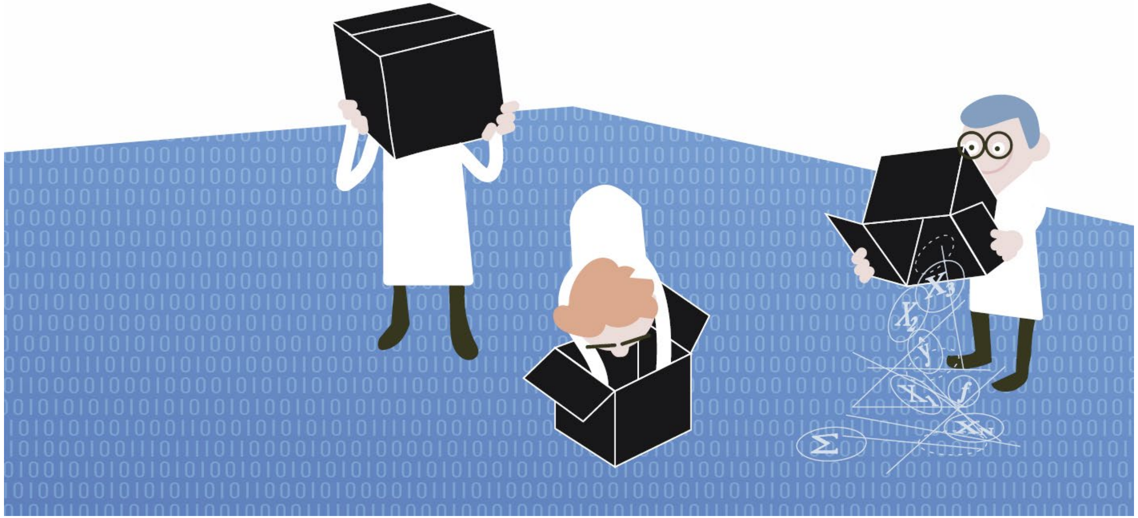
eXplego: Hvordan velge riktig metode for å forklare kunstig intelligens?



21.09.2023 – Anders Løland

eXplego: An interactive tool that helps you select appropriate XAI-methods for your explainability needs*

Martin Jullum[1,*], Jacob Sjødin[2], Robindra Prabhu[2] and Anders Løland[1]

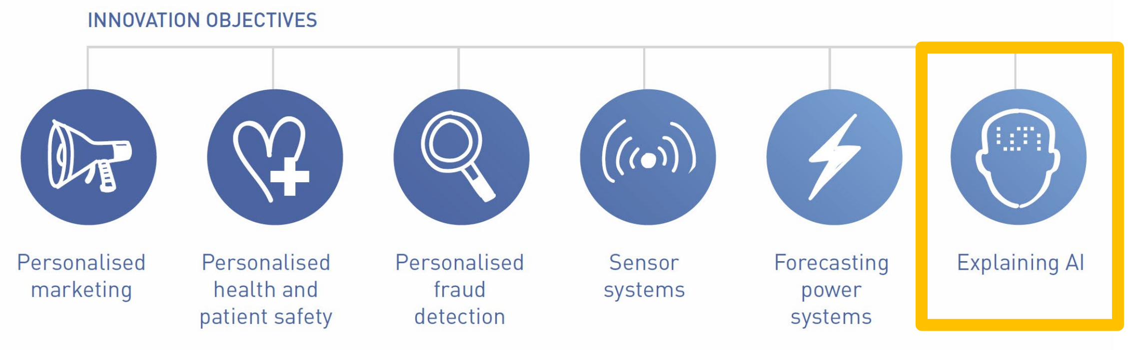
[1] *Norwegian Computing Center, P.O. Box 114, Blindern, NO-0314 Oslo, Norway*

[2] *NAV IT Utvikling og Data, Arbeids- og velferdsdirektoratet, Fyrstikkalléen 1, 0661 Oslo, Norway*

Abstract

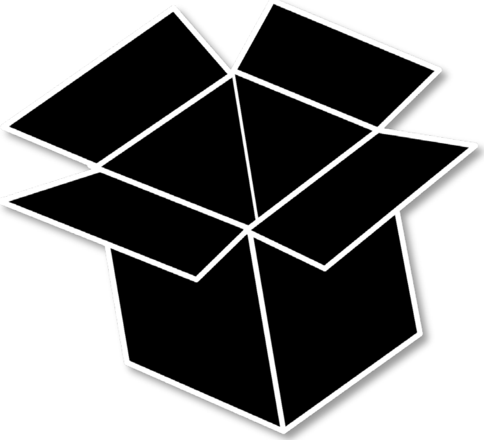
The growing demand for transparency, interpretability, and explainability of machine learning r
and AI systems has fueled the development of methods aimed at understanding the properti
behavior of such models (XAI). Since different methods answer different explainability questio
crucial to understand the kind of explanation the different XAI-methods provide, and in what situ
they should be used. We introduce **eXplego**, an interactive tree-structured tool designed to

BigInsight

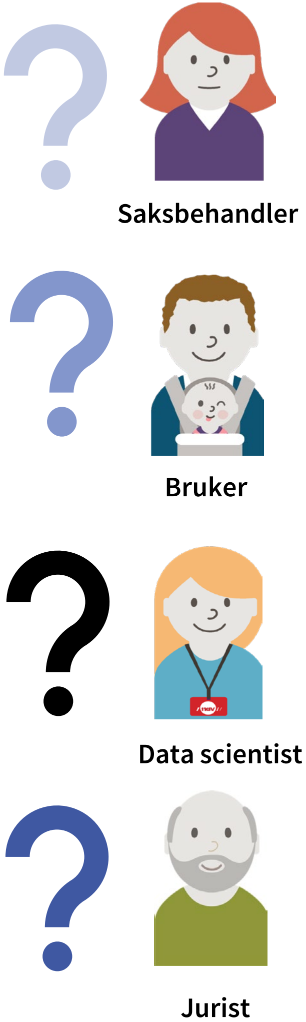


nav

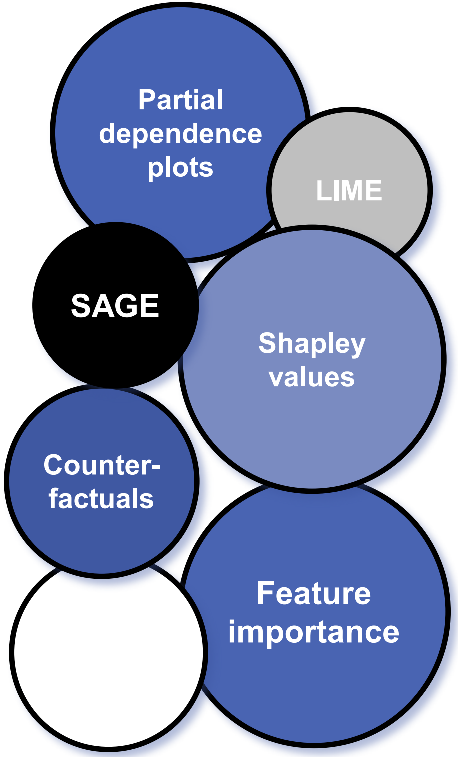
NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER




beslutning


Saksbehandler
Bruker
Data scientist
Jurist

eXplego


Partial dependence plots
LIME
SAGE
Shapley values
Counter-factuals
Feature importance

