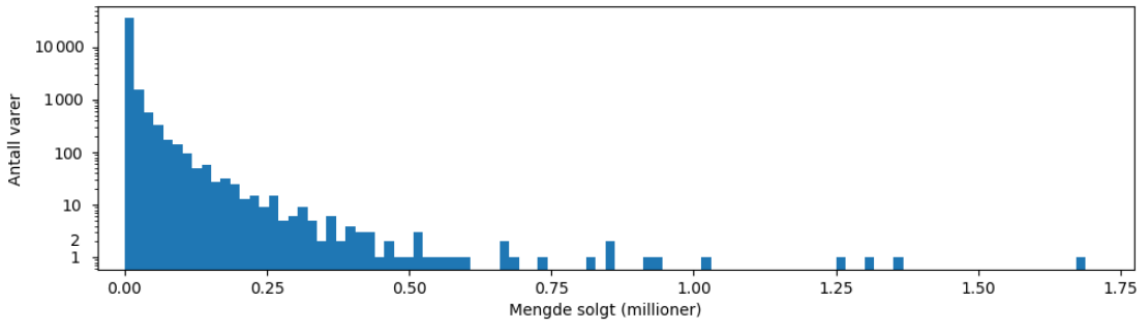
Når det gjelder salt som næringsstoff, er det vanskelig å fastslå hvor mye av det kjøpte kvantumet som faktisk konsumeres. Mye salt forsvinner i avløpet når man koker pasta eller trekker fisk, og en del brukes til formål utenfor matlaging. Som et kompromiss har vi valgt å utelate rene saltvarer som selges i beholdere over ett kilogram.

Varer med uspesifisert matinnhold er fjernet fra utvalget. Eksempler på slike varer inkluderer «SALATBAR», «DAGENS BAGUETT» og «KJØTT-TAST».

Denne vareutvelgelsen resulterte i et utvalg på om lag 40 000 unike varer i 2018-årgangen.

I grafikken under ser vi fordelingen av antall varer *y* med mengden millioner solgte enheter *x*. Mengdeenhetene kan være ulike for hver vare, så denne figuren gir mest et bilde av hvor mange varer som er «lite solgt» og «mye solgt» i butikkene vi har i utvalget.

**Figur 2.1 Logaritmisk histogram av mengde solgte enheter per unike id**




Vi kan se at det er mange varer som er lite solgt, og som ikke vil ha stor betydning for de endelige tallene i statistikken. En del av metodene vi har brukt for å få et fullstendig tallgrunnlag, går i stor grad ut på å finne verdier for de minst solgte varene. For de mest solgte varene har UiO bistått med

korrekte næringsverdier, slik at de mer usikre imputeringsmetodene får mindre betydning for sluttresultatet.

## 2.2. Vareinformasjon

### Energi- og næringsinnhold

For å kunne beregne total mengde energi og næring for befolkningen, må vi vite hvor mye energi og næring hver vare inneholder. Mattilsynet har innført en rekke krav til næringsdeklarasjon for ferdigpakket mat (Mattilsynet, 2023). Det finnes ikke digitale versjoner av alle disse næringsdeklarasjonene, men et stort antall er å finne blant et utvalg kilder:

- Tradesolution AS, datauttrekk fra 2020
- www.kassal.app, datauttrekk fra 2024
- Ulike nettsteder, som for eksempel www.unil.no

I disse kildene identifiseres hver vare med GTIN-strekkode.

Det er ikke krav til produsentene om å oppgi komplett informasjon om energi- og næringsinnhold, så selv om en vare finnes blant disse kildene, er det ikke sikkert at alle næringsstoffer har oppgitt verdi.

For varer som er identifisert med PLU-koder, er det vanskeligere å finne direktekilder til vareinformasjon. Eksempler på slike varer er «AGURK STK», «WIENERBRØD» og «BRIE LØSVEKT». Disse varene ble sortert etter hvor mye de er solgt og med bistand fra UiO har vi funnet representative varer i Mattilsynets matvareoppslag (Matvaretabellen).

For de mindre solgte matvarene og for manglende opplysninger, har vi brukt maskinlæring til å predikere energi- og næringsinnhold. Mer informasjon om dette kommer senere i dokumentet.

All tilgjengelig matvareinformasjon blir sammenstilt i ett datasett, hvor hver rad representerer ett datapunkt (*vare-id*, *variabel*, *verdi*, *kilde* og *dato*). Dette datasettet sorteres etter kildepreferanse og dato for å unngå dubletter. Hvis vi får tilgang til nye kilder til vareinformasjon, kan disse kumulativt legges til i denne filen.

