UDBYTTEFORUDSIGELSE I VINTERHVEDE
VED BRUG AF MACHINE LEARNING

SEGES INNOVATION

STØTTET AF

SEGES INNOVATION

Promilleafgiftsfonden for landbrug

UDBYTTEFORUDSIGELSE I VINTERHVEDE VED BRUG AF MACHINE LEARNING

Af Mette Kramer Langgaard, Lasse Rose Malskær og Jacob Hein

1.1 Sammendrag

Det er vigtigt at kunne estimere det forventede udbytte i marken med en vis præcision, hvis for eksempel kvælstofbehovet til afgrøden skal fastsættes korrekt. I 2020 blev første version af udbytteprognosen i vinterhvede implementeret i management programmet CropManager.

I 2021-22 er der, i projektet "*Lær af verdens største forsøgsareal"* støttet af Promilleafgiftsfonden, indsamlet et datagrundlag fra 2016-2021 for at forbedre udbytteprognosemodellerne i vinterhvede. Målet med aktiviteterne var at undersøge, om udbyttet kan forudsiges med en god nok præcision til, at landmanden kan regulere kvælstoftildelingen til marken midt i maj. Prognosemodellerne er udviklet på baggrund af udbyttedata fra mejetærskere, satellitdata, terrænhøjdedata, vejrdata, jordtype, lokalitet, sort og hovedafgrøden 5 år tilbage.

Resultaterne af dataanalysen viser, at udbytteprognosemodeller kan forudsige udbyttet i vinterhvede med en nøjagtighed (MAE) af prædiktionen på 6,5 og 5,5 hkg pr. ha på markniveau henholdsvis 4. maj og 27. juli. De nye modeller er nu så nøjagtige, at landmanden kan anvende udbytteprognosen til at regulere kvælstoftildelingen i vinterhvede ved tredje tildeling i midten af maj. Prognosen skal forsat anvendes med forsigtighed på sandjord og i vækstsæsoner som afviger fra 2016-2021.

INDHOLD

1.2 Baggrund

Kvælstofbehovet på marken beregnes bl.a. ud fra jordtype, forfrugt, vanding og tilført husdyrgødning m.m. Det forventede udbytte indgår ligeledes i fastsættelsen af kvælstofbehovet på markniveau, og hvis udbyttet fejlestimeres med 10 hkg pr. ha, har landmanden sandsynligvis over- eller undergødsket vinterhveden med ±15 kg kvælstof pr. ha. Derfor er det relevant at kunne forudsige det forventede udbytte i marken inden sidste kvælstoftildeling i vækststadium 37 (midt maj), så kvælstofbehovet i marken kan reguleres efter afgrødens behov på lokaliteten. En valid udbytteprognose kan også anvendes til planlægning af logistik i høst, forventet indkøb af supplerende foder mv.

Petersen et al., 2023 har i et studie i vårbyg med 9 års udbyttedata, RVI-målinger og vejrdata kunne forudsige udbyttet i en forsøgsmark med en nøjagtighed (Mean Absolut Error = MAE) på 3,8 hkg pr. ha omkring en måned før høst. Resultaterne viser potentialet i udbytteprædiktion i kornafgrøder, og vigtigheden af registrering af udbytter

Første version af udbytteprognosen i vinterhvede blev implementeret i CropManager i 2020 (se resultaterne af analysen her). I 2021-22 er der i Promilleprojektet *"Lær af verdens største forsøgsareal"* indsamlet flere udbyttedata og flere parametre er tilføjet modellen. Formålet er at udvikle en robust model som på tværs af år, jordtype og lokalitet kan forudsige udbyttet i vinterhvede med en god nok præcision til, at landmanden bl.a. kan regulere kvælstoftildelingen til marken i vækststadium 37 (start maj).

1.3 Metode

Datagrundlag

Der er anvendt både offentlige og fortrolige datakilder til at udvikle udbytteprognosen i vinterhvede. Modellen er testet med følgende data: markpolygoner, udbyttedata fra mejetærskere, Sentinel-2 L1C data, terrænhøjdedata, vejrdata, jordtype, vinterhvedesort, hovedafgrøden 5 år bagud, og koordinater (lokalitet). Alle data er konverteret til WGS84/UTM32N (EPSG:32632) og inddelt i grid på 10 x 10 meter, hvilket giver et datasæt på 293.829 observationer (pixels). I udvikling af modellen indgår i alt 791 variabler.

Udbyttedata

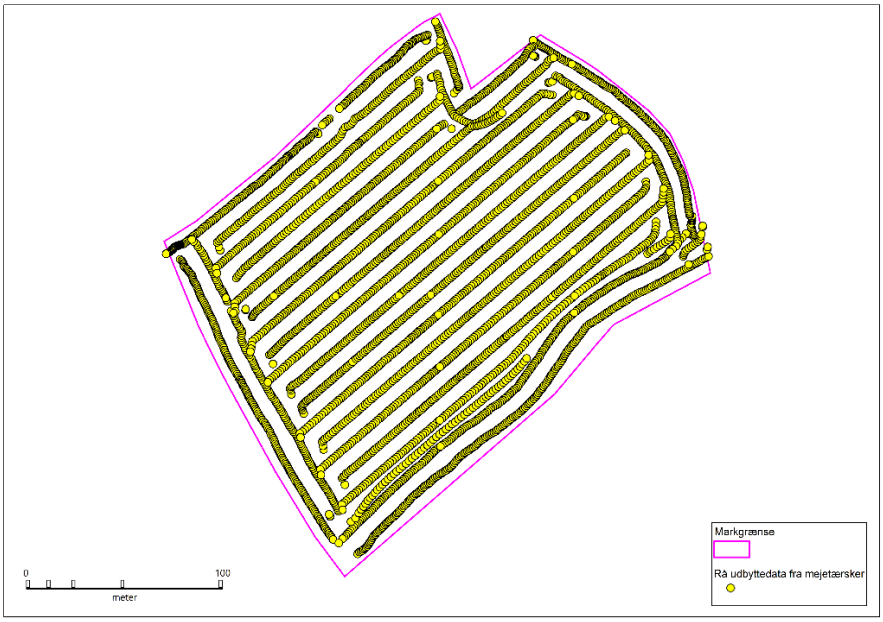
Positionsbestemte udbyttedata fra mejetærskere er indsamlet fra 2016-2021. Data er indsamlet på to tidspunkter; i 2018 og igen i 2022. I 2018 blev der indsamlet og oprenset udbyttedata fra 2016 og 2017, og i 2022 blev der indsamlet og oprenset udbyttedata fra 2016-2021. Efter oprensning var der i alt 2938 ha (287 marker) vinterhvede fra i alt 19 bedrifter fra 2016-2021.

Oprensningsprocessen består af følgende trin:

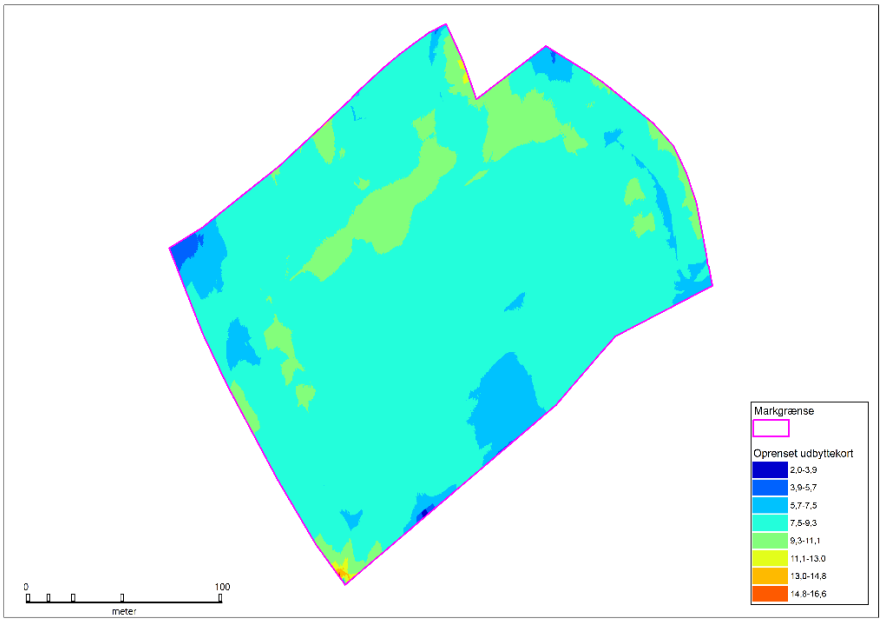
- Sortering af datapunkter i en mark efter tidsstempel
- Udbytteregistreringer udenfor intervallet 10-250 hkg pr. ha fjernes
- Statistiske afvigelser fjernes ved brug af distance-to-yield ratio (-log(distance/yield)
- Et glidende gennemsnit beregnes og datapunkter med udbyttemålinger > standardafvigelse x 2,5 fjernes. Dette gentages fire gange.

