Task definition

Diseases and Health in Cattle Herds

Exported on 11/01/2018

Table of Contents

1 Project attributes

Project name	Diseases and Health in Cattle Herds
Project case number	7699
Project task number	30
Project start date	01 Jan 2018
Project due date	31 Dec 2018
Project status	IN PROGRESS

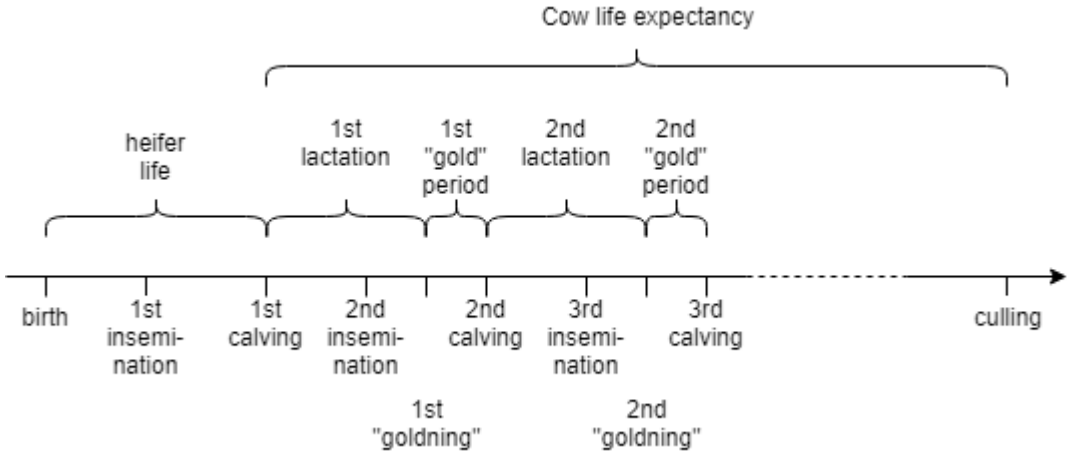
2 Stakeholders

Name	Organization
Jens Bligaard	SEGES
Johannes Frandsen	SEGES
Lars Arne Hjort Nielsen	SEGES
Peter Raundal	SEGES
Henrik Læssøe Martin	SEGES
Jørgen Nielsen	SEGES
Per Hejndorf	SEGES
Christian Schou Oxvig	SEGES
Peter Fogh	SEGES

3 Task description

In the Kvægdatabase (KvægDB), one of the largest agricultural databases, data related to health status and disease treatments of individual cattle herds have been collected for several years. Detailed registrations are required for herds that are part of a formalized healthcare process. Using machine learning, we expect to be able to analyze and predict which individuals and herds are most at risk of health problems. Likewise, it will be possible to analyze the treatment process, enabling us to say more about which treatments are the most optimal in the individual situation. By combining this data with information on survival, release time, lactation, etc., it will be possible to predict the cows and herds where treatment is most profitable. The goal is to describe a decision support tool that can propose the most optimal decision in each case.

This task deals with predicting some future quantity related to an individual cattle at some point the during its life course. More specifically, we attempt to predict the time to culling at the time of the first calving. This time interval, we denote "cow life expectancy". A sketch of the typical cattle life course is included below.



4 Tasks prioritization

4.1 Main tasks for the data science team

- **Data collection**:
  - The data science team must collect diseases and health-related data (according to the data described below) for at least 5,000 cattle.
- **Analyses**:
  - Based on the collected data, the data science team must train one or more machine learning algorithm(s) to predict the cow life expectancy of individual cattle, i.e. the time from first calving to the time of culling.
  - The data science team must use two different feature sets in training the machine learning algorithm(s); one feature set which includes weight features and one that does not.
- **Evaluation**:
  - Based on the analysis results, the data science team must evaluate the proposed models in order to assess their validity and performance. In particular, the data science team must provide a measure of the uncertainty of the predicted cow life expectancy.
- **Report**:
  - Based on the data collection, analysis results, and evaluation, the data science team must write an IMRAD report covering the results.