# 3. Vekt og volum

I grunnlagsdataene har hver rad én mengdevariabel for antall solgte enheter per vare per butikk per uke. I dette delkapitlet forklares hvordan vi har brukt dette til å finne hvor mange kilogram som er kjøpt av varen og hvordan vi har beregnet spiselig del av den solgte mengden

## 3.1. Metodebeskrivelser

**Metode 1**
Vekt og volum hentes fra varedeklarasjoner i EPD-databasen til Tradesolution AS. Dataene oppgir antall enheter og typen enhet det er referert til. Alle varer som angir mengdeenhet i vekt eller volum, hentes fra datasettet og standardiseres til kilo eller liter. De mest solgte varene med mengdetyper som «porsjon», «stykke» osv., tildeles en gjennomsnittsvekt per enhet som deretter omregnes til kilogram.

**Metode 2**
Det er mange varer som har spesifisert vekt og volum i varenavnets tekstfelt i dataene fra butikkutvalget til dagligvarekjedene: «GULRØTTER 500g» og «REKER ½ kg». For andre varer må det beregnes totalvekt som et produkt av antall mindre enheter: «ØL 6x0,5L» regnes som $6 \cdot 0{,}5\ \mathrm{L} = 3{,}0\ \mathrm{L}$, og tilsvarende vekt blir dermed 3,0 kg. Varer der vekten er oppgitt som et intervall, eksempelvis «PINNEKJØTT 1-1,5 kg», tildeles en vekt basert på middelverdien av øvre og nedre grense: 1,25 kg. Vi har flere tilgjengelige tekstfeltkilder å søke etter størrelsesangivelser, så ved uoverensstemmelser mellom disse kildene, har vi laget en prioritert liste for hvilke kilder som har forrang. Det er stor variasjon i hvordan slike mengdestørrelser skrives i tekstfeltet, og det har vært krevende å lage løsninger som passer for alle tilfeller.

**Metode 3**
Varer med ikke-null desimalendelse i totale salgstall, regnes som løsvektvarer og får enheten kilogram. For at ikke feilregistreringer inkluderer varer feilaktig med denne metoden, har vi beregnet andel uker varen har ikke-null desimalendelse og satt 30% som en terskel.

**Metode 4**
Varer som ikke kan få verdi for vekt/volum fra tekst eller fra Tradesolution AS-data, blir estimert fra gjennomsnittsvekt fra varer i samme matvaregruppe (COICOP).

**Metode 5**
For visse matvaretyper vil det ikke bli korrekt å oversette volum (liter) direkte til vekt (kilogram). Iskrem, olje og syltetøy er eksempler på slike varetyper, der mengdebeskrivelsen oppgitt i volum må konverteres ved egne omregningsfaktorer (egenvekt). For andre varer det er naturlig å oppgi at volum (f.eks. drikkevarer) settes ved antall liter likt antall kilogram. (Mattilsynet, u.d.)

**Metode 6**
De få varene som ikke får tildelt vekt etter metodene over, tildeles en gjennomsnittsvekt for alle matvarer på omtrent 0,5 kg.

Alle mengdestørrelser normaliseres til kilogram slik at energi- og næringsinnhold per 100 gram kan multipliseres med 10 for å beregne totale mengder.

Resultatet av disse metodene er gjennomgått og korrigert etter søk for store avvik (uteliggere).

Vi har sørget for å oppnå nøyaktig vekt på de mest solgte varene, slik at de mindre nøyaktige metodene ikke gir betydelig utslag i de publiserte statistikktabellene.

## 3.2. Spiselig mengde

For å finne spiselig mengde av kjøpte matvarer, har vi benyttet variabelen «spiselig andel» og multiplisert med total mengde kjøpt vare. Spiselig andel tar hensyn til hvor mye av matvaren som består av ikke-spiselige deler som skall, bein, frø og liknende. Eksempelvis har vannmelon en spiselig andel på 46 % (Matvaretabellen - Rå vannmelon, 2024). Varer med manglende opplysninger om spiselig andel har fått verdien 100%.

Beregningen av spiselig mengde må gjøres, fordi både de deklarerte energi- og næringsverdiene på ferdigpakket mat og varer i Matvaretabellen oppgis per 100 gram spiselig mengde.