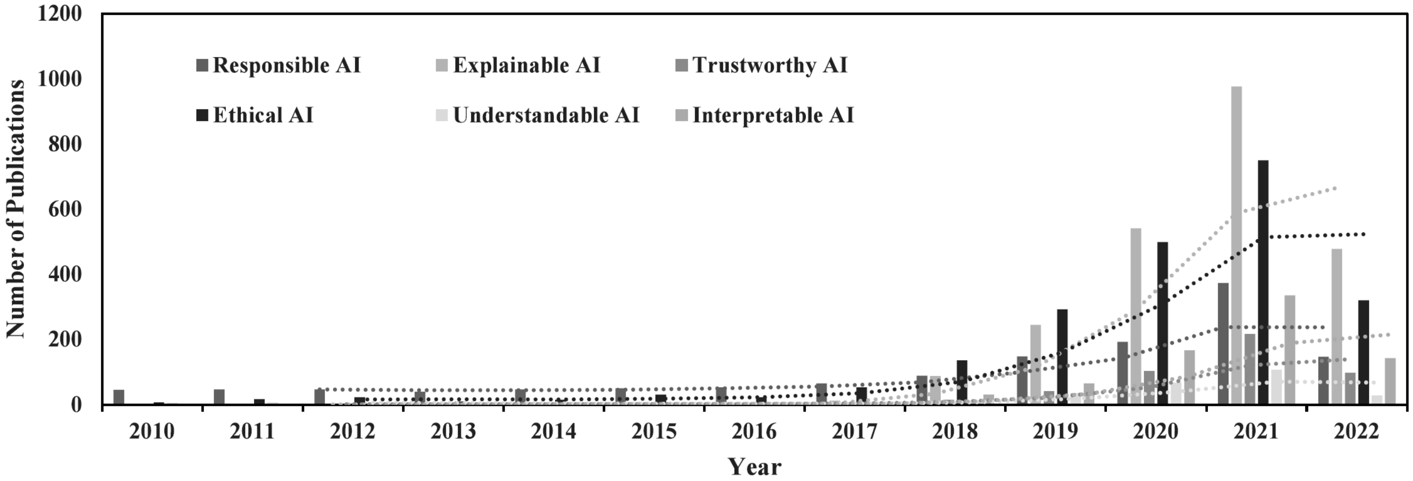


Fig. 30. The evolution of the total number of publications on XAI over time. The dotted lines show the trend over the previous three years using a moving average. These statistics were retrieved from the Scopus database in June 2022.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2018, accepted September 4, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2870052

Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

AMINA ADADI AND MOHAMMED BERRADA
Computer and Interdisciplinary Physics Laboratory, Sidi Mohammed Ben Abdellah University, Fez 30050, Morocco
Corresponding author: Amina Adadi (amina.adadi@gmail.com)

ABSTRACT
adoption of art
more algorithr
use of AI-base
allows powerf
explainable A
AI-based syste
disruption. Th
aspects of the
we review the
major researcl

Artificial Intelligence 298 (2021) 103502

Contents lists available at ScienceDirect

Artificial Intelligence

ELSEVIER

www.elsevier.com/locate/artint

Explaining individual predictions when features are dependent: More accurate approximations to Shapley

Kjersti Aas *, Martin Jullum, Anders Løland

Norwegian Computing Center, P.O. Box 114, Blindern, N-0314 Oslo, Norway

ARTICLE INFO

Article history:
Received 3 October 2019
Received in revised form 5 January 2021
Accepted 29 March 2021
Available online 31 March 2021

Keywords:
Feature attribution
Shapley values
Kernel SHAP
Dependence

ABSTRACT

Explaining complex or seemingly simple machine lear
problem. We want to explain individual predictions f
interpretable explanations. Shapley value is a game t
this purpose. The Shapley value framework has a seri
and can in principle handle any predictive model
efficient approximation to Shapley values in higher d
methods, this approach assumes that the features ar
currently suffer from inclusion of unrealistic data in
the explanations may be very misleading. This is
model is used for predictions. In this paper, we
handle dependent features. We provide several exam

BigInsight

← BACK TO ALL EVENTS

EXPLAINING AI SEMINAR: RICH CARUANA (MICROSOFT RESEARCH)

Thursday, March 23, 2023
09:00 – 10:00

Speaker: Rich Caruana (Microsoft Research)

Location: Click here to join the meeting (Microsoft Teams)

Title: Friends Don't Let Friends Deploy Black-Box Models: The Importance of Intelligibility in Machine Learning

Abstract: In machine learning often tradeoffs must be made between accuracy and intelligibility: the most accurate models usually are not very intelligible, and the most intelligible models usually are less accurate. This can limit the accuracy of models that can safely be deployed in mission-critical applications such as healthcare where being able to understand, validate, edit, and trust models is important. EBMs (Explainable Boosting Machines) explego.nr.no

eXplego

Explainability is key requisite for trustworthy AI, but selecting the right XAI-method to accompany your model development can be a challenging task. **eXplego** is a decision tree toolkit that provides developers with interactive guidance to help select an appropriate XAI-method for their particular use case.

This is a collaborative project between Norsk Regnesentral and the Norwegian Labour and Welfare Administration (NAV), funded by BigInsight.