Udbyttedata blev herefter interpoleret ved brug af algoritmen Inverse distance (IDW) i python (library GDAL) og normaliseret (se figur 1), hvilket betyder, at der for hver 10 x 10 meter i marken er estimeret et udbytte i det pågældende år.

Råt udbyttedata (punktdata) fra mejetærsker



Oprenset og interpoleret udbyttedata



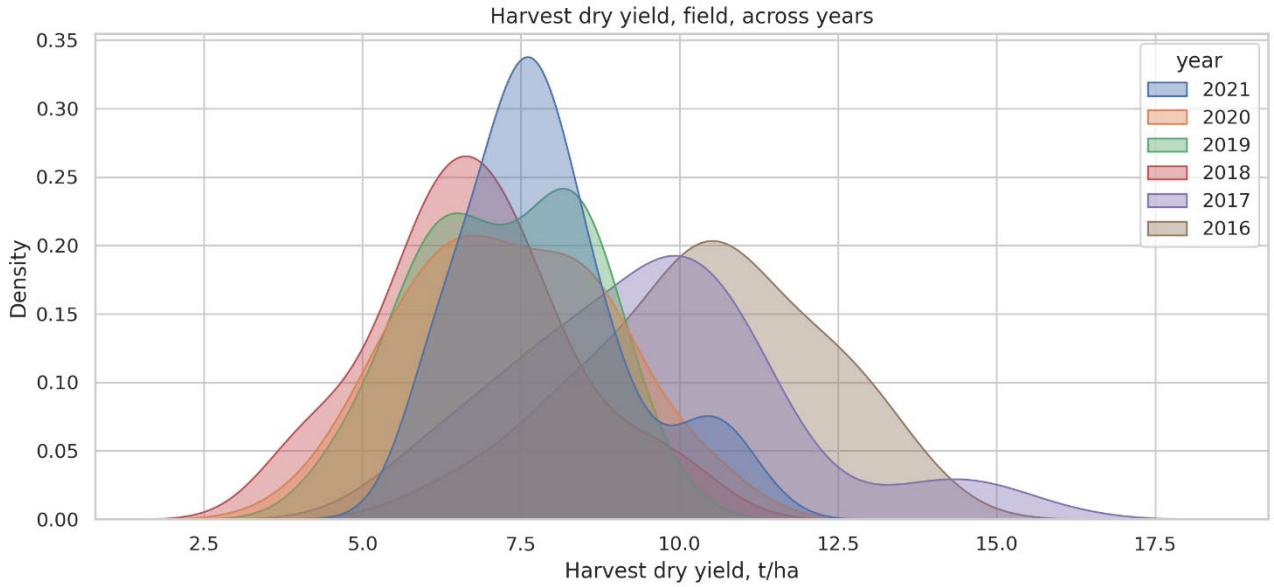
Figur 1. Oprensning og interpolering af råt udbyttedata. Øverst ses rå udbyttedata fra en mejetærsker (punktdata) og nederst udbyttedata oprenset og interpoleret. Når punktdata interpoleres, betyder det, at der overalt i marken er estimeret et udbytte ud fra punktdata. Billedet til højre er derudover klassificeret i 8 udbytteniveauer.

Tabel 1 og figur 2 viser fordelingen af udbyttedata mellem årene. Mængden af udbyttedata er størst i 2017 og 2021, og der er mindst data fra høstårene 2019 og 2020. Det gennemsnitlige udbytte er lavest i 2018, hvor tørken generelt påvirkede høstresultaterne i store dele af landet. På figur 2 ses det, at fordelingen af udbytterne i 2016 og 2017 adskiller sig fra de andre år med højere udbytteniveauer for en stor del af markerne.

Tabel 1. Udbyttedata fra mejetærskere fra 2016-2021 fordelt på år med antal marker, antal bedrifter som har bidraget med udbyttedata, samlet areal (ha), antal pixels af 10 x 10 meter samt det gennemsnitlige udbytte (hkg pr. ha) med standardafvigelsen (SD) i parentes.

År	Udbyttedata fra mejetærskere				
	Antal marker	Antal bedrifter	Areal, ha	Antal Pixel[1]	Gennemsnitlige udbytte, hkg pr. ha
2016[2]	33	7	289	28,898	104 (18)
2017[2]	95	15	856	22,491	98 (22)
2018	35	6	356	35,580	68 (15)
2019	26	5	221	22,062	72 (13)
2020	29	4	233	23,322	73 (16)
2021	69	5	984	98,356	79 (13)
Sum:	**287**		**2,938**	**293,829**	

1) Pixel på 10 x 10 m.
2) 69 % af udbyttedataet fra 2016 og 74 % af udbyttedataet fra 2017 er indsamlet i 2018, mens resten er indsamlet i 2022.



Figur 2. Fordelingen af udbyttedata på markniveau de enkelte høstår fra 2016-2021.

Satellitdata

I analysen er der anvendt satellitdata fra Sentinel 2 (L1C) som er downloaded via en service fra Sentinel-Hub. Sentinel 2 består af to satellitter (S2A og S2B), som levere data fra 13 spektrale bånd, der alle er anvendt i udviklingen af udbytteprognosen (Se tabel 2). Der er anvendt satellitbilleder fra 9. marts til 27. juli i årene 2016-2021. Satellitbilleder med skyer er blevet fjernet ved brug af S2_cloudless algoritmen, som er indstillet til at fjerne satellitbilleder med > 70% sandsynlighed for skyer (Sinergise Laboratory for geographical information systems, 2022).

Tabel 2. Spektrale bånd tilgængelig fra Sentinel 2 (S2A og S2B) samt båndbredde (nm) og opløselighed (m). Bånd markeret med orange indgår i beregningen af vegetationsindeksene NDVI, NDRE og MSAVI2.

	S2A		S2B		
Bånd nummer	**Centrale bølgelængde (nm)**	**båndbredde (nm)**	**Centrale bølgelængde (nm)**	**Båndbredde (nm)**	**Spatial resolution (m)**
B01	442.7	21	442.2	21	60
B02	492.4	66	492.1	66	10
B03	559.8	36	559	36	10
B04	664.6	31	664.9	31	10
B05	704.1	15	703.8	16	20
B06	740.5	15	739.1	15	20
B07	782.8	20	779.7	20	20
B08	832.8	106	832.9	106	10
B08a	864.7	21	864	22	20
B09	945.1	20	943.2	21	60
B10	1373.5	31	1376.9	30	60
B11	1613.7	91	1610.4	94	20
12	2202.4	175	2185.7	185	20

Ud fra bånd B04, B05 og B08 er vegetationsindeksene NDVI, NDRE og MSAVI2 udregnet.