# 4. Maskinlæring til prediksjon av manglende verdier for energi- og næringsstoffer

På tross av at vi har tilgjengelig informasjon om et stort antall matvarer, er det allikevel behov for å finne metoder til å fylle ut (imputere) manglende verdier. Metoden for dette må være sterk nok til å kunne brukes på tvers av årganger og minimere behovet for manuelt utfylte verdier. Tabellen under viser andelen av verdier som må predikeres, både som andel av antall varer og andel av totalvekt. Det er stort sett verdiene for varene som er lite solgt som må predikeres. Derfor er andelen mye lavere for totalvekt solgt enn for antall varer (se figur 2.1).

**Tabell 4.1 Prosentandel varer med manglende verdier: Antall og totalvekt**

| | Prosentandel av totalt antall varer med manglende verdier | Prosentandel av total vekt for varer med manglende verdier |
|---|---|---|
| Energi, kcal | 30,0 | 3,2 |
| Energi, kJ | 29,9 | 3,2 |
| Karbohydrat, g | 30,0 | 3,3 |
| Sukkerarter, g | 30,1 | 3,3 |
| Feitt totalt, g | 30,4 | 3,4 |
| Metta fettsyrer, g | 31,4 | 4,3 |
| Kostfiber, g | 60,2 | 32,5 |
| Protein, g | 30,4 | 3,4 |
| Salt, g | 31,2 | 4,7 |
| Alkohol, g | 3,4 | 0,6 |

Det er få predikerte alkoholverdier, fordi vi satt 0 gram alkohol på alle varer som er plassert i hovednivånivå 1 (matvarer og alkoholfrie drikkevarer) i COICOP18. (Standard for COICOP2018, 2024).

Kostfiber er ofte ikke oppgitt hverken på fysisk innpakning, eller i våre kilder. Vi har testet å imputere 0 g / 100 g kostfiber for varer med andre oppgitte verdier, men gikk bort fra det fordi det blir for konservativt. Kostfiber er derfor næringsstoffet statistikken gir mest usikre tall for.

Under arbeidet med bongdata som datagrunnlag, fikk vi bistand fra Norsk Regnesentral, gjennom BigInsight-prosjektet, til å utvikle en maskinlæringsmodell for å predikere energi- og næringsinnholdet i varer basert på varetekst. Denne modellen brukte *k* Nearest Neighbors med Jaccard similarity for å klassifisere lignende matvarer, og en RandomForestRegressor for å predikere verdiene. (Redelmeier & Løland, 2022)

Endringer i datagrunnlaget gjorde at det ble behov for å tilpasse maskinlæringsmodellen. Vi tok utgangspunkt i eksisterende ideer og utviklet en ny modell basert på XGBoost (eXtreme Gradient Boosting). Vi har brukt Python med bibliotekene scikit-learn og XGBoost til prediksjon, samt Pandas til preprosessering av tokens.

Lenke til internt SSB GitHub-repo. Første publisering har tag «1.0.0»:
www.github.com/statisticsnorway/stat-helse-kosthold

## 4.1. Preprosessere tokens

Med maskinlæring kan vi trene en modell på eksisterende data og predikere næringsverdier utfra tekstinput. Teksten kalles en *tokenstreng* bestående av *tokens*. Til vårt formål har vi satt en token til å være ett ord (tekst adskilt med mellomrom).

Til test- og treningsdata har vi kun brukt varer som finnes i datagrunnlaget og et uttrekk fra Matvaretabellen (Matvaretabellen). Her er et eksempel på fem varer med tilhørende tekstinformasjon fra ulike kilder:

**Figur 4.2 Eksempel på tekstinformasjon for fem varer**

| EANTEKST | TEKST | varenavn | Markedsnavn | GRUPPETEKST | gruppenavn | gruppe_navn |
|---|---|---|---|---|---|---|
| RUNDSTYKKER FINE 6PK REMA | RUNDSTYKKER FINE 6STK REM | | | BAKEVARER BAGUETTER OG RUNDSTYKKER | | BAGUETTER HALVSTEKTE |
| GO-TAN CHILI HVITLØK WOK 240ML | GO-TAN CHILI HVITLØK WOK 240ML | go-tan chili hvitlok | go-tan chili hvitløk wok 6x240ml | Asiatisk mat | Sauser våte andre | |
| CHEERIOS HAVRE 500G NESTLE | CHEERIOS HAVRE 500G NESTLE | cheerios havre 500g | | FROKOSTCEREALER | FROKOSTCEREALER | |
| NORDFJORDSKINKE KRYDRET 2 | NORDFJORDSKINKE KRYDRET 2 | nordfjordsk. Kryd. | Nordfjordskinke krydret 250g | PÅLEGG OG KJØTTVARER | | Ferskt Pålegg |
| PIZZABUNN GLUTENFRI 350G BRISK | PIZZABUNN GLUTENFRI 350G BRISK | | | GLUTENFRIE BAKEVARER – BRØD | | PIZZA – DYPFRYST |

Målet for preprosesseringen er å slå sammen disse kolonnene til en tokenstreng som best mulig reflekterer hva slags vare en rad representerer.