4.2 Prioritized additional tasks (if time permits)

1. The data science team may incorporate "besætnings_id "and "sundhedsforløbsmodul" as variables in the models to model potential differences in treatment practices."
2. The data science team may attempt to establish the earliest time in the cattle life cycle (earlier than first calving) at which one can predict the time of culling.
3. The data science team may train one or more machine learning algorithm(s) to predict the course of the culling of individual cattle.
4. The data science team may train one or more machine learning algorithm(s) to predict the life production of individual cattle (kg EKM or kg værdistof).
5. The data science team may include the additional health-related data in the prediction.
6. The data science team may utilize time-sequence machine learning models and data.
7. The data science team may include the additional data attributes from the KvægDWH.

4.3 Not doing

- The data science team does not include climate data in the prediction.
- The data science team does not build a recommendation system that automatically suggests the optimal treatment.
- The data science team does not consider profitability in the analyses.
- The data science team does not analyze diseases and treatments across herds - only for the individual cattle.
- The data science team does not train a machine learning algorithm to predict the age at first calving.
- The data science team does not train a machine learning algorithm to predict the age at first insemination.

4.4 Risks and blockers

- **The legislation regulating the requirements for diseases treatments registrations has changed significantly over time (in particular around 2006, 2010, and 2015) which may have resulted in highly inconsistent registration practices over time. Such inconsistent data quality may potentially make it impossible to train machine learning algorithms that generalize to a larger data set.**
- **The diseases treatments registrations practices differ significantly between farms. Such inconsistent data quality may potentially make it impossible to train machine learning algorithms that generalize to a larger data set.**

4.5 Time estimation

As seen in the table below, we expect that we will use all of the allocated hours to solve the task defined in this document.

Time	Hours
Allocated time	850
Scoping time (already spent)	300
Estimated time needed for project management (meetings, task refinements, etc.) (to be spent)	130
Estimated time needed for solving the task (to be spent)	420
Remaining time	0

4.6 Description of end product

This is a proof-of-concept of a machine learning system which predicts the life expectancy of individual cattle. If successful, such a machine learning system is to function as a decision support tool for a farmer or a veterinarian in selecting the optimal treatment of the individual cattle. It is expected that a final version of such a machine learning system is to be included in the Dairy Management System (DMS).

4.7 Current state-of-the-art

The iCull project was a GUDP financed (5.574.182 kr.) collaborative effort by Dianova, KU Sund, DTU Veterinærinstituttet, DTU Compute, and SEGES running from 01/11/2013 to 31/10/2015 [1]. The main contact at SEGES in the iCull project was Erik Rattenborg. Focus was on creating a support tool that selects cattle that should be culled in order to implement a financially optimal strategy for culling. Two main areas of interest were investigated in the project: 1. Simulation of culling of cattle in herds suffering from the disease paratuberculosis. 2. Prediction of the future value of cattle in terms of performance, cell count, and health status.

Various academic studies have focused on the use of machine learning for modelling and predicting diseases in cattle [2] [5], predicting insemination outcomes [3], and classifying cattle behavioral patterns [4]. See also [6], [7], [8].