(1)
$$NDVI = \frac{(B08 - B04)}{(B08 + B04)}$$

(2)
$$NDRE = \frac{B08 - B05}{B05 + B05}$$

(3)
$$MSAVI2 = \frac{2 \cdot B08 + 1 - \sqrt{(2 \cdot B08 + 1)^2 - 8 \cdot (B08 - B04)}}{2}$$

For at have satellitdata gennem hele vækstsæsonen i 2016-2021 er data lineær interpoleret gennem tid, og satellitdata (13 spektrale bånd + NDVI, NDRE og MSAVI2) for hver 14 dag fra 9. marts til 27. juli indgår i analysen. Derudover er den relative ændring fra 9. marts beregnet for alle variablerne. I alt indgår der 336 variabler fra satellitdata i analysen.

Den danske højde model

Den danske højdemodel (DEM) beskriver terrænets højde i forhold til det gennemsnitlige havniveau (opløsning på 0,4 meter) (Styrelsen for Dataforsyning og infrastruktur, 2022). I analysen er der for hver 10 x 10 meter pixel i de 287 marker beregnet den relative højde i punktet i forhold til det laveste punkt i marken. Derudover indgår hældningen (fra 0-90 grader samt procentvise hældning) og orienteringen af hældningen (0° = nord, 90 °= øst, 180° = syd og 270° = vest) i hver 10 x 10 meter pixel i marken. Der indgår derfor 5 variabler til analysen fra højdemodellen. Når observationerne aggregeres til markniveau, anvendes kun højden i forhold til havniveauet.

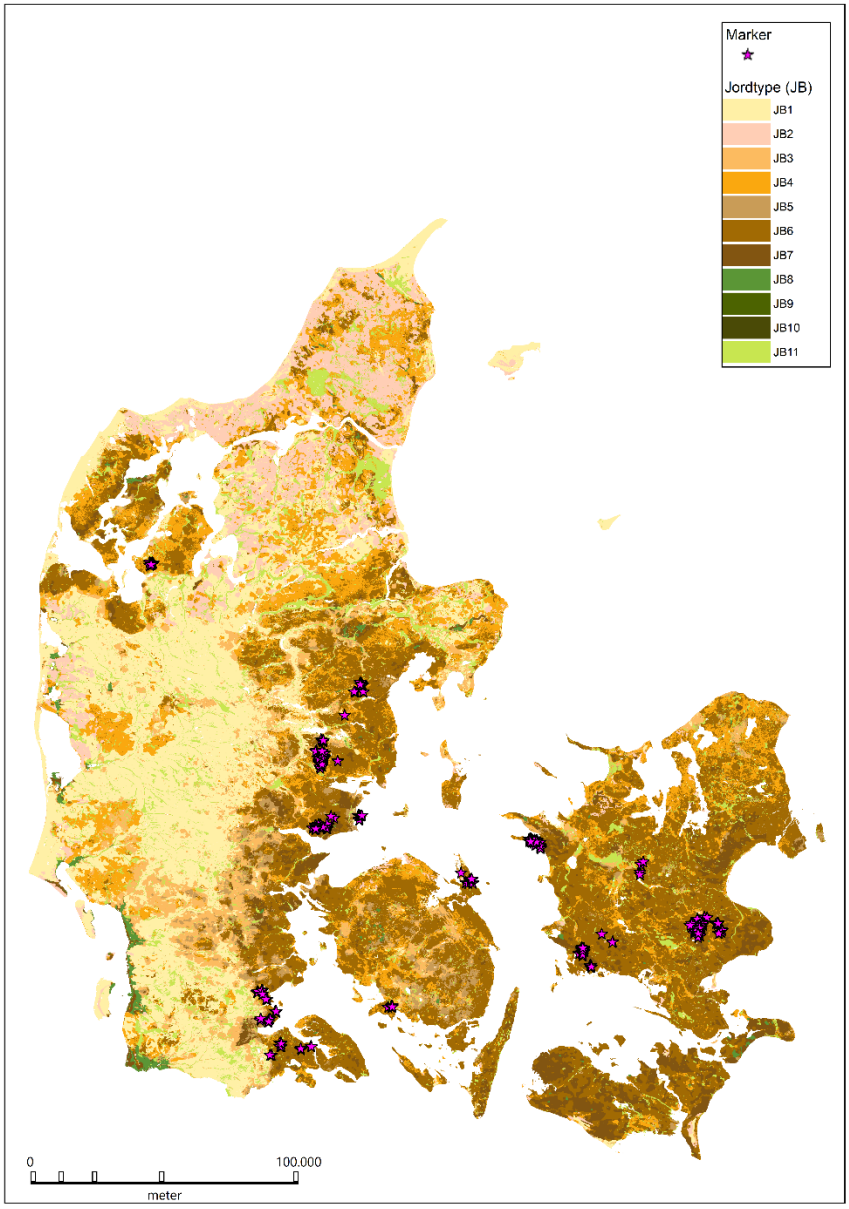
Vejrdata

Vejrdata som luft- og jordtemperatur, nedbør og indstråling er hentet fra DMI i en opløsning på 10 x 10 km og summeret for hver 14 dag (Danmarks Meteorologiske Institut, 2022). Derfor anvendes den samme måling for f.eks. nedbør i hver 10 x 10 meter pixel i marken og for marker, som ligger indenfor samme zone. For alle vejrparametre er der for hver 14. dag beregnet: gennemsnit, standard afvigelse (SD), minimum, maksimum. I alt indgår der 440 variabler fra vejrdata i analysen frem til 27. juli.

Jordtype

For hver af de 287 marker er den dominerende jordtype (JB) på marken anvendt som variable til udvikling af modellen. Det betyder, at alle pixels indenfor en mark har samme JB-nummer. Tabel 3 og figur 3 viser fordelingen af marker på de forskellige jordtyper fra 2016-2021. Det ses, at udbytteprognosen primært er udviklet med data fra JB4-7, mens sandjorde (JB1-3), svær lerjorde (JB8), meget svær lerjorde (JB9), siltjorde (JB10) og humusjorde (JB11) er underrepræsenteret eller helt mangler i datasættet.

Tabel 3 og Figur 3. For hver af de 287 marker fra 2016-2021 ses fordelingen af markerne på JB2-JB8 i de enkelte år.



	År					
JB	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
1	-	-	-	-	-	-
2	-	1	-	-	-	-
3	1	3	-	1	1	-
4	6	15	11	9	8	11
5	1	9	-	-	1	1
6	11	44	19	7	12	40
7	14	23	5	9	7	16
8	-	-	-	-	-	1
9	-	-	-	-	-	-
10	-	-	-	-	-	-
11	-	-	-	-	-	-
12	-	-	-	-	-	-

Vinterhvedesort

I Dansk Markdatabase (MarkOnline) er indhentet informationer om vinterhvedesorten på de enkelte marker. Ud af de 287 marker er sorten registreret på 163 marker (se tabel 4). Sytten forskellige sorter er registreret fra 2016-2021 (Benchmark, Chevignon, Creator, Graham, Informer, Kalmar, Kvium, KWS Cleveland, KWS Dacanto, KWS Lili, KWS Scimitar, Mariboss, Ohio, Pistoria, Sheriff, Substance og Torp). Derudover er der anvendt en sortblanding på i alt 9 marker fra 2017 og 2021, hvor blandingen anvendt er ukendt, hvilket betyder at kategorien kan dække over forskellige blandinger. Ingen af de 17 sorter er repræsenteret i alle høstår. Der indgår én kategorisk variable fra markdatabasen i analysen.

Tabel 4. Sorter registreret i vinterhvedemarkerne anvendt til udvikling af udbytteprognosen i vinterhvede fordelt på år.