Her er noen av operasjonene som gjøres:

- Slå sammen alle tekstkolonner:

  `»GO-TAN CHILI HVITLØK WOK 240ML GO-TAN CHILI HVITLØK WOK 240ML go-tan chili hvitlok go-tan chili hvitløk wok 6x240ml Asiatisk mat Sauser våte andre»`
- Gjør om alt til små bokstaver, fjern uønskede tegn og bytt ut ikke-norske tegn med tilsvarende norske:

  `'go-tan chili hvitløk wok 240ml go-tan chili hvitlok hvitløk wok 6x240ml asiatisk mat sauser våte andre'`
- Ta bort alle mengdestørrelser fra teksten og sørge for at forkortelser som `'u sukker'` og `'m skinn'` gjøres om til `'uten sukker'` og `'med skinn'`:

  `'go-tan chili hvitløk wok  go-tan chili hvitlok hvitløk wok  asiatisk mat sauser våte andre'`
- Fjern enkeltbokstaver og hvite mellomrom:

  `'go-tan chili hvitløk wok go-tan chili hvitlok hvitløk wok asiatisk mat sauser våte andre'`
- Slå sammen alle etterfølgende par av ord med understrek (bigrams) for å få med betydning som `'uten_sukker'`. Dette steget lager mange nye tokens og kan se ut til å lage støy, men modellen ser bort i fra sjeldne tokens slik at kun dobbeltord som dukker opp flere ganger blir med i modellen.

  `'go-tan chili hvitløk wok go-tan chili hvitlok hvitløk wok asiatisk mat sauser våte andre go-tan_chili chili_hvitløk hvitløk_wok wok_go-tan go-tan_chili`

```
chili_hvitlok hvitlok_hvitløk hvitløk_wok wok_asiatisk asiatisk_mat mat_sauser
sauser_våte våte_andre'
```

- Legg på COICOP18-gruppenummer i tre varianter til slutt i tokenstrengen. Dette gir modellen flere nivåer av gruppetilhørighet og kan være god tilleggsinformasjon om hva slags vare det er.

```
'go-tan chili hvitløk wok go-tan chili hvitlok hvitløk wok asiatisk mat sauser
våte andre go-tan_chili chili_hvitløk hvitløk_wok wok_go-tan go-tan_chili
chili_hvitlok hvitlok_hvitløk hvitløk_wok wok_asiatisk asiatisk_mat mat_sauser
sauser_våte våte_andre 01194 0119 011'
```

Vi har valgt å ikke fjerne vanlige ord (stop words) som for eksempel «og», «er», «på» og «grønn», fordi det er vanskelig å se rekkevidden av slike valg. Å ta bort et ord som er irrelevant i én sammenheng, kan være viktig i en annen. Dette er vanskelig å vite på forhånd. Derfor har vi fokusert mer på å tilpasse eksisterende ord til norsk skrivemåte, for eksempel «entrecôte» til «entrecote», og kun fjernet tokens som inneholder vekt, volum og antall. Fargeord er et eksempel på denne problemstillingen. Vi forsøkte å fjerne alle fargeord, men oppdaget at uttrykk som «rød saus» og «brun saus» ble redusert til «saus», noe som gjorde det vanskeligere for modellen å skille mellom dem. Et annet eksempel er «grønn kiwi» og «grønn tuborg», hvor «grønn» beskriver både frukt og øl, men de har svært ulike næringsverdier. Modellen vil forhåpentligvis lære at «grønn» er en dårlig token, ettersom den vil oppdage liten samvariasjon på tvers av varer med denne beskrivelsen.

Vi har valgt å ikke fjerne produktmerker som «tine», «rema1000», «jacobs» osv, fordi det kan hende at modellen finner disse ordene som nyttige på en måte vi ikke har forutsett. I tillegg blir det tilfeldig hvilke produktmerker vi finner og tar ut blant de 40 000 varene som skal prosesseres.

Merk at maskinlæringsmodellen tolker tall i teksten som ren tekst, ikke som numerisk verdi. Derfor fjerner vi alle tall etterfulgt av en vektstørrelse. Å fjerne alle tall vil ødelegge varenavn som «7-up», og fett- og alkoholinnholdig informasjon som «melk 0,5% fett» og «pils 4,5%».