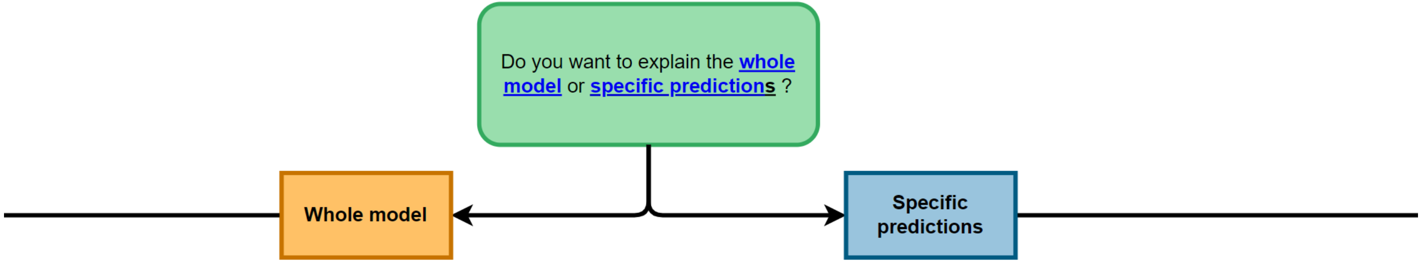
TOGGLE USER INSTRUCTIONS

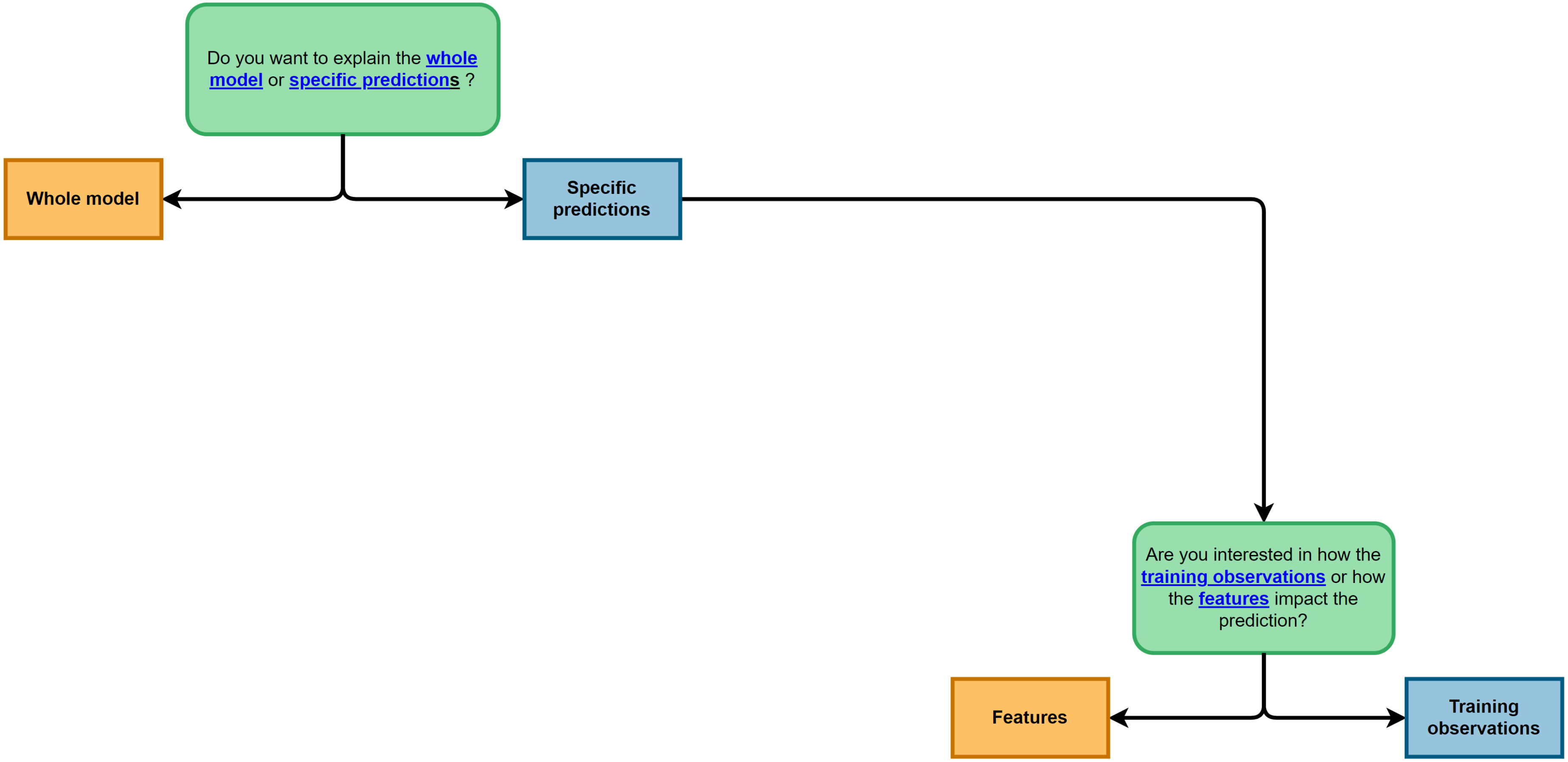
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