Vinterhvedesorter	År						
	2016	2017	2018	2019	2020	2021	Total
Torp	4	48	15	2	-	-	69
Benchmark	2	14	1	2	-	-	19
Graham	-	-	-	1	3	7	11
Informer	-	-	-	-	3	8	11
Sortsblanding	-	4	-	-	-	5	9
KWS Lili	-	3	1	-	3	-	7
Creator	2	4	-	-	-	-	6
Pistoria	-	6	-	-	-	-	6
Sheriff	-	-	2	3	-	-	5
Kalmar	-	-	1	3	-	-	4
KWS Dacanto	4	-	-	-	-	-	4
Chevignon	-	-	-	-	3	-	3
Kvium	-	-	-	-	2	-	2
Mariboss	2	-	-	-	-	-	2
Substance	-	2	-	-	-	-	2
KWS Cleveland	1	-	-	-	-	-	1
KWS Scimitar	-	-	-	-	1	-	1
Ohio	-	-	-	1	-	-	1
Ingen sortregistrering	18	14	15	14	14	49	124

Markpolygoner og sædskifte

Landbrugsstyrelsen udarbejder hvert år et korttema over, hvilke afgrøder der dyrkes på de pågældende marker det enkelte år på baggrund af Enkeltbetalingsansøgningerne. Markpolygoner med afgrødetypen ”Vinterhvede” og ”Vinterhvede, brødhvede” er anvendt i analysen, og oplysninger om hovedafgrøden 5 år tilbage er ligeledes hentet fra disse korttemaer for de 287 marker. Der indgår derfor seks variabler fra Landbrugsstyrelsen data til analysen, da afgrødetypen slås sammen til én variable (vinterhvede).

Lokalitet

Den geografiske placering (Northing og Easting koordinater) af hver pixel indgår også i datasættet og tilfører 2 variabler til analysen.

Machine learning model

Machine learnings (ML) algoritmen Gradient Boosting er anvendt til udbytteprognosen i vinterhvede (Prokhorenkova et al., 2019). Algoritmen anvender binære beslutningstræer og virker ved at kombinere forskellige beslutningstræer til én model.

Prognosedatoer
I analysen testes flere modellers evne til at prædiktere udbyttet i vinterhvede på forskellige tidspunkter gennem vækstsæsonen. I vinterhvede kan landmanden vælge at regulere anden kvælstoftildelingen medio april (vækststadium 30-32) og/eller tredje kvælstoftildeling start/medio maj (vækststadium 37), hvorfor modellernes evne til at prædiktere udbyttet testes 6. april og 4. maj. Derudover testes forudsigelser 1. juni og lige før høst 27. juli i denne analyse.

Validering
I databehandlingen deles datasættet i to, hvor størstedelen (ofte 85%) anvendes til at træne modellen, og en mindre del af data (ofte 15%) anvendes til at validere modellen. Den gennemsnitlige absolutte fejl (MAE), root mean square error (RMSE), R^2 og standardafvigelsen af den absolutte fejl (SD af AE) bruges som mål for nøjagtigheden af udbytteprædiktionen:

(1) $$MAE = \sum_{i=1}^{n} \frac{|h_i - p_i|}{n}$$

Hvor *h* er det målte udbytte registreret med mejetærskeren (oprenset og interpoleret udbytte) i punktet, *p* er det estimerede udbytte (prognosemodellen) og *n* er antallet af observationer i datasættet.

(2) $$RMSE = \sqrt{(p-h)^2}$$

Hvor *h er* det målte udbytte med mejetærskeren i punktet og *p* er det estimerede udbytte.

Hver gang udbytteprognosemodellen ændres eller der tilføjes nyt data, trænes modellen på ny og den nye model valideres igen. Dette kaldes ”eksperimenter”. For hvert eksperiment vurderes det, hvor godt modellen præsterer (MAE, RMSE, R^2 og SD af AE) i forhold til det målte udbytte fra mejetærskerne.

Eksperimenter
Der er fortaget forskellige eksperimenter i analysen for at udvikle den bedste model til at forudsige udbyttet i vinterhvede. Se tabel 5. Prognosemodellerne testes ofte på både pixel- og markniveau, da begge niveauer er interessante inden for præcisionsjordbrug. I eksperiment 1-3 og 7-8 er forudsigelserne for hver pixel summeret op på markniveau.

I eksperiment 1 forudsiges udbyttet 27. juli. Modellen anvender alle 791 variabler, og datasættet består af 293.829 observationer af 10 x 10 meter pixels, som er delt i et 85/15 split, hvor 85% af data anvendes til at træne modellen på, og 15% anvendes til at validere modellen på. Split af data er randomiseret, hvilket betyder at data fra en mark kan være anvendt i både træning og validering af modellen.

I eksperiment 2 forudsiges udbyttet 6. april, 4. maj, 1. juni og 27. juli. Data er nu delt på markniveau, så data fra en mark nu enten er i trænings- eller valideringsdatasættet.

I eksperiment 3 testes betydningen af udvalgte data (vejrdata, terrænhøjde og satellitdata m.m.) ved at fjerne variablerne fra modellen. Derudover testes betydningen af at fjerne variabler fra datasættet, som er tæt korreleret med andre variabler eller variabler med lav forklaringsgrad for forudsigelsen af udbyttet. Dette gøres for at mindske støj, og for at optimere modellernes performance.

I eksperiment 4 aggregeres data til markniveau inden træning og validering, hvilket betyder at antallet af observationer er gået fra 293.829 til 287.

I eksperiment 5 krydsvalideres modellerne med år (data splittes på år). Det betyder, at data fra ét høstår tages ud af træningsdata, og i stedet anvendes til at validerer modellen. Dette gøres for alle høstårene, hvor prædiktionerne bliver gemt, og senere kombineret i et vægtet gennemsnit. Udbyttet forudsiges 4. maj og 27. juli.

I eksperiment 6 fjernes data indsamlet i 2018 fra datasættet, og der anvendes kun data indsamlet i 2022 (195 marker). Ellers er eksperimentet identisk med eksperiment 5.

Eksperiment 7 er identisk med eksperiment 5 ud over, at modellen trænes og valideres på pixelniveau i stedet for markniveau.

Eksperiment 8 er identisk med eksperiment 6 ud over, at modellen trænes og valideres på pixelniveau i stedet for markniveau.

Tabel 5. Viser hvordan eksperiment 1-8 varierer indenfor prognosedato, antal variabler, antal observationer samt split af data mellem træning- og valideringsdatasæt.

Eksperiment	Prognosedato	Variabler	Observationer	Split af data
1	27. juli	Alle	293.829 pixel	Randomiseret (15 % i valideringsdatasættet)
2	6. april	Alle	293.829 pixel	markniveau (ca. 40 marker i valideringsdatasættet)
	4. maj			
	1. juni			
	27. juli			
3	27. juli	Eliminering af variabler	293.829 pixel	markniveau (ca. 40 marker i valideringsdatasættet)
		DMI		
		DTM		
		Satellit		
4	27. juli	Aggregeret til markniveau + Alle	287 marker	markniveau (ca. 40 marker i valideringsdatasættet)
		Aggregeret til markniveau + Eliminering af variabler		
5	4. maj	Aggregeret til markniveau - + Eliminering af variabler	287 marker	Krydsvalidering med år
	27. juli			
6	4. maj	Aggregeret til markniveau - + Eliminering af variabler	195 marker	Krydsvalidering med år (kun inkluderet data indsamlet i 2022)
	27. juli			
7	4. maj	Eliminering af variabler	293.829 pixel	Krydsvalidering med år
	27. juli			
8	4. maj	Eliminering af variabler	220.644 pixel	Krydsvalidering med år (kun inkluderet data indsamlet i 2022)
	27. juli			

1.4 Resultater og diskussion

Der er udført 8 eksperimenter på datasættet for at finde frem til den bedste udbytteprognosemodel til vinterhvede. Tabel 6 viser nøjagtigheden (MAE, RMSE, SD af AE) af forudsigelserne i eksperiment 1-8 på valideringsdatasættet i hkg vinterhvede pr. ha på mark- og pixelniveau samt R^2.