For å håndtere tekstdata bruker vi funksjonen TfidfVectorizer fra pakken scikit-learn. Dette verktøyet omdanner tekst til en numerisk form som datamaskinen kan forstå. Det fungerer ved å vektlegge ord basert på hvor ofte de dukker opp i dokumentet og hvor sjeldne de er i hele datasettet. Vanlige ord som dukker opp i mange dokumenter får lavere vekt, mens mer unike ord får høyere vekt. Mye av dette arbeidet handler om å systematisk fjerne, eller tilpasse ord som beskriver hver vare, uten å ødelegge for andre varer. Her er innstillingene vi bruker:

```
vec = TfidfVectorizer(
    ngram_range=(1, 1),  # bigrams er manuelt laget i tidligere steg
    max_features=None,   # ingen øvre grense på antall tokens
    min_df=2             # bruk bare tokens som er brukt 2 eller flere ganger
)
```

Valgene som tas under preprosesseringen er basert på Norsk Regnesentrals kode og er stegvis korrigert på nytt datagrunnlag gjennom prøving og feiling. Alle varene i datagrunnlaget får en tokenstreng behandlet på lik måte. Disse strengene danner grunnlaget for modelltrening og prediksjon av næringsverdier.

## 4.2. Valg av modell

Vi har valgt å bruke XGBoost (eXtreme Gradient Boosting) til å predikere energi- og næringsverdier. En studie fra 2023 tok i bruk denne modellen til å predikere næringsverdier for en matvare basert på varenes ingredienslister (Hu, Ahmed, & L'Abbe, 2023). Denne modellen ga mye raskere kjøringer enn modellen Norsk Regnesentral benyttet (RandomForestRegressor) og ga tilfredsstillende resultater. Vi har valgt å ikke ta i bruk ingredienslister til prediksjonen, fordi en stor del av varene i datagrunnlaget manglet denne informasjonen.

## 4.3. Resultater

For hvert næringsstoff filtreres varer med manglende verdier ut, slik at vi sitter igjen med alle varer vi har næringsverdien til. Dette datasettet deles tilfeldig inn i et treningssett på 80% av radene og et testsett på de resterende 20%. Modellen trenes på treningssettet og predikerer verdier i testsettet, slik at vi kan sammenlikne faktiske verdier med modellens prediksjoner. Disse sammenlikningene gjøres ved å beregne $R^2$, også kjent som determinasjonskoeffisienten, et tall mellom 0 og 1 som angir hvor godt modellen forklarer variasjonen i de faktiske verdiene.[2]

- $R^2 = 1$ indikerer at modellen forklarer all variasjon i de faktiske verdiene
- $R^2 = 0$ indikerer at modellen ikke forklarer noe av variasjonen i de faktiske verdiene

Her er et eksempel på fem av prediksjonene på testsettet ved en modellkjøring:

**Figur 4.1 Prediksjon av kilokalorier per 100 gram vare for fem tilfeldige varer**

Variabel kcal (25928) varer

| tokens_bi | kcal | kcal_pred |
|---|---|---|
| solidox tyggis rene tenne godteri og snacks tyggegummi sukkerfri solidox_tyggis tyggis_rene rene_tenne tenne_godteri godteri_og og_snacks snacks_tyggegummi tyggegummi_sukkerfri 01189 0118 011 | 187.0 | 257.606598 |
| mack isbjørn øl eng fl lys pilsner isbjorn engl ma mac 4.5\% pils alkoholinnholdig drikke mack_isbjørn isbjørn_øl øl_eng eng_fl fl_isbjørn isbjørn_mack mack_fl isbjørn_lys lys_pilsner pilsner_mack mack_isbjorn isbjorn_fl fl_mack isbjørn_fl fl_engl engl_ma ma_mack isbjørn_isbjørn engl_mac mac_isbjørn mack_4.5\% 4.5\%_pils pils_alkoholinnholdig alkoholinnholdig_drikke 02130 0213 021 | 39.0 | 47.312675 |
| potetgull salt&vinegar maarud salt & vinegar snacks og chips potetchips potetgull_salt&vinegar salt&vinegar_maarud maarud_salt salt_& &_vinegar vinegar_potetgull maarud_snacks snacks_og og_chips chips_potetchips 01199 0119 011 | 515.0 | 510.739380 |
| fløtepudding karmøy karmoy flotepudding fiskepudding industripakket fløtepudding_karmøy karmøy_karmoy karmoy_flotepudding flotepudding_fløtepudding karmøy_fiskepudding fiskepudding_industripakket 01133 0113 011 | 105.0 | 158.801178 |
| smør croissant premium annen wienerbakst - halvstekt smør_croissant croissant_premium premium_annen annen_wienerbakst wienerbakst_- -_halvstekt 01113 0111 011 | 404.0 | 302.073547 |

R-squared: 0.76

Kolonnen «kcal» viser den faktiske verdien og «kcal_pred» angir modellens predikerte verdi. Nederst til venstre kan vi se $R^2 = 0.76$ for denne kjøringen på testsettet.