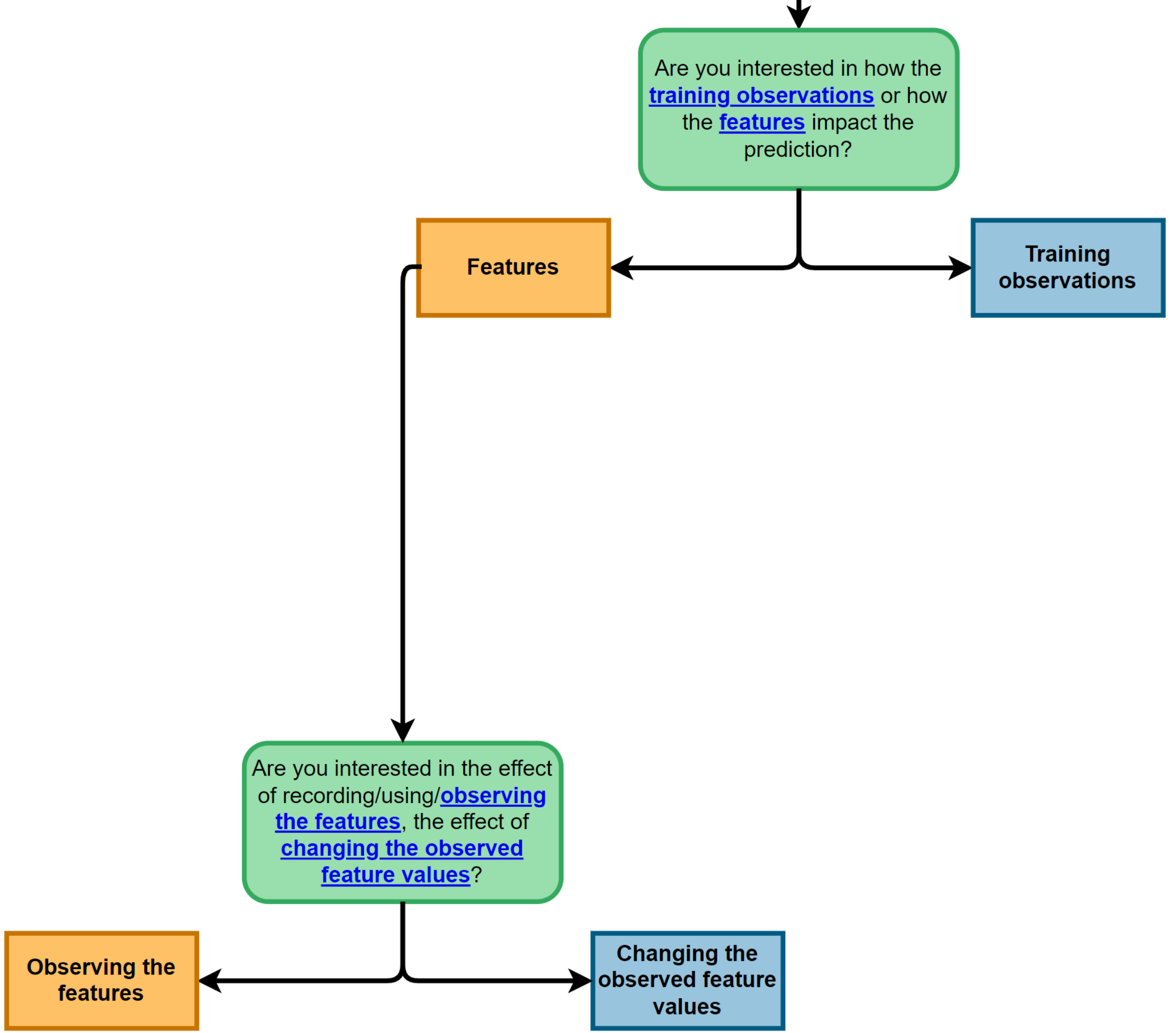
SHOW ENTIRE TREE

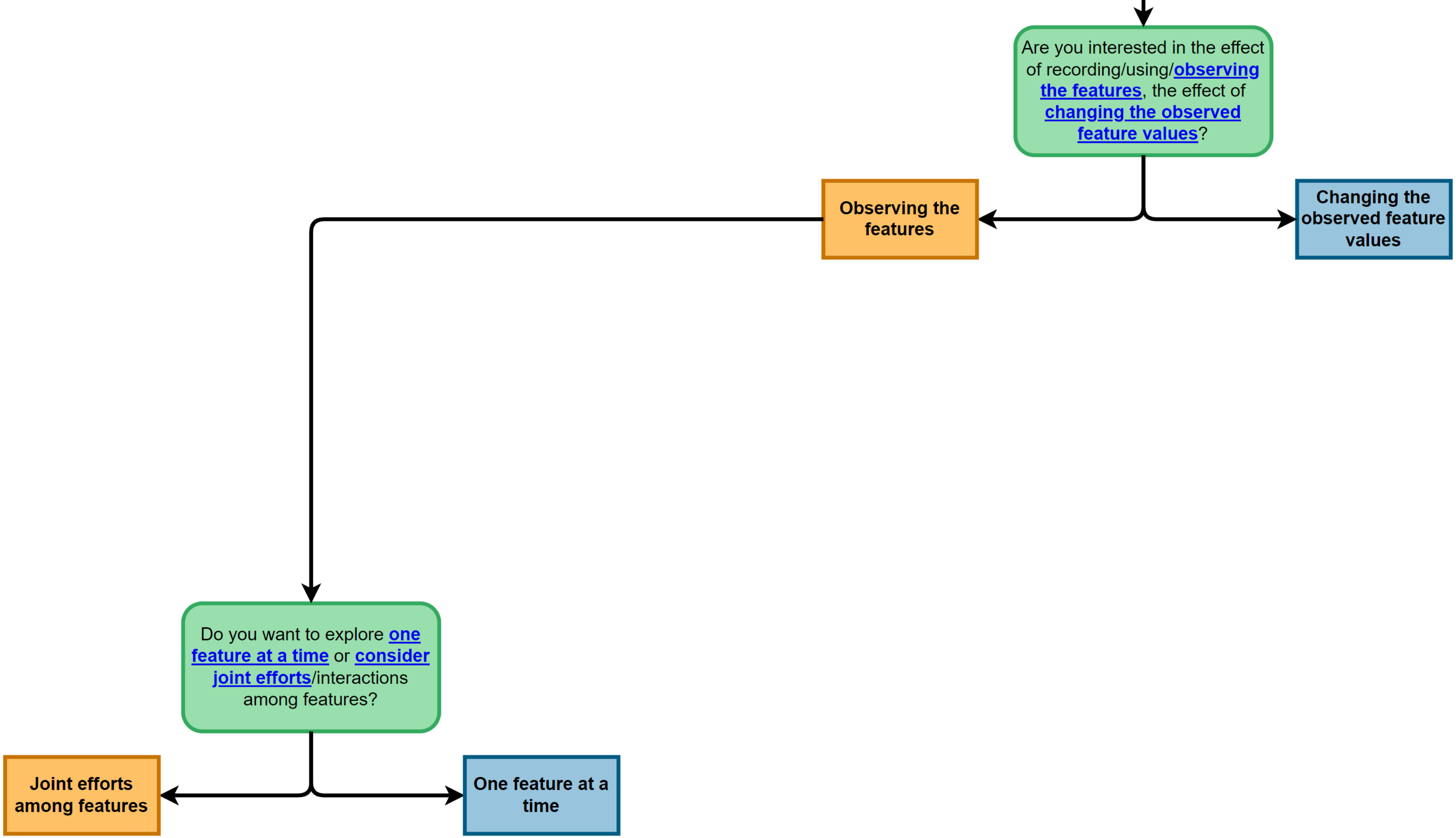
RESET DISPLAY

SHOW ENTIRE TREE WITH EXAMPLES

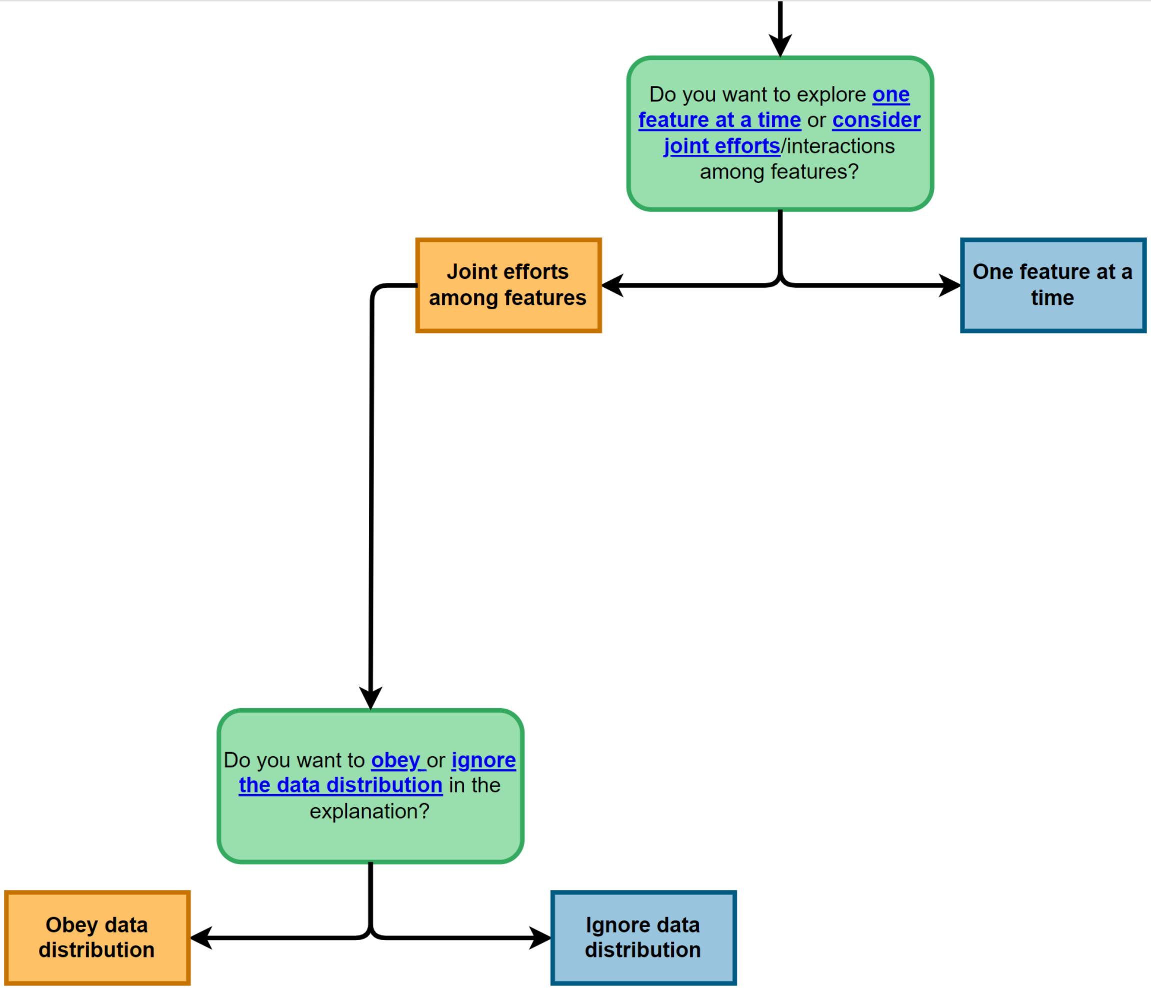



Do you want to explain the whole model or specific predictions ?
Whole model
Specific predictions
Are you interested in how the training observations or how the features impact the prediction?
Features
Training observations

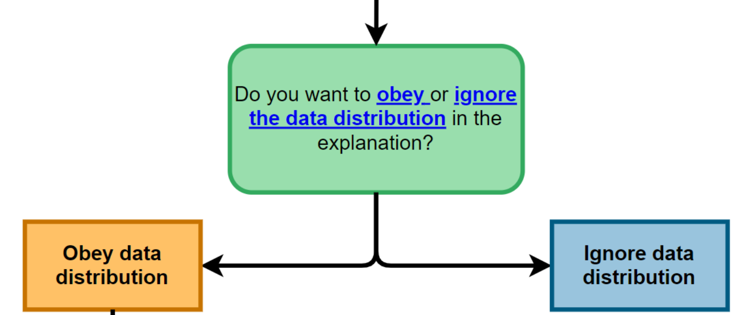

Are you interested in how the training observations or how the features impact the prediction?
Features
Training observations
Are you interested in the effect of recording/using/observing the features, the effect of changing the observed feature values?
Observing the features
Changing the observed feature values



Are you interested in the effect of recording/using/observing the features, the effect of changing the observed feature values?
Observing the features
Changing the observed feature values
Do you want to explore one feature at a time or consider joint efforts/interactions among features?
Joint efforts among features
One feature at a time



Do you want to explore one feature at a time or consider joint efforts/interactions among features?
Joint efforts among features
One feature at a time
Do you want to obey or ignore the data distribution in the explanation?
Obey data distribution
Ignore data distribution



Conditional Shapley values

Short method description: Uses Shapley values from game theory to decompose the prediction outcome onto the different features. Uses conditional expectation to measure feature contribution, and obeys the data distribution when estimating these.

Interpretation: The Shapley value of a feature says (roughly!) how of the prediction changed due to the feature being observed.

Software: shapr (R)

Original paper: Aas et al. (2021)

Other resources: Sec 5.1 of review paper by Chen et al. (2022) describing the difference between conditional and marginal Shapley values.

Nice to know:

- Estimation of the conditional expectations measuring the feature contribution is a research topic on its own. For continuous unimodal data, we recommend the Gaussian or "empirical" approach, for mixed data we recommend the ctree approach.

Artificial Intelligence 298 (2021) 103502

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Explaining individual predictions when features are dependent: More accurate approximations to Shapley values ☆

Kjersti Aas *, Martin Jullum, Anders Løland

Norwegian Computing Center, P.O. Box 114, Blindern, N-0314 Oslo, Norway

A R T I C L E I N F O

Article history:
Received 3 October 2019
Received in revised form 5 January 2021
Accepted 29 March 2021
Available online 31 March 2021

Keywords:
Feature attribution
Shapley values
Kernel SHAP
Dependence

A B S T R A C T

Explaining complex or seemingly simple machine learning models is an important practical problem. We want to explain individual predictions from such models by learning simple, interpretable explanations. Shapley value is a game theoretic concept that can be used for this purpose. The Shapley value framework has a series of desirable theoretical properties, and can in principle handle any predictive model. Kernel SHAP is a computationally efficient approximation to Shapley values in higher dimensions. Like several other existing methods, this approach assumes that the features are independent. Since Shapley values currently suffer from inclusion of unrealistic data instances when features are correlated, the explanations may be very misleading. This is the case even if a simple linear model is used for predictions. In this paper, we extend the Kernel SHAP method to handle dependent features. We provide several examples of linear and non-linear models

BigInsight

INNOVATION OBJECTIVES

- Personalised marketing
- Personalised health and patient safety
- Personalised fraud detection
- Sensor systems
- Forecasting power systems
- Explaining AI

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

UNIVERSITETET I OSLO

nav

Skatteetaten

DNB

Gjensidige

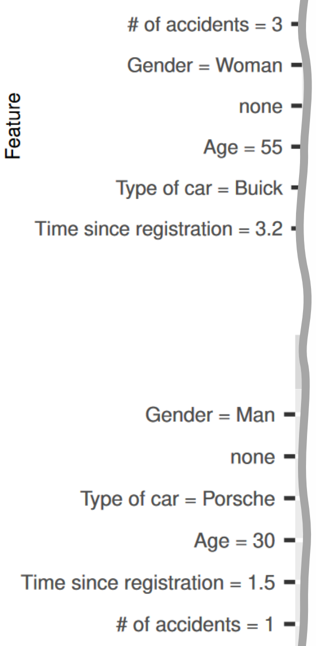
Shapley-verdier

Jeg vil forklare prediksjonen.

kontrafaktiske forklaringer

Jeg vil forklare beslutningen.