Eksperiment 1: I eksperiment 1 kan pixel fra samme mark indgå i både træning og validering af udbytteprognosemodellen, hvilket resulterer i en MAE på 0,6 hkg pr. ha på markniveau 27. juli og en R^2 tæt på 1. Tilfældige deling af data i henholdsvis trænings- eller valideringsdata vurderes ikke at give et realistisk billede af prognosemodellens prædiktionsevne i en ubekendt mark. Denne praksis er dog tidligere set, hvorfor testen er udført til sammenligning med de resterende eksperimenter.

Eksperiment 2: I eksperiment 2 forudsiges udbyttet 6. april, 4. maj, 1. juni og 27. juli. Modsat eksperiment 1 er data delt på markniveau, hvilket betyder, at alle pixel i en mark enten er i trænings- eller valideringsdatasættet. Dette resulterer i en MAE på henholdsvis 6,7, 6,2, 5,9 og 5,6 hkg pr. ha på markniveau 6. april, 4. maj, 1. juni og 27. juli. På pixelniveau ses en MAE på 9,7-9,2 hkg pr. ha fra 6. april til 27. juli.
Der observeres en markant stigning i MAE, når modellen ikke er blevet trænet på data fra samme mark, som den forudsiger udbyttet i, hvilket er forventelig. MAE falder jo tættere på høst, udbyttet forudsiges, hvilket bl.a. kan forklares med, at datagrundlaget for prognosemodellen stiger med vækstsæsonen (satellit- og vejrdata), og at der dermed er færre ubekendte data i den resterende vækstsæson (f.eks. ubekendte kommende vejrforhold). Fra 6. april til 27. juli falder MAE med 1,1 hkg pr. ha. Resultaterne viser ligeledes, at MAE stiger, når modellen skal forudsige udbyttet på pixelniveau (positionsbestemt) i forhold til, når prognoserne summeres på markniveau.

Eksperiment 3: I eksperiment 3 testes bl.a. effekten af at fjerne enten vejrdata (DMI), terrænhøjdedata (DTM) eller satellitdata fra prognosemodellen. Hvis resultaterne sammenholdes med eksperiment 2 den 27. juni, så ses det, at alle variabler fra disse udvalgte datagrupper har betydning for prognosemodellen. Særligt vejrdata og satellitdata er vigtige for modellen. MAE stiger til 5,7-7,4 hkg pr. ha 27. juni, hvis udvalgte data fjernes fra modellen.
En del af de 791 variabler i eksperiment 2 er tæt korreleret eller har en lav forklaringsgrad i prognosemodellen. I eksperiment 3 er de mindst vigtige variabler i modellen fjernet, hvilket resulterede i en MAE på 4,9 hkg pr. ha 27. juli. Modellen performer en del bedre, når antallet af variabler i modellen reduceres, hvilket formentlig skyldes, at støjen fra de mange datakilder mindskes. Senere i rapporten vises de vigtigste variabler for de bedste prognosemodeller.

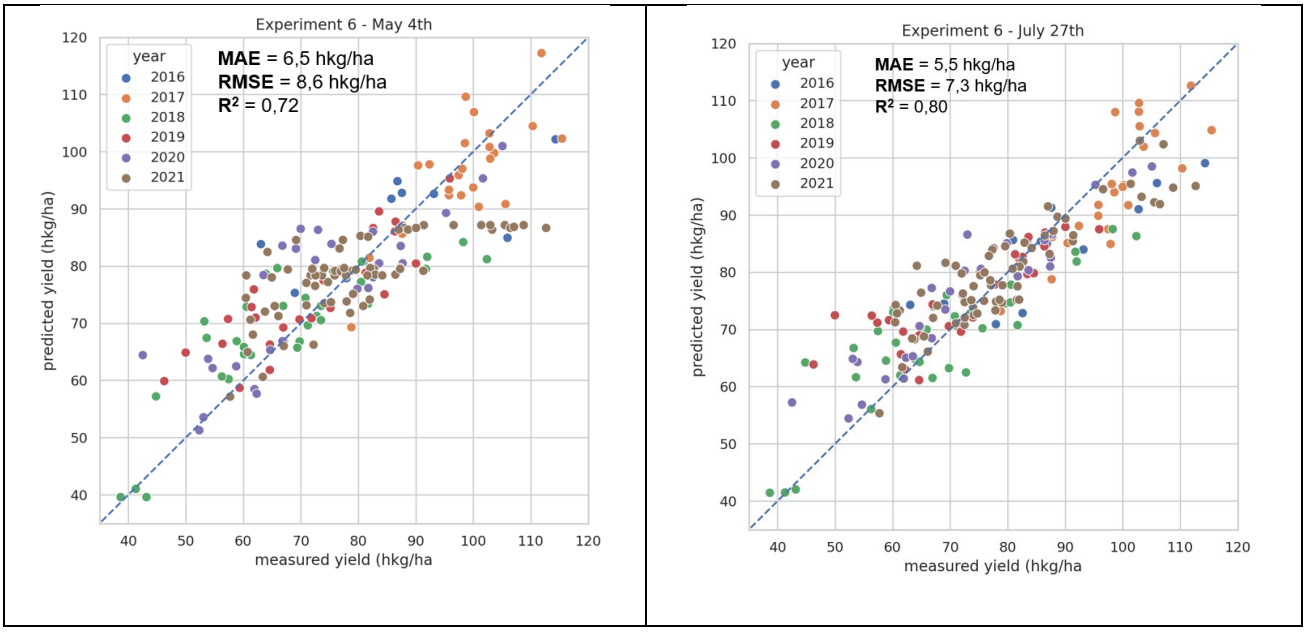
Eksperiment 4: i eksperiment 4 testes effekten af at aggregere observationer til markniveau inden træning og validering af modellen. Dette resulterer i, at MAE falder til 4,1 hkg pr. ha 27. juli, når de mindst vigtige variabler samtidig fjernes fra prognosemodellen.
I prognosemodellerne i eksperiment 1-4 kan data fra alle år være i både trænings- og valideringsdatasættet, hvilket kan give en bias. Selvom variablen ”år” ikke er inkluderet i datasættet, kan andre variabler som f.eks. vejrdata eller satellitdata fungere som proxy for ”år”. Herved kan prognosemodellerne indirekte få ”kendskab” til udbytteniveauet de pågældende år.

Tabel 6. Viser nøjagtigheden af udbytteprædiktionen (MAE, RMSE, R^2 og SD af AE) i eksperiment 1-8 på markniveau og på pixelniveau (10 x 10 meter). Prognosemodellerne i vinterhvede er udviklet på træningsdatasættet, og præcisionen (MAE, RMSE, R^2 og SD af AE) af modellerne vurderes på valideringsdatasættet.

Eksperiment	Prognosedato	Variabler	MAE, hkg/h		RMSE		R^2		SD af AE	
			Validering		Validering		Validering		Validering	
			Mark	Pixel	Mark	Pixel	Mark	Pixel	Mark	Pixel
1	27. juli	Alle	**0,6**	4,1	**0,9**	6,2	**1,00**	0,92	**0,7**	4,7
2	6. april	Alle	**6,7**	9,7	**9,3**	12,6	**0,74**	0,56	**6,5**	8,1
	4. maj		**6,2**	9,4	**8,5**	12,2	**0,79**	0,59	**5,8**	7,7
	1. juni		**5,9**	9,3	**7,9**	12,2	**0,82**	0,59	**5,3**	7,8
	27. juli		**5,6**	9,2	**7,5**	11,9	**0,83**	0,61	**5,0**	7,6
3	27. juli	Eliminering af variabler	**4,9**	8,5	**7,1**	11,2	**0,85**	0,65	**5,1**	7,3
		DMI	**7,4**	9,6	**10,0**	12,7	**0,71**	0,56	**6,7**	8,2
		DTM	**5,7**	9,2	**7,7**	11,9	**0,83**	0,61	**5,1**	7,6
		Satellit	**5,9**	9,4	**8,2**	12,2	**0,80**	0,59	**5,8**	7,8
4	27. juli	Aggregeret til markniveau + Alle	**5,4**		**7,1**		**0,85**		**4,7**	
		Aggregeret til markniveau + Eliminering af variabler	**4,1**		**5,4**		**0,91**		**3,6**	
5	4. maj	Aggregeret til markniveau + Eliminering af variabler	**9,0**		**11,8**		**0,69**		**7,6**	
	27. juli		**8,8**		**11,9**		**0,68**		**8,0**	
6	4. maj	Aggregeret til markniveau + Eliminering af variabler	**6,5**		**8,6**		**0,72**		**5,6**	
	27. juli		**5,5**		**7,3**		**0,80**		**4,7**	
7	4. maj	Eliminering af variabler	**10,2**	14,0	**13,0**	18,6	**0,65**	0,45	**7,3**	11,7
	27. juli		**9,4**	13,6	**12,1**	18,1	**0,68**	0,47	**7,6**	11,8
8	4. maj	Eliminering af variabler	**7,1**	10,8	**9,4**	14,2	**0,66**	0,41	**6,2**	9,2
	27. juli		**6,8**	10,7	**9,1**	14,1	**0,68**	0,42	**6,1**	9,2