[2] En feil i teksten er rettet den 13. august 2024. Tidligere sto det at R2 er korrelasjonskoeffisient; dette er nå korrigert til determinasjonskoeffisient.

Her er tilsvarende kjøring for karbohydrat:

**Figur 4.2 Prediksjon av gram karbohydrat per 100 g vare for fem tilfeldige varer**

Variabel karbo (25919) varer

| | tokens_bi | karbo | karbo_pred |
|---|---|---|---|
| 0 | spekeskinke brett stranda westfaler fjordskinke kjøttvarer og pølser kjøtt spekevarer spe hel ben spekemat kjøl spekeskinke_brett brett_stranda stranda_westfaler westfaler_spekeskinke spekeskinke_fjordskinke fjordskinke_brett brett_kjøttvarer kjøttvarer_og og_pølser pølser_kjøtt kjøtt_spekevarer spekevarer_spe spe_kjøtt spekevarer_hel hel_ben ben_spekemat spekemat_kjøl 01123 0112 011 | 0.0 | 1.415710 |
| 1 | milky way mini godteri og søtsaker sjokolade stykksaker uinnpk. drops milky_way way_mini mini_godteri godteri_og og_søtsaker søtsaker_sjokolade sjokolade_stykksaker stykksaker_uinnpk. uinnpk._sjokolade sjokolade_og og_drops 01185 0118 011 | 76.6 | 58.077374 |
| 2 | lofoten laks backloin back loin fisk fersk industrip indpk kjøl lofoten_laks laks_backloin backloin_lofoten backloin_back back_loin loin_fisk fisk_fersk fersk_industrip industrip_indpk indpk_kjøl 01131 0113 011 | 0.0 | 2.117081 |
| 3 | sanasol diverse ingredienser sanasol_diverse diverse_ingredienser | 51.0 | 10.694229 |
| 4 | musli blackberry date&qinoa axa quinoa kornprodukter og frokostblanding kornblandinger musli_blackberry blackberry_date&qinoa date&qinoa_axa axa_musli blackberry_quinoa quinoa_axa axa_kornprodukter kornprodukter_og og_frokostblanding frokostblanding_kornblandinger 01114 0111 011 | 53.0 | 57.513184 |

R-squared: 0.82

Tabellen under viser modellens $R^2$-resultater sammenstilt med andre metoder for å estimere næringsverdier.

**Figur 4.3 Sammenlikninger av $R^2$ for ulike metoder**

| | XGBoost Prediksjon | COICOP18 | COICOP6 | Alle varer |
|---|---|---|---|---|
| Energi, kcal | 0,73 | 0,57 | 0,65 | 0,00 |
| Energi, kJ | 0,77 | 0,59 | 0,66 | 0,00 |
| Feitt totalt, g | 0,72 | 0,53 | 0,62 | -0,01 |
| Metta fettsyrer, g | 0,72 | 0,53 | 0,58 | -0,01 |
| Karbohydrat, g | 0,81 | 0,68 | 0,72 | -0,01 |
| Sukkerarter, g | 0,75 | 0,55 | 0,60 | -0,01 |
| Protein, g | 0,72 | 0,58 | 0,61 | -0,01 |
| Salt, g | 0,47 | 0,12 | 0,19 | 0,00 |
| Kostfiber, g | 0,59 | 0,32 | 0,32 | 0,00 |
| Alkohol, g | 0,68 | 0,88 | 0,00 | 0,00 |

- **XGBoost Prediksjon:** Gjennomsnittlig $R^2$-verdi av fem modellkjøringer med en fordeling på 80%-20% av trenings- og testsett.
- **COICOP18:** Vi beregner gjennomsnittsverdier for alle varer i samme COICOP18-gruppe og varenes faktiske verdier måles mot dette gjennomsnittet. Vi måler $R^2$ mellom disse verdiene. Merk at ikke alle varer har COICOP-gruppering, og at disse ikke blir med i sammenlikningene.
- **COICOP6:** Vi gjør det samme som for COICOP18, bare med COICOP6-gruppering (en tidligere versjon av matgrupperingen). Ingen varer med alkohol har COICOP6.
- **Alle varer:** Gjennomsnittsverdier for alle varer.