Shaple ske er



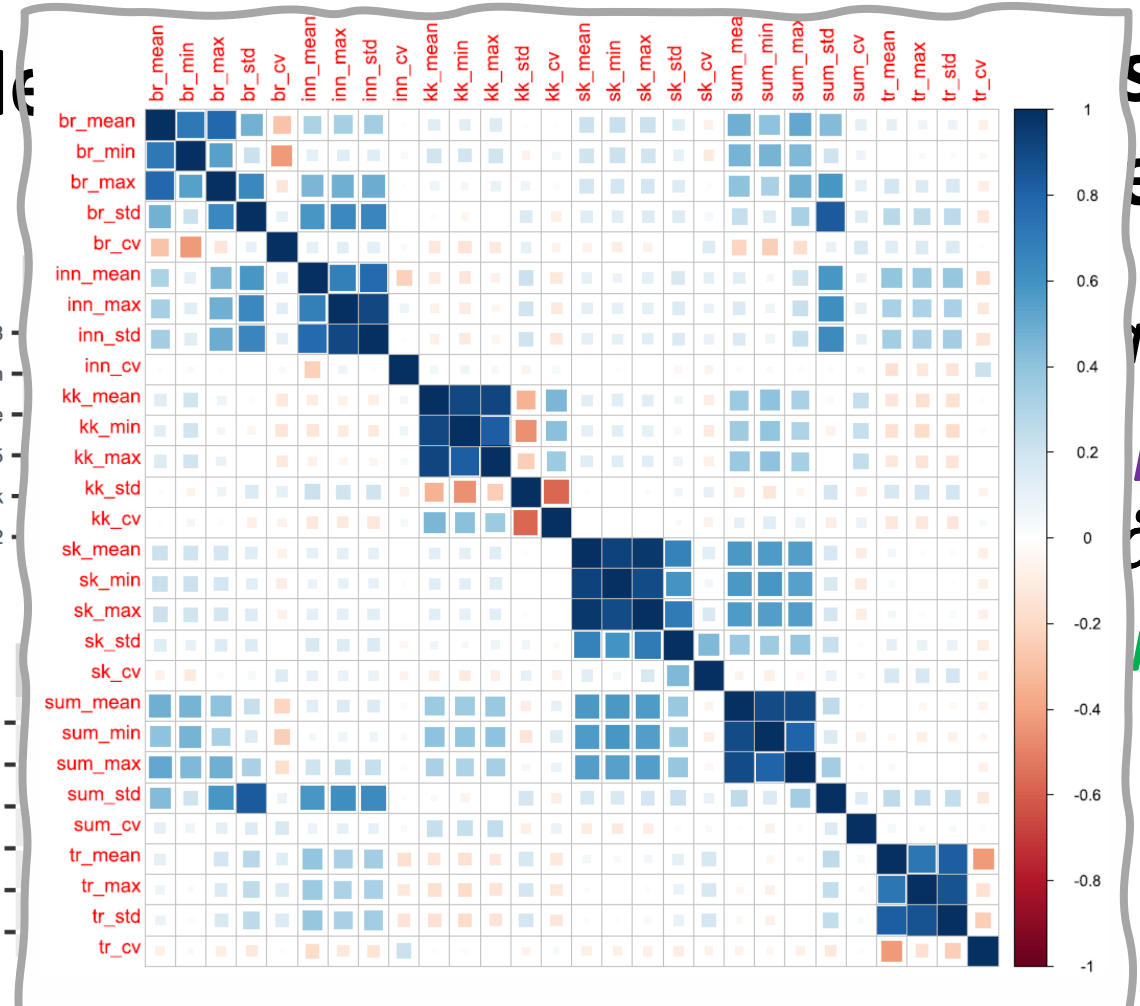


Fig. 7. Kendall's τ correlation matrix for the real data set.

adde hatt

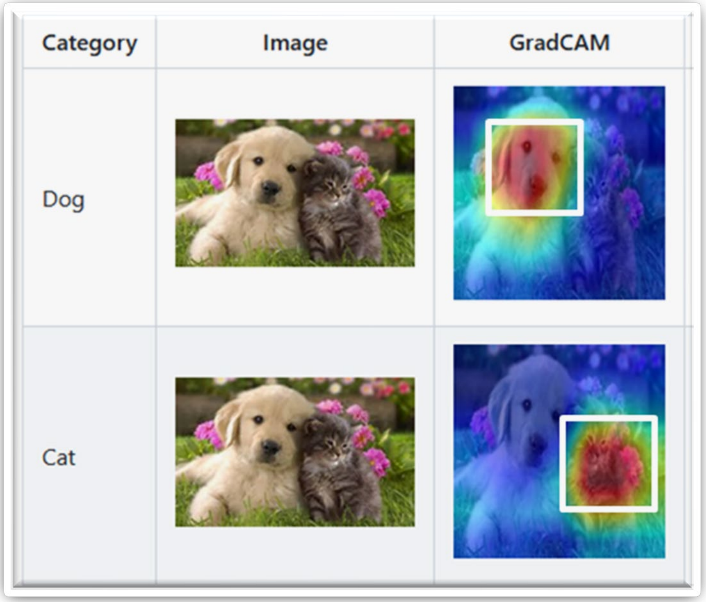
på et år,

ått

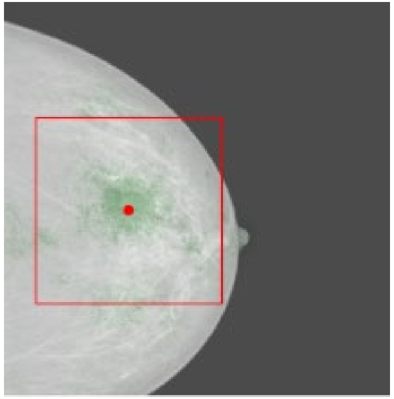
ikring.”

utfall.