Eksperiment 5: i eksperiment 5 krydsvalideres med år. Resultaterne viser at MAE stiger til 9,0 og 8,8 hkg pr. ha 4. maj og 27. juli.
Eksperimentet er den ultimative test for udbytteprognosen i vinterhvede og illustrerer, hvordan modellen performer under realistiske forhold, hvor modellen beregner en prognose et år, hvor den ikke har kendskab til udbytteniveauet.

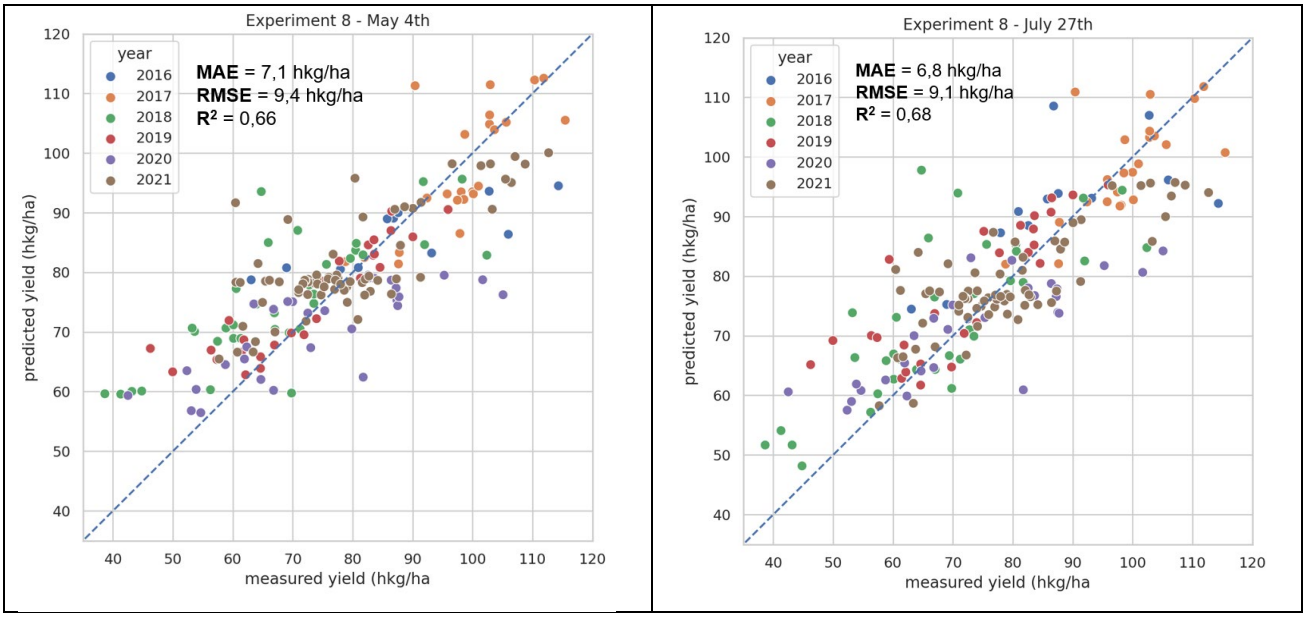
Eksperiment 6: Eksperiment 6 er identisk med eksperiment 5, men anvender kun data indsamlet i 2022. Figur 3 viser det estimerede udbytte som funktion af det målte udbytte (markniveau) fra 2016-2021. Det ses, at MAE falder til 6,5 og 5,5 hkg pr. ha på markniveau henholdsvis 4. maj og 27. juli, når udbyttedata indsamlet i 2018 fjernes fra modellen.
Det kan ikke forklares, hvorfor udbyttedata indsamlet i 2018 adskiller sig fra udbyttedata indsamlet i 2022. Det kan dog ikke udelukkes, at metoderne anvendt til at ekstrahere og oprense data kan afvige mellem de to indsamlingstidspunkter på trods af standardisering af processerne. Derfor vurderes resultaterne fra eksperiment 6 at være retvisende for, hvor godt udbytteprognosen forventes at kunne ramme udbyttet i vinterhvede på markniveau med ovennævnte datagrundlag.



Figur 3. viser det estimerede udbytte (hkg pr. ha) i eksperiment 6 som funktion af det målte udbytte (hkg pr. ha) fra 2016-2021 på markniveau (195 observationer) henholdsvis 4. maj (til venstre) og 27. juni (til højre). Figurerne inkluderer MAE, RMSE og R^2.

Eksperiment 7: Eksperiment 7 er identisk med eksperiment 5, men prognoserne trænes og valideres på pixelniveau. Resultaterne viser en MAE på 10,2 og 9,4 hkg pr. ha på markniveua henholdsvis 4. maj og 27. juni.
Når modellerne trænes og valideres på pixelniveau, bliver modellen dårligere til at forudsige udbyttet i vinterhvede i forhold til prædiktion på markniveau på trods af at datagrundlaget er markant større (293.829 observationer 27. juli). Dette indikerer at det store antal observationer på hver 10 x 10 meter også skaber en del støj, hvilket reflekteres i de højere MAE på de udvalgte prognosedatoer.

Eksperiment 8: Eksperiment 8 er identisk med eksperiment 6, men prognoserne beregnes på pixelniveau. Resultaterne viser en MAE på 7,1 og 6,8 hkg pr. ha på markniveua henholdsvis 4. maj og 27. juni. Hvis modellen skal forudsige udbyttet inden for hver 10 x 10 meter af en mark, stiger MAE til 10,8 og 10,7 hkg pr. ha 4. maj og 27. juni.



Figur 4. Viser det estimerede udbytte (hkg pr. ha) i eksperiment 8 som funktion af det målte udbytte (hkg pr. ha) fra 2016-2021 på markniveau henholdsvis 4. maj (til venstre) og 27. juni (til højre) (Datagrundlaget for eksperiment 8 er på 220.644 observationer frem til 27. juni, men i disse figurer er det estimerede udbytte vist på markniveau (195 observationer). Figurerne inkluderer MAE, RMSE og R2.

Udbytteprognosemodellen fra eksperiment 6 har vist en MAE, som er lav nok til, at modellen på markniveau kan anvendes til at regulere kvælstoftildelingen til vinterhvede i vækststadium 37 (start maj). Præcisionen kom ikke ned på niveauet fundet i vårbyg i studiet af Petersen et al., 2023, men her har de udviklet en model på én mark ud fra 9 års data modsat denne analyse, hvor udbytteprognosen 4. maj bygger på 195 marker.
Hvis modellen skal anvendes af landmænd, er det dog vigtigt at nævne de forhold, hvor prognosen skal anvendes med forsigtighed. F.eks. er data fra sandjord ikke repræsenteret i datasættet, og der generelt mangler data fra Vest- og Nordjylland. Derudover er modellen udviklet på data fra vækstsæsonerne 2016-2021, hvor nogle år er bedre repræsenteret end andre. Nøjagtigheden af prædiktionen forventes derfor ringere under forhold som afviger fra datagrundlaget i modellen.