Her kan vi se at for alle verdier, foruten alkohol, gir modellens prediksjoner i gjennomsnitt en høyere $R^2$ enn for de andre metodene. COICOP18 gir gode verdier alkohol, fordi det er i denne grupperingen skiller mellom alkoholinnholdig drikke og ikke (Standard for COICOP2018, 2024). Alkohol er også næringsstoffet vi trenger færrest predikerte verdier for, så dette gir ikke så stort utslag på sluttresultatene.

Det er også interessant å se at COICOP-grupperingene er en relativt god indikator på næringsverdier og at salt og kostfiber gir lavere $R^2$ på tvers av alle metodene.

Før vi publiserte tallene ble alle tilgjengelige varer brukt til trening av modellen, altså ingen 80%-20%-fordeling.

I noen tilfeller predikerer modellen negative næringsverdier. Disse verdiene blir i postprosesseringen justert til 0. Dette er sannsynligvis en konsekvens av at tokenstrengen inneholder mange tokens som bidrar negativt i betydelig grad. I resultatene over er negative prediksjoner tatt med i beregningen av $R^2$.

## 4.4. Mulige forbedringer og videre utvikling

En naturlig fortsettelse av dette arbeidet vil være å forsøke ulike hyperparametere i XGBoost-modellen og å gjøre flere tilpasninger i preprosesseringen. Det finnes metoder for å normalisere tokens til en standardform (stemming). Det vil for eksempel si å omforme kjente norske substantiv til bestemt form entall, slik at «banan» og «bananer» blir samme token «banan», og dermed likestilles av modellen. Dette har vi ikke prøvd ut enda.

Vi kan også gjøre forsøk med å bruke alle tilgjengelige datapunkter til modelltrening. Vi har, som tidligere nevnt, valgt å kun bruke matvarer i utvalget for å sikre en likest mulig tokenstreng. Det kan imidlertid være at denne antakelsen er feil, og at vi bør inkludere så mange datapunkter som mulig. Dette kan testes videre ved senere publiseringer. Det gjenstår også å sjekke hvorvidt inkludering av Matvaretabellen i modelltreningen bør gjøres eller ikke, da disse får en annen type tokenstreng enn varene i datagrunnlaget. Se eksempelet med Sanasol i Figur 4-2, hvor denne varen ikke har COICOP18 nummerering i token-strengen.

Det finnes fysiske sammenhenger mellom flere av de predikerte verdiene. For eksempel er kcal og kJ begge mål for energi med ulik enhet. Vi har valgt å predikere begge hver for seg uten omregning, og har gjort tester på det ferdige resultatet og sjekket at omregningen stemmer godt. Det finnes også formler som regner om fett, karbohydrat, alkohol og kostfiber til energi. Dette er ikke hensyntatt i denne publiseringen og er et mulig forbedringspunkt senere.

# 5. Norsk matvaregruppering

Helsedirektoratet, som har finansiert utviklingen av statistikken, har ønsket å bruke en matvaregruppering som er bedre tilpasset norske forhold. Den internasjonale standarden, COICOP, skiller for eksempel ikke godt nok mellom grovt- og fint brød og fet- og mager fisk.

UiO har, i forbindelse med publiseringen av denne statistikken, utviklet en ny norsk matvaregruppering. Denne består av 15 hovedkategorier med to undernivåer[3].

For å kunne plassere hver matvare i disse gruppene, har vi tatt utgangspunkt i hver vares COICOP-gruppering og brukt tilgjengelig informasjon om næringsinnhold, ingredienser og navnesøk. Her er et eksempel på bruk:

**Eksempelkode**

```
#- 2.3 Salta/røykt/tørka kjøt ------------------------------------------------
        coicop18 == «01123» ~ case_when(
            !match.ingred(«kalkun|kylling») ~ case_when(
    #-> 2.3.1 Raudt heilt kjøt, ≥6% salt
                salt >= 0.06                                        ~ «2.3.1»,
    #-> 2.3.2 Spekepølse, salami, ≥4-5,9% salt
                salt >= 0.04 & salt < 0.06                          ~ «2.3.2»,
    #-> 2.3.3 Lettsalta kjøt, bacon mv, >1,9-3,9% salt
                salt > 0.019 & salt < 0.04                          ~ «2.3.3»,
    #-> 2.3.4 Kjøtpålegg, raudt, ≤1,9% salt
                salt <= 0.019                                       ~ «2.3.4»,
    #-> 2.3.6 Salta/røykt/tørka kjøt, ukjent salt
                is.na(salt)                                         ~ «2.3.6»,
                TRUE ~ «2.3.1-2.3.4, 2.3.6 NEC»
            ),
    #-> 2.3.5 Kjøtpålegg, kvitt
            match.ingred(«kalkun|kylling»)                          ~ «2.3.5»,
            TRUE ~ «2.3 NEC»
        ),
```