4.7.1 Reference list

	Title	Author	Link
1	iCull – Optimeret udnyttelse af koens værdi før udsætning	Mogens Madsen, Dianova, et al	http://www.icull.dk/ http://orbit.dtu.dk/en/projects/icull(3801feee-66a8-4214-8fd3-daedf91eeb1a).html http://lbst.dk/tvaergaaende/gudp/gudp-projekter/2013/nyt-it-vaerktoej-skal-hjaelpe-landmanden-med-at-vurdere-syge-koeers-fremtid/
2	Embedding system dynamics in agent based models for complex adaptive systems	Teose et al.	https://dl.acm.org/citation.cfm?id=2283818 (PDF: http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/download/3304/3730)
3	Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms	Shahinfar et al.	https://ac.els-cdn.com/S0022030213008059/1-s2.0-S0022030213008059-main.pdf?_tid=e0253be8-6213-4331-9945-0e2d36e1f3ad&acdnat=1519985555_1d17216ab78ffa9c5ff5567f53b57f5c
4	Dynamic cattle behavioural classification using supervised ensemble classifiers	Dutta et al.	https://www.sciencedirect.com/science/article/pii/S0168169914003123
5	Modelling the spatial distribution of Fasciola hepatica in dairy cattle in Europe	Ducheyne et al.	https://biblio.ugent.be/publication/7093862
6	Simulating the Epidemiological and Economic Impact of Paratuberculosis Control Actions in Dairy Cattle	Kirkeby et al.	doi: 10.3389/fvets.2016.00090
7	A Robust Statistical Model to Predict the Future Value of the Milk Production of Dairy Cows Using Herd Recording Data	Græsbøll et al.	doi: 10.3389/fvets.2017.00013
8	Models to Estimate Lactation Curves of Milk Yield and Somatic Cell Count in Dairy Cows at the Herd Level for the Use in Simulations and Predictive Models	Græsbøll et al.	doi: 10.3389/fvets.2016.00115

4.8 Current related SEGES projects

	Project name	Relation to this project
1	Kviebarometret	This project investigated the relationship between sickness and environment of the heifer's first month and its lifespan (i.e. the cow life expectancy of the individual cattle). This project does not utilize machine learning, but traditional data analysis and visualization to draw conclusions.

5 Tentative time schedule

This project is part of the time schedule of the Data Science team.

6 Data description

The available data consists of registrations for individual animals in a tabular format.

6.1 Data sources

Source name	Source format	Storage and usage restrictions	Additional information
Kvægdatabasen (MEDIO)	Oracle Database	Read only access. No special storage and usage restrictions.	The MEDIO database is a snapshot of the production data taken a few times a year. Thus, the DB does not contain up-to-date production data.

6.2 Data attribute information regarding the heifer

We only query data regarding "Dansk Holstein" cattle born from 2000-01-01 and onward.

	Data source	Data type	Range	Additional information	Features	Feature type	Feature description
1	Kvægdatabasen	Integer	1-5	Danish description: størrelse ved fødsel Ordered classified size. Numer 5 is 'ophørt' and outside order.	foedstrkode	categorical(1,2,3,4), ordered	"Kviens fødselsstørrelsekode" Ordered size (1 small - 4 large). Remove any 5 (udgået). Note that the scale has changed at some point from a 5 level scale to a 4 level scale.
2	Kvægdatabasen	datetime		Danish description: fødselsdato Perhaps converted to day of year.	foedselsaar foedselsdag foedselsdagiugen foedselsmd	integer(2000;2013) categorical(1,...,31), categorical(0,...,6) categorical(1,...,12),	"Kviens fødselsaar" "Dag i måneden for kviens fødsel" "Dag i ugen for kviens fødsel" "Måned i året for kviens fødsel" Day-of-month may be excluded.

	Data source	Data type	Range	Additional information	Features	Feature type	Feature description
3	Kvægdatabasen	integer		Danish description: alder i dage ved 1. inseminering	alder_insem	integer(300;900)	"Kviens alder i dage ved hendes 1. inseminering" NULLs may be interpreted as "the use of a private bull" or may be removed
4	Kvægdatabasen	integer		Danish description: alder i dage ved 1. kælvning	alder_kaelvning	integer(550;1200)	"Kviens alder i dage ved hendes 1. kælvning"
5	Kvægdatabasen		1 - inf	Danish description: antal insemineringer før kælvning	insem_antal	categorical(1,2,3,4,5+), ordered	"Antal insemineringer af kviens inden hendes 1. kælvning"
6	Kvægdatabasen	integer	0 - inf	Danish description: antal sygregisteringer for specifikke sygdomme	lksk_11plus, lksk_2, lksk_28plus, lksk_41, lksk_53, lksk_72	bool bool bool categorical(0,1-3,4+), ordered bool bool	At least one registration or none. NULLs = none At least one registration or none. NULLs = none At least one registration or none. NULLs = none Ideally only registrations with more than 8 days in between should be counted At least one registration or none. NULLs = none At least one registration or none. NULLs = none "Antal sygeregistreringer for kvien med sygdomskoden lksk_* fra dens fødsel og frem til den første kælvning"