Vigtige variabler for udbytteprædiktion i vinterhvede
I flere eksperimenter er der udført "feature eliminering", hvilket betyder at antallet af variabler i modellen er reduceret for at optimere modellerne yderligere. Figur 5 viser op af y-aksen de vigtigste variabler for modellerne i eksperiment 6 og 8 (for 4. maj og 27. juli). Det ses, at listen med vigtige variabler for de fire prognoser varierer, men mange variabler går igen i de fire prognosemodeller i eksperiment 6 og 8.

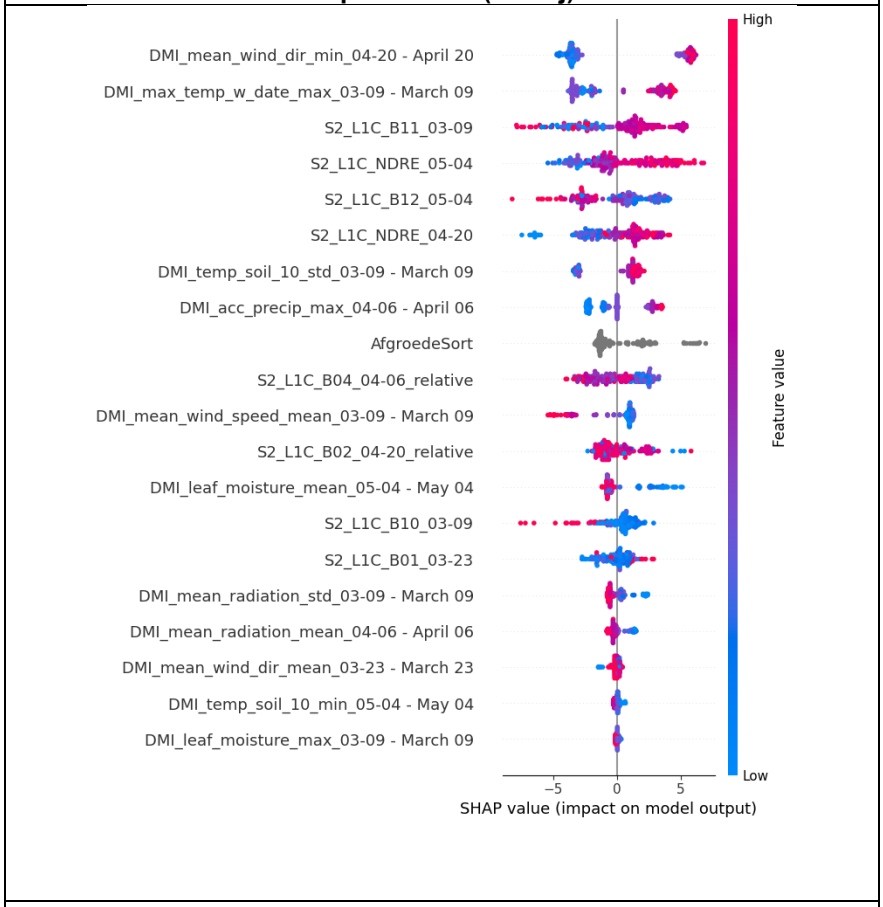
Figur 5 viser ligeledes SHAP-værdier (hkg pr. ha) som udtrykker, hvordan hver enkelt variabel påvirker det prædikterede udbytte for hver enkelt observation (mark) i datasættet. Farven viser om den pågældende variabel er høj (pink) eller lav (blå). Hvis man tager udgangspunkt i variablen "*S2_L1C_NDRE_04-20"* i eksperiment 8 den 4. maj, så medfører høje NDRE-værdier 20. april høje prædikteret udbytter og lave NDRE-værdier lave udbytter. Det ses også at de højeste udbytter observeres ved mere gennemsnitlige NDRE-værdier 20. april, hvilket kunne skyldes, at risikoen for lejesæd øges ved højere biomassemålinger sidst i april. Hvis alle SHAP-værdierne for alle variablerne summeres fås det gennemsnitlige prædikterede udbytte i datasættet.

Figur 5 viser, at vejrdata, satellitdata, markens lokalitet, sorten samt afgrøde på marken to år tidligere er vigtige datakilder for modellerne.

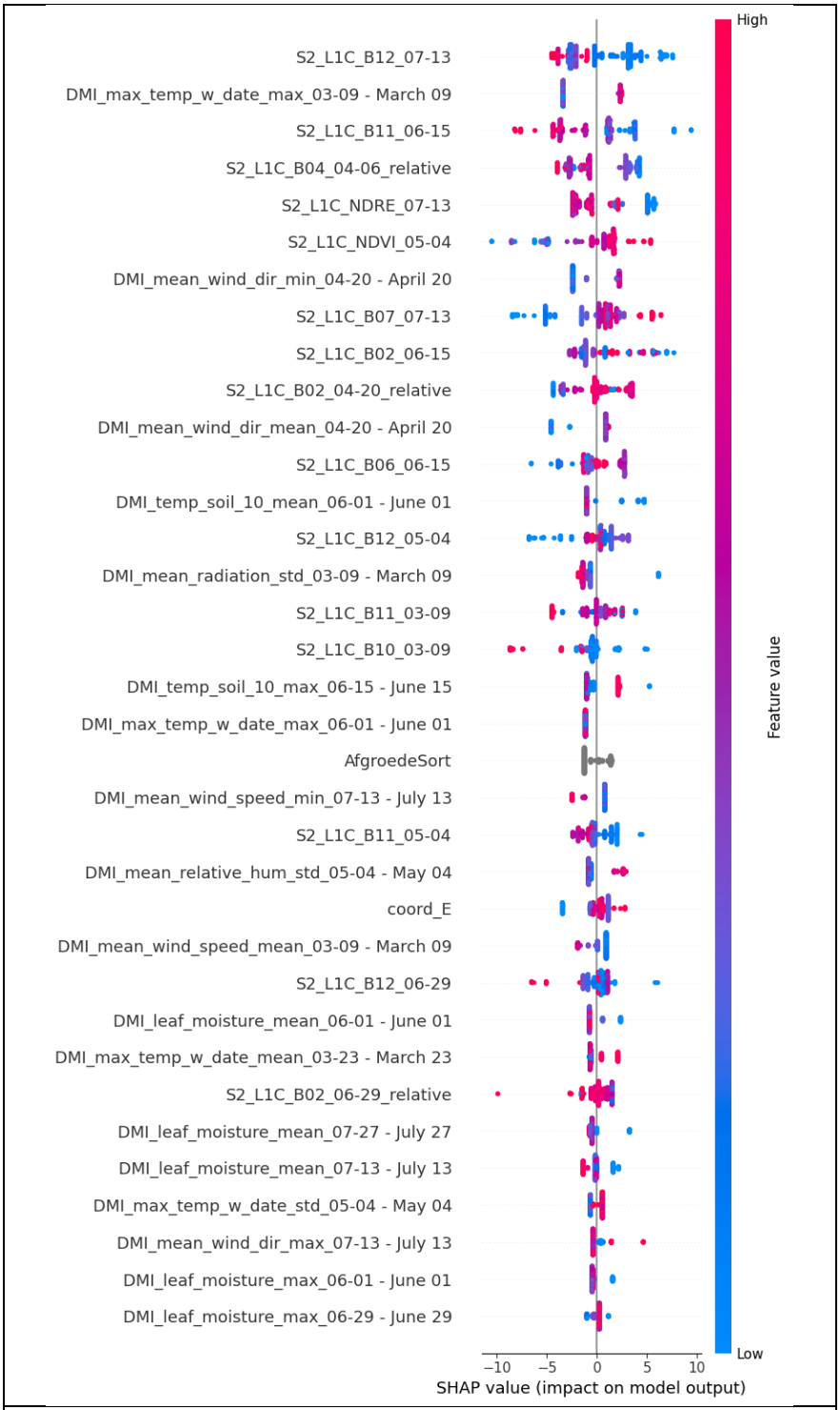
Ifølge figur 5 er variabler som f.eks. vejrparametrene ”minimum vind” og ”SD af jordtemperatur” 9. mart er vigtige for prognosemodellen i eksperiment 6 (4. maj). Det kan skyldes, at variablerne fungerer som en proxy for ”år”. Figur 5 kan derfor kun anvendes til at øge forklaringsgraden af vigtigheden af udvalgte variabler, samt hvordan værdierne påvirker det prædikteret udbytte.