NEC-kategoriene (Not Elsewhere Classified) vises ikke eksplisitt i de publiserte tabellene, men de blir tatt med i totalsummen av nivåene over.

De mest solgte varene som av denne metoden ble plassert i NEC-grupper, ble i flere omganger plassert i passende grupper av UiO. Dette håper vi å kunne klassifisere med maskinlæring ved senere publiseringer.

Predikerte næringsverdier blir ikke brukt til å fordele varene i matvaregrupperingen. Det er laget egne grupper for varer med mangelfulle opplysninger om næringsverdier (se gruppe 2.3.6 i eksempelet over).

[3] https://www.ssb.no/klass/klassifikasjoner/716

# 6. Vekting og skalering av datagrunnlaget

Grunnlagsdataene fra 2018 er hentet fra et utvalg av 169 dagligvarebutikker. Disse butikkene varierer i størrelse og prissammensetning og det er grunn til å anta at nordmenns handlevaner varierer avhengig av hvilken butikk de handler i. Vi har derfor valgt å vekte utvalget av butikker ved å justere kjedeomsetningen i utvalget mot kjedeomsetningen på landsbasis (ratemodellering). Omsetningstall har vi funnet i VoF (Virksomhets- og foretaksregisteret, 2024), nærmere bestemt omsetningstallene for næring 47.111 Butikkhandel med bredt vareutvalg med hovedvekt på nærings- og nytelsesmidler.

Her er en skisse av hvordan tallene skaleres etter vekting:

| | |
|---|---|
| $u$: | antall uker i datamaterialet |
| $U$: | antall uker i året (52) |
| $B$: | befolkningstall i Norge ved utgangen av fjerde kvartal |
| $d$: | antall dager i året, justert for skuddår |
| $n$: | vektor med totale næringsmengder i Norge etter omsetningsvekting |
| $N$: | næringsmengder per nordmann per dag (publiserte tall) |

$$N = \frac{n}{B \cdot d} \cdot \frac{U}{u}$$

# Referanser

Hu, G., Ahmed, M., & L'Abbe, M. (2023). *Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods.* Hentet fra resarchgate.net: https://www.researchgate.net/profile/Guanlan-Hu/publication/366539470_Natural_language_processing_and_machine_learning_approaches_for_food_categorization_and_nutrition_quality_prediction_compared_to_traditional_methods/links/6462761b434e26474feb1d0b/Natur

*Kodeliste for matvaregruppering.* (2024). Hentet fra ssb.no/klass: https://www.ssb.no/klass/klassifikasjoner/716

*Mattilsynet*. (u.d.). Hentet fra www.mattilsynet.no: www.mattilsynet.no/mat-og-drikke/matvaretabellen/mal-vekt-og-porsjonsstorleikar

Mattilsynet. (2023, 01 27). *Mattilsynet*. Hentet fra mattilsynet.no: https://www.mattilsynet.no/mat-og-drikke/merking-av-mat/naeringsdeklarasjon-for-ferdigpakket-mat

*Matvaretabellen.* (u.d.). Hentet fra www.matvaretabellen.no

*Matvaretabellen - Rå vannmelon*. (2024, juni 15). Hentet fra www.matvaretabellen.no: https://www.matvaretabellen.no/vannmelon-ra/

Redelmeier, A., & Løland, A. (2022). *Predicting a food product's missing nutritional.* Hentet fra Nordic Statistical Meeting: https://www.nsm2022.is/s/PREDICTING-A-FOOD-PRODUCTS-MISSING-NUTRITIONAL-VALUES-USING-MACHINE-LEARNING.pdf

*Standard for COICOP2018*. (2024). Hentet fra https://www.ssb.no/klass: https://www.ssb.no/klass/klassifikasjoner/691

*Virksomhets- og foretaksregisteret.* (2024). Hentet fra https://www.ssb.no/a/metadata/om_datasamlinger/virksomhets-_og_foretaksregisteret/bof.html