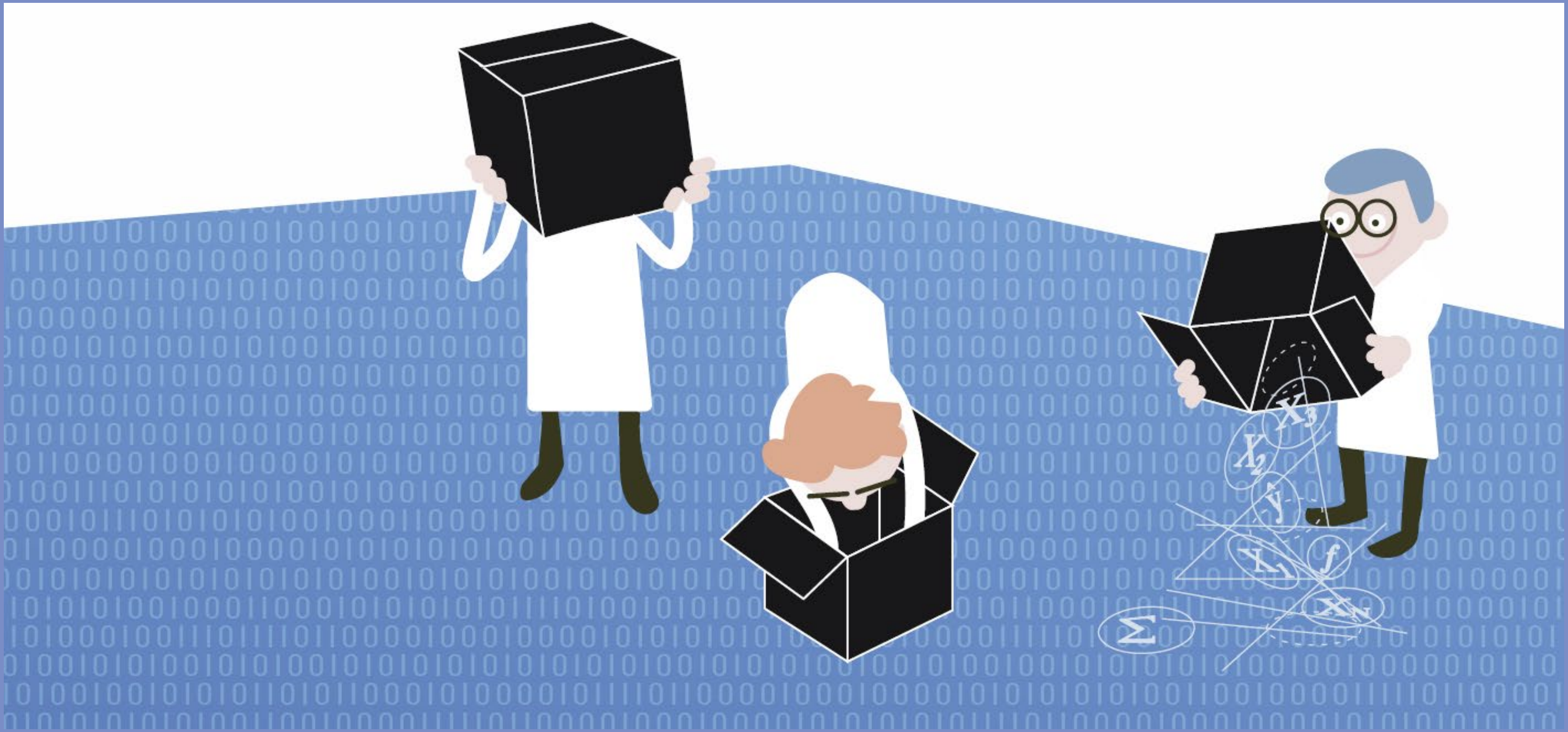
Forklaringer av bilder



Eksempel for katter og hunder



Vi kan gjøre tilsvarende for mammografi

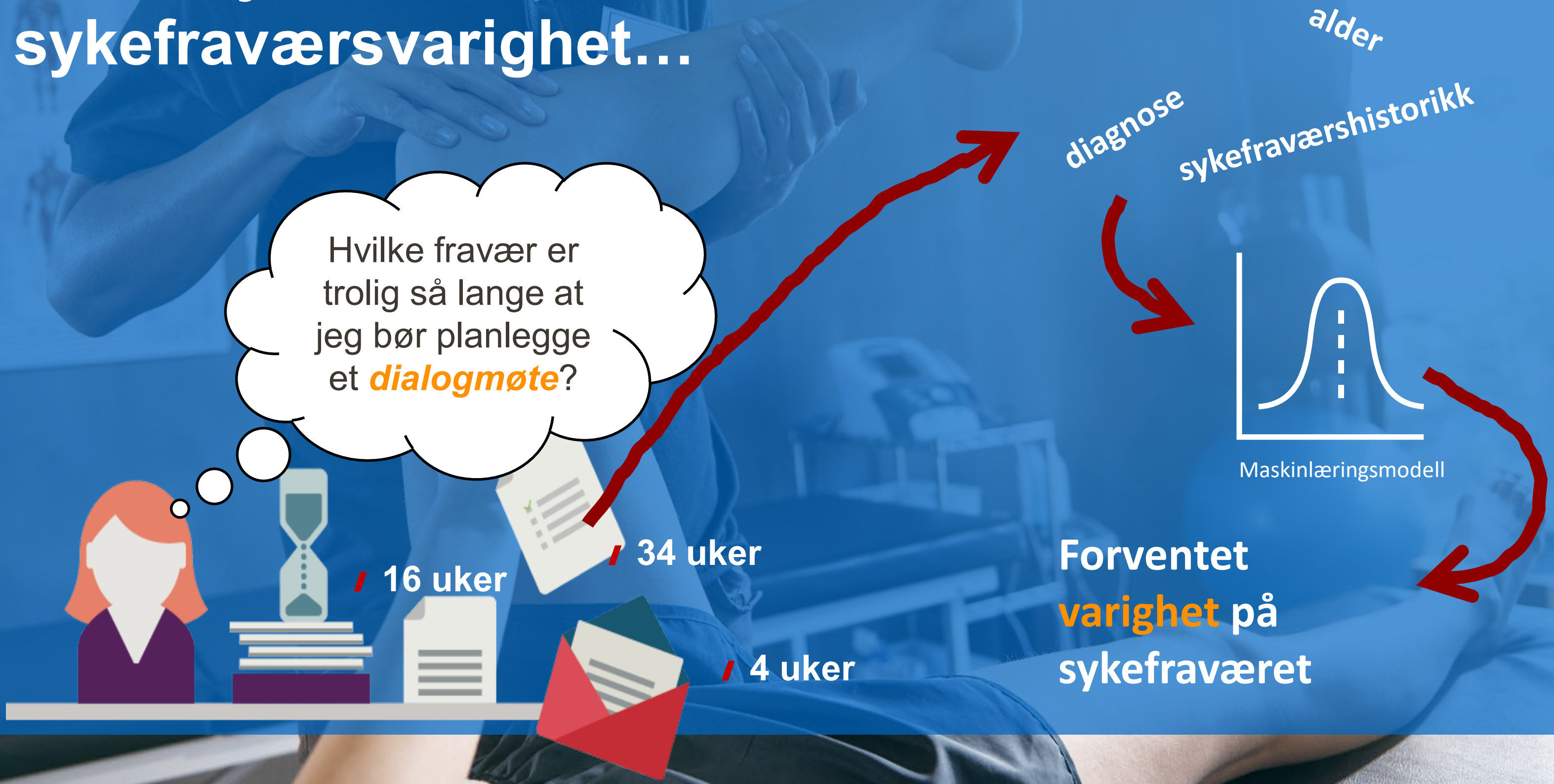

X_3
X_2
$\hat{y}$
X_1
f
X_N
Σ

«Forklar det den som kan!»