	Data source	Data type	Range	Additional information	Features	Feature type	Feature description
7	Kvægdatabasen	float		Danish description: fødselsvægt	foedselsvaegt	float(25;60)	"Kviens vægt ved fødsel (målt på dagen eller op til 3 dage efter) This attribute seems to include a lot of standard figures.
8	Kvægdatabasen	integer		Danish description: levetid i dage fra første kælvning til udsætning	ko_level alder	integer(0;7200)	"Antallet af dage fra kviens første kælvning til den udsættes (død, slagtet eller aflivet)" *This is the response variable.*
9	Kvægdatabasen	integer		Danish description: farens NTM (Nordic Total Merit) index	far_NTM	float(-inf;inf)	The dataset is divisioned into two groups (one with mean around 0 and one with mean around 100) corresponding to two different approaches to registration of NTM. These are not comparable. Neither are NTM values for different years. Thus, the NTM values must be normalized on a year basis, i.e. all data for 2000 must be normalized, all data for 2001 must be normalized, and so forth. Alternatively, this attribute may be excluded from the analysis.

7 Data attribute information regarding the mother of the heifer

	Data source	Data type	Range	Additional information	Features	Feature type	Feature description
1	Kvægdatabasen	integer	0-5	Danish description: kælvningsforløb Ordered classification. Numer 0 is 'ophørt'	mor_forloebskode	categorical(1,2,3-5), ordered	"Forløbskode for kviens fødsel" Remove all where forloebskode == 0
2	Kvægdatabasen	integer		Danish description: kælvningsnummer / laktationsnummer	mor_kaelvningsnr	categorical(1,2,3,4,5,6-10), ordered	"Kviens mors kælvningsnummer ved kviens fødsel"
3	Kvægdatabasen	integer		Danish description: mors alder i dage før første kælvning	mor_alder_foerste_kaelvning	integer(550; 1200)	"Kviens mors alder i dage ved morens første kælvning"
4	Kvægdatabasen	integer		Danish description: goldperiodelængde.	mor_goldperiode	categorical(0-28,29-42,43-56,57-119) + one flag value (-1) representing NULL and the ~600 having a goldningsdato for first borns (mor_kaelvningsnr==1)	"Længden af kviens mors goldperiode i antal dage før kviens fødsel"

	Data source	Data type	Range	Additional information	Features	Feature type	Feature description
5	Kvægdatabasen	integer	0 - inf	Danish description: antal sygregisteringer for specifikke sygdomme	mor_lksk_12 mor_lksk_21 mor_lksk_22 mor_lksk_91	bool bool bool bool	At least one registration or none. NULLs = none At least one registration or none. NULLs = none At least one registration or none. NULLs = none At least one registration or none. NULLs = none "Antal sygeregistreringer for moren med sygdomskoden lksk_* i perioden 1 uge før til 3 uger efter kviens fødsel"

8 Additional information

This project is part of the PAF project: "Øget konkurrencekraft i landbruget gennem brug af kunstig intelligens." - Ansøgning til promilleafgiftsfonden for landbrug 2018.pdf

8.1 Lists of diseases treatments:

Relevant diseases treatments for cows:

LKSYGKODE	Danish description
2	Børbetændelse
4	Efterbyrd
11	Yverbetændelse (slås sammen med 12, 14, 15, 94, 95 og 179)
12	Yverbetændelse, goldperioden
14	Yverbetændelse efter læsion
15	Yverbetændelse, akut
20	Løbeudvidelse (slås sammen med 23, 96 og 97)
21	Ketose
22	Kælvningsfeber
23	Løbedrejning
72	Fluemastitis
90	Børkrængning (kan evt. udelades)
91	Børslyngning (kan evt. udelades)
92	Kejsersnit (kan evt. udelades)
94	Yverbetændelse, brandig
95	Yverbetændelse, subklinisk

LKSYGKODE	Danish description
96	Løbedrejning, højresidig
97	Løbedrejning, venstresidig
179	Yverbetændelse med lammelse

* Mother related diseases marked blue. Diseases marked with orange are left out since they are deemed not relevant for the current analysis.

