KURSPLAN

Förklarbar artificiell intelligens, 5 hp

Explainable AI, 5 credits

Kurskod: DT8060

Akademin för informationsteknologi

Nivå: Avancerad nivå

Välj kursplan

Version

2026-01-19 - Tills vidare

Fastställd av: Forsknings- och utbildningsnämnden , 2024-09-18 och gäller studenter antagna vårterminen 2026.

Huvudområde med fördjupning

Datateknik, Avancerad nivå, har endast kurs/er på grundnivå som förkunskapskrav. (A1N)

Behörighetskrav

Högskoleingenjörsexamen i datateknik inklusive ett självständigt arbete 15 hp eller Teknologie kandidatexamen i huvudområdet datateknik inklusive ett självständigt arbete 15 hp. 7,5 hp programmering och 7,5 hp matematik inklusive linjär algebra. Engelska 6 eller Engelska nivå 2. Undantag ges för kravet på svenska, för dig med utländska betyg.

Kursens inplacering i utbildningssystemet

Kursen ges som en fristående kurs.

Mål

Kursens mål är att studenten utvecklar kunskaper och färdigheter inom en mängd olika aspekter inom förklarbar AI (XAI) inklusive: behovet av och vikten av att förklara olika AI-metoder, taxonomin för XAI och klassiska och välkända XAI-metoder. Studenten ska utveckla kunskap av både teoretiska och praktiska slag.

Efter avslutad kurs ska studenten kunna:

Kunskap och förståelse

- redogöra för kategorisering av XAI-metoder
- redogöra för olika välkända XAI-metoder
- diskutera olika mätetal för att utvärdera XAI-metoder

Färdighet och förmåga

- självständigt implementera XAI-metoder för en given AI-metod för att öka förklarbarheten
- utifrån givna ramar välja en relevant XAI-metod för en given AI-metod
- avväga mellan olika aspekter inom XAI såsom modellprestanda och förklaringsbarhet

Värderingsförmåga och förhållningssätt

- utvärdera XAI-metoder utifrån olika egenskaper inklusive precision & trohet, robusthet, osäkerhet och representativitet
- utvärdera kvaliteten på artificiell intelligens förklaringen utifrån ett mänskligt perspektiv genom att överväga egenskaper såsom begriplighet, selektivitet och kontrastivitet

Innehåll

Kursen omfattar följande ämnen:

- Introduktion till förklarbar AI, vad XAI är, varför det är viktigt samt relaterade terminologier
- Bred taxonomi av XAI-metoder inklusive Intrinsic jämfört med post hoc, modellspecifik jämfört med modellagnostisk och lokal jämfört med global
- Avvägning mellan noggrannhet och förklarbarhet samt förklaringar som är anpassade till mänsklig förståelse
- Egenförklarbara modeller inklusive linjär regression, logistisk regression, generaliserad linjär modell (GLM), generaliserad additiv modell (GAM) och beslutsträd.
- XAI-metoder inklusive, Partial Dependence Plot (PDP), Conformal Prediction, Individual Conditional Expectation (ICE), Feature Importance, Saliency Maps, Local Interpretable Model-Agnostic Explanations (LIME), SHAP, Integrated Gradient (IG)
- Utvärdering av förklaringsbarhet

Undervisningsspråk

Undervisningen bedrivs på engelska.

Undervisning

Både föreläsningar och laborationer kommer att hållas online. Laborationerna är designade i Python och utvecklade för att göra de koncept som ges under föreläsningarna förklarbara. Videor av föreläsningarna kommer också att läggas ut online via högskolans lärplattform för lärande i egen takt.

Undervisningen är på engelska och helt online.

Betygsskala

Tvågradig skala (UG): Underkänd (U), Godkänd (G)

Examinationsformer

Examination sker genom laborationer och skriftlig tentamen. Laborationer utförs i Python och inlämnas i form av Jupyter Notebooks.

2301: Skriftlig tentamen, 2,5 hp
Tvågradig skala (UG): Underkänd (U), Godkänd (G)

2302: Praktiska uppgifter, 2,5 hp
Tvågradig skala (UG): Underkänd (U), Godkänd (G)

Undantag från angiven examinationsform

Om särskilda skäl finns får examinator göra undantag från angiven examinationsform och medge att en student examineras på annat sätt. Särskilda skäl kan till exempel vara beslut om riktat pedagogiskt stöd.

Kursvärdering

I kursen ingår kursvärdering. Denna är vägledande för utveckling och planering av kursen. Kursvärderingen dokumenteras och redovisas för studenterna.

Kurslitteratur och övriga läromedel

Välj litteraturlista

2025-01-20 – Tills vidare

Litteraturlista 2025-01-20 – Tills vidare

Beslutad av: Forsknings- och utbildningsnämnden, 2024-09-18.

Molnar, Christoph. *Interpretable Machine Learning*. Leanpub 2019
Tillgängnlig online:
https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html

Rothman, Denis. *Hands-On Explainable AI (XAI) with Python*. Packt 2020

Forskningsartiklar inom XAI (vilka kommer att delas ut under kursens gång).