Hvordan jobbe med forklarbarhet i produktutviklingen?

Robindra Prabhu, data scientist
NAV IT Utvikling og Data
21.09.23

Prediksjon av sykefraværsvarighet…

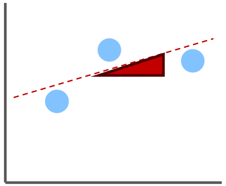


Data scientist

…vent litt!

? Variablene våre er jo avhengige!

? Noen av variablene er jo ganske tekniske...



Gradient of a fit to the last n sick leave gradations

? Information overload!
Blir nødt til å gruppere bidrag fra ulike variabler.

JOSS The Journal of Open Source Software

shapr: An R-package for explaining machine learning models with dependence-aware Shapley values

Nikolai Sellereite[1] and Martin Jullum[1]

1 Norwegian Computing Center

DOI: 10.21105/joss.02027

Software
- Review
- Repository
- Archive

Editor: Yuan Tang
Reviewers:
- @frycast
- @expectopatronum

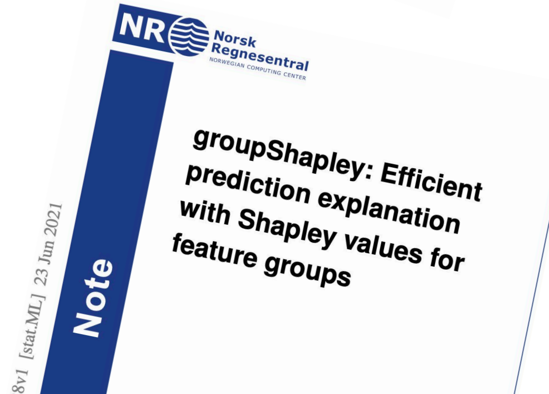
Submitted: 10 December 2019
Published: 05 February 2020

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

A common task within machine learning is to train a model to predict an unknown outcome (response variable) based on a set of known input variables/features. When using such models for real life applications, it is often crucial to understand why a certain set of features lead to a specific prediction. Most machine learning models are, however, complicated and hard to understand, so that they are often viewed as "black-boxes", that produce some output from some input.

Shapley values (Shapley, 1953) is a concept from cooperative game theory used to distribute fairly a joint payoff among the cooperating players. Štrumbelj & Kononenko (2010) and later Lundberg & Lee (2017) proposed to use the Shapley value framework to explain predictions by distributing the prediction value on the input features. Established methods and implementations for explaining predictions with Shapley values like Shapley Sampling Values (Štrumbelj & Kononenko, 2014), SHAP/Kernel SHAP (Lundberg & Lee, 2017), and to some extent TreeSHAP/TreeExplainer (Lundberg et al., 2020; Lundberg, Erion, & Lee, 2018), assume that the features are independent when approximating the Shapley values. The R-package shapr, however, implements the methodology proposed by Aas, Jullum, & Løland (2019), where predictions are explained while accounting for the dependence between the features, resulting in significantly more accurate approximations to the Shapley values.



NR Norsk Regnesentral NORWEGIAN COMPUTING CENTER

Note

groupShapley: Efficient prediction explanation with Shapley values for feature groups

8v1 [stat.ML] 23 Jun 2021

Saksbehandler

Grupperte Shapley-verdier

Sannsynlig friskmelding om **14 uker og 2 dager** ± 7 dager

Sannsynlig friskmelding

15.01 **18.03** **17.04** **01.05** **12.09.2020**

Frist dialogmøte 2

Maksdato

Dette trekker varigheten opp

1. Sykmeldingsgrad
2. Bosted
3. Yrke

Dette trekker varigheten ned

1. Diagnose
2. Lege
3. Alder

Detaljert informasjon ^

“Hvorfor det? Ikke det jeg ville vektlagt…”

Om faktorene

Sykmeldingsgrad

- graden som brukes i sykmeldingen ved uke 17
- gjennomsnittlig sykmeldingsgrad fram til uke 17
- forholdet mellom sykmeldingsgraden i siste og nest siste sykmelding

Bosted

- kommunenummer
- gjennomsnittlig lengde på sykefravær for innbyggerne i kommunen
- arbeidsledighet i kommunen måneden før personen har vært sykmeldt i 17 uker

Yrke

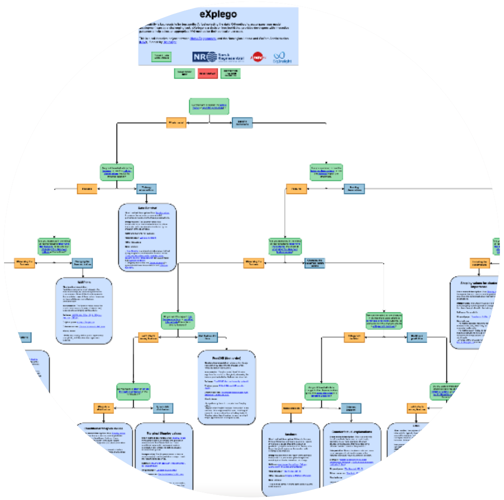
Juristen

- Forutberegnelighet
- Saklighet
- Kontradiksjon
- Offentlighet

"

En gjennomgang av noen utvalgte lokale forklaringsmetoder for komplekse modeller, **herunder approksimasjoner og kontrafaktiske forklaringer**, tyder på at slike metoder ikke tilfredsstiller forvaltningslovens krav til begrunnelse. Det tilsier at enklere og mer forklarbare modeller bør velges i automatiserte avgjørelser.

Hvordan jobbe med forklarbarhet i produktutviklingen?



Bruk XAI. Bruk eXplego!

Gir deg en god start!

Design rundt kontekst

Sjekk hvordan løsningen mottas, brukes og oppfattes

Tverrfaglighet

Innvolver flere fagmiljøer og test på ulike brukergrupper