Relevant diseases treatments for calves:

LKSYGKODE	Danish description
28	Tarmbetændelse(slås sammen med 51)
41	Lungebetændelse
51	Diarre
53	Navlebetændelse

8.2 Additional data

Data attributes specified in the below table are only included in the analyses if time permits.

	Data source	Attribute description (Danish)	Granularity of registration	Number of registrations per animal	Data type	Range	Additional information
1	Kvægdatabase	flytninger, indenfor bedrift	individual	zero or more	Integer	1-50	Unordered classification 18 classes. Note that class 1 and 16 are relevant for 'Fytninger'.
2	Kvægdatabase	indkøb fra fremmed bedrift	individual	zero or more	Integer	1-50	Unordered classification 18 classes. Note that class 1 and 16 are relevant for 'Fytninger'.

	Data source	Attribute description (Danish)	Granularity of registration	Number of registrations per animal	Data type	Range	Additional information
3	Kvægdatabase	fødselsfordeling Distribution of what? sex, month, size.	herd	Several	?	?	besætningens kælvningsfordeling fx beskrevet som kælvningstæthed pr. måned/ procentvis fordeling
4	Kvægdatabase	kælvningsinterval in days	mother	one	integer	?	mellem fødsel af aktuelle individ og fødsel af forrige individ
5	Kvægdatabase	vægt	individual	zero or more			alle vejninger i individets levetid Expected to be registered for very few individuals.
6	Kvægdatabase	alder v. sygdom	individual	zero or more			alder ved hver sygdomskode
7	Kvægdatabase	drægtighedsundersøgelse, løbning, ilægning og kælvning.	individual	one			
8	Kvægdatabase	insemineringsintervaller, kvie	individual	zero or more			fx beskrevet som et gennemsnit at alle insemineringsintervaller for det enkelte individ
9	Kvægdatabase	insemineringsintervaller, ko	individual	zero or more			fx beskrevet som et gennemsnit at alle insemineringsintervaller for det enkelte individ

	Data source	Attribute description (Danish)	Granularity of registration	Number of registrations per animal	Data type	Range	Additional information
10	Kvæg-data-warehouse	antal insemineringer før drægtighedd, ko	individual	zero or more			fx beskrevet som et gennemsnit af antal insemineringer pr. drægtighed
11	Kvægdatabase	kælvningsdatoer	individual	zero or more			
12	Kvægdatabase	kælvningsmåneder	individual	zero or more			
13	Kvægdatabase	celletal	individual	zero or more			over tid
14	Kvægdatabase	ketosescore	individual	zero or more			alle ketosescore Few expected registrations.
15	Kvægdatabase	huldscore	individual	zero or more			alle huldscorer Few expected registrations.
16	Kvægdatabase	huldændring i **goldperiode**	individual	zero or more			
17	Kvægdatabase	paratb status	mother	zero or more			Expected to be registered for 20%-25% of all herd (maximally).
18	Kvægdatabase	klovregistreringer	individual	zero or more			
19	Kvægdatabase	klovbeskæringer	individual	zero or more			Not systematically registered. Only expected to be available for less than 50% of all individuals.

	Data source	Attribute description (Danish)	Granularity of registration	Number of registrations per animal	Data type	Range	Additional information
20	Kvæg-data-warehouse	spredning på afstand fra kælvning-første inseminering	individual	zero or more			alle tidligere kælvninger
21	Kvægdatabase	ydelse, opnået minus mål	individual	zero or more			