Figur 5. Viser hvilke variabler der er vigtigst for det prædikteret udbytte i de fire modeller i eksperiment 6 og 8 (4. maj og 27. juli). Figurerne viser hvordan de forskellige variabler påvirker det prædikteret udbytte i hkg pr. ha (SHAP-værdier).

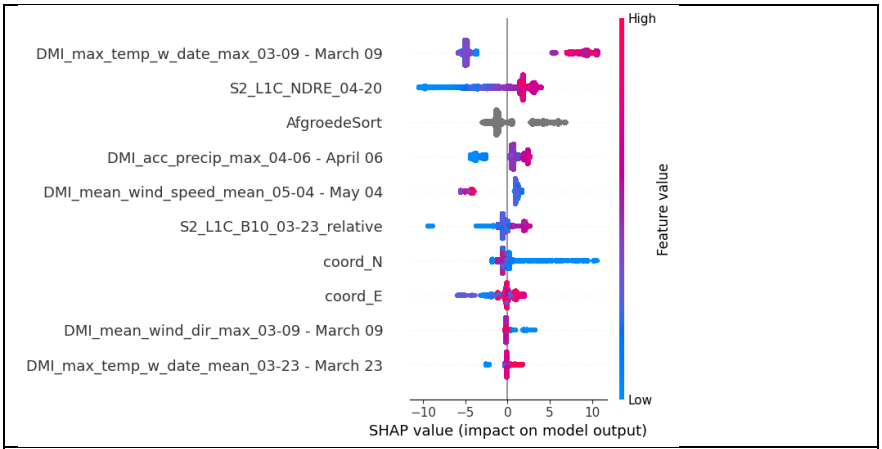
Experiment 6 (4. maj)



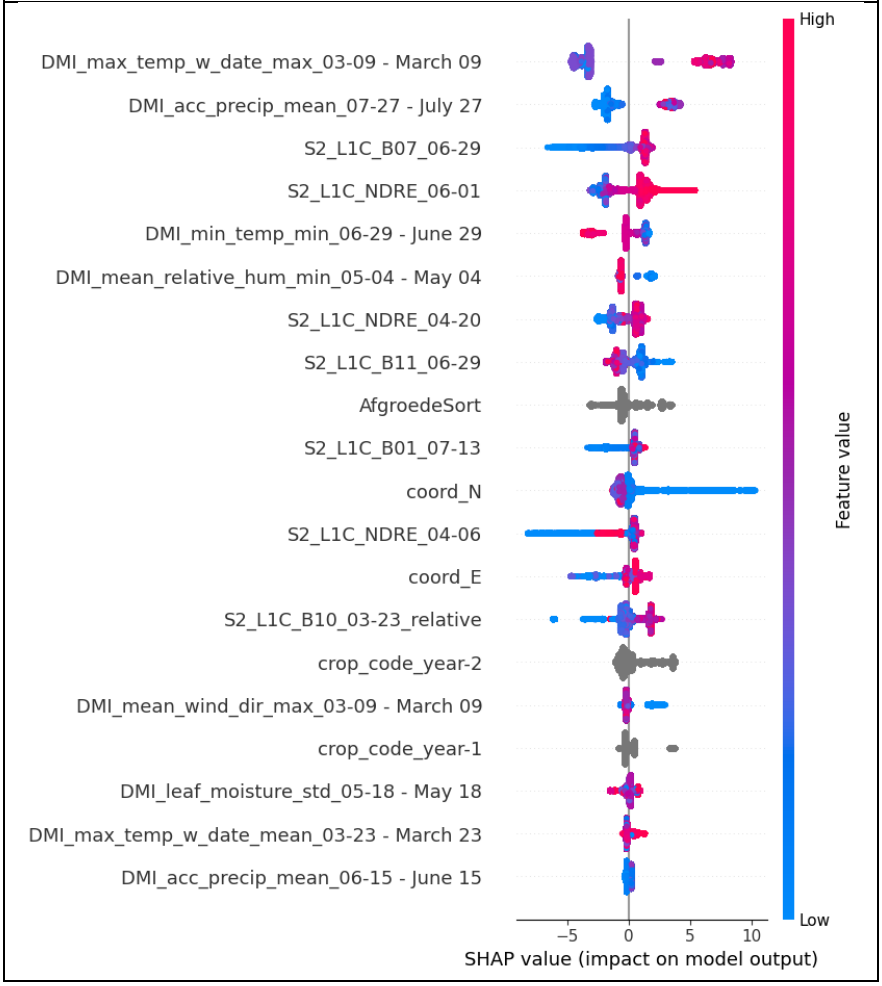
Eksperiment 6 (27. juli)



Experiment 8 (4. maj)



Experiment 8 (27. juli)



1.5 Konklusion

Indsamling af data fra seks vækstsæsoner samt tilføjelse af nye datakilder har forbedret grundlaget for udvikling af mere robuste prognosemodeller i vinterhvede. Resultaterne viser at:

- Det er muligt at forudsige udbyttet i vinterhvede på markniveau med en MAE på 6,5 og 5,5 hkg vinterhvede pr. ha henholdsvis 4. maj og 27. juli.

- Nøjagtigheden på udbytteprognosen 4. maj er god nok til, at landmanden kan anvende prognosen som en del af inputtet til at regulere kvælstoftildelingen i vinterhvede ved tredje tildeling start maj.

- Ved udbytteforudsigelse indenfor marken (for hver 10 x 10 meter) stiger MAE til 10,8 og 10,7 hkg pr. ha 4. maj og 27. juli sammenlignet med prædiktion på markniveau.

- Prognosen skal anvendes med forsigtighed på sandjord og i vækstsæsoner som afviger fra 2016-2021.

1.6 Referencer

Petersen, C. T., Langgaard, M. K., & Petersen, S. D. (2023). Yield prediction in spring barley from spectral reflectance and weather data using machine learning. Soil Use and Management, 00, 1–13. *https://doi.org/10.1111/sum.12902*

Sinergise Laboratory for geographical information systems, Ltd. (2022). Sentinel 2 data using the Sentinel-Hub services: https://sentinel-hub.com/, last accessed December 2022.

Styrelsen for Dataforsyning og infrastruktur (2022). Danmarks Højdemodel - Terræn. https://dataforsyningen.dk/data/930

Danmarks Meteorologiske Institut. (2022). Weather data: https://www.dmi.dk/frie-data/

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). CatBoost: unbiased boosting with categorical features. Non-peer reviewed pre-print at ARXIV.1706.09516

UDBYTTEFORUDSIGELSE I VINTERHVEDE VED BRUG AF MACHINE LEARNING

Udgivet af
SEGES Innovation P/S
Agro Food Park 15, Skejby
DK 8200 Aarhus N

Forfattere
Mette Kramer Langgaard, SEGES Innovation P/S
Lasse Rose Malskær, SEGES Innovation P/S
Jacob Hein, SEGES Innovation P/S

Kontakt
Mette Kramer Langgaard

M +45 6120 2701

Foto
Mette Kramer Langgaard, SEGES Innovation P/S

April 2023

STØTTET AF
Promilleafgiftsfonden for landbrug