Rapport

AI – Utvecklingsprocessen och data

ES2024-01

Januari 2024

Innehåll

1. Sammanfattning

Enligt regeringsformens grundläggande principer ska den offentliga makten utövas utifrån saklighet och med målet att gynna den enskildes personliga, ekonomiska och kulturella välfärd. Varje myndighet har i sin verksamhet att förhålla sig till grundläggande principer i lagstiftningen såsom enkelhet, snabbhet, kostnadseffektivitet och rättssäkerhet. Artificiell intelligens (AI) kan bidra till myndighetens efterlevnad av dessa principer och det finns stora ekonomiska vinster och välfärdsnyttor med AI i alla sektorer inom offentlig förvaltning. AI-utveckling med maskininlärning (ML) kräver dock ytterligare förmågor än traditionell utveckling.

I denna rapport beskrivs exempel på olika utvecklingsprocesser för AI/ML och vilka fördelar respektive nackdelar som finns vid användning av dessa processer. Vidare beskrivs kortfattat roller som förekommer i en utvecklingsprocess för AI/ML, t.ex. verksamhetsexpert, verksamhetsarkitekt, it-arkitekt, arkivarie, utvecklingsteam, informationssäkerhetsansvarig, informationsägare, data scientist, data engineer, jurist, användare och roller i förvaltning.

Rapporten beskriver också en generell utvecklingsprocess och hur datahantering sker i denna utifrån följande uppdelning:

- Initiala överväganden inför en AI/ML-utveckling
- Beskrivning av de olika utvecklingsstegen
- Övervakning och livscykelhantering av AI/ML-system i produktion och förvaltning

Vid *initiala överväganden* behöver myndigheten bl.a. göra en bedömning av verksamhetsbehovet, om utvecklingsinsatsen ligger inom myndighetens uppdrag, om myndigheten ska utveckla själv eller införskaffa och om utvecklingsinsatsen bör genomföras alls t.ex. utifrån etiska aspekter.

AI/ML-utveckling innehåller ofta olika steg och i denna rapport utgår vi från en övergripande beskrivning med följande *utvecklingssteg*.

- Identifiera och samla in data
- Förbereda data
- Välja algoritm och modell
- Träna, validera och testa samt utvärdera och finjustera
- Förbereda för produktion och förvaltning samt driftsätta

Vanligen itereras stegen, dvs. flera av stegen återupprepas och en del steg är iterativa i sig. Vilken datahantering som förekommer i de olika stegen varierar både mellan myndigheter och mellan olika projekt inom samma myndighet. I samtliga steg är det viktigt att beakta rättsliga, säkerhetsmässiga och etiska aspekter. I utvecklingsprocessen behöver myndigheten ta ställning till förklarbarhet, tillförlitlighet, icke-diskriminering och bias.

När ett AI/ML-system är i *produktion och förvaltning* är det viktigt att ansvaret och befogenheterna för AI/ML-systemet är tydligt definierade och att det finns fastställda rutiner för hantering, dokumentation m.m.

Några reflektioner kring myndigheters AI-ML-utveckling är att vid maskininlärning är det fråga om förutsägelser och det kan vara svårt att nå fram med budskapet att resultatet som kommer ut av datahanteringen handlar om sannolika utfall snarare än att något är sant eller falskt. Därför kan det finnas skäl att tala om *risker i förhållande till konsekvens* samt *frekvens i förhållande till allvarlighet*. Andra utmaningar är att få tillgång till tillräcklig mängd data för att genomföra en AI/ML-utveckling. Många gånger krävs också stora manuella insatser i form av kunskaps- och insiktshöjande åtgärder, skapa tränings- och valideringsdata, träningen i sig, datahantering, tolkning, juridiska bedömningar, integritets- och säkerhetsbedömningar m.m. Det finns också en risk att användningen av AI bidrar till negativa känslor eftersom AI/ML-systemet saknar människans förmåga till t.ex. intuition. Därför är det viktigt att en AI/ML-utveckling är väl förankrad och följer en viss process.

Genom samverkan och gemensam kunskapshöjning kan förhoppningsvis flera verksamheter komma igång med sin AI/ML-utveckling och använda AI/ML för att möta lagstiftningens grundläggande principer om enkelhet, snabbhet, kostnadseffektivitet och rättssäkerhet.

2. Inledning

Enligt regeringsformens grundläggande principer[1] ska den offentliga makten utövas utifrån saklighet och med målet att gynna den enskildes personliga, ekonomiska och kulturella välfärd. Därtill finns principer i förvaltningslagen[2] som myndigheter ska uppfylla vid ärendehandläggning. Finns tekniska möjligheter att åstadkomma enkelhet, snabbhet, kostnadseffektivitet och rättssäkerhet bör en myndighet använda sådan teknik om det inte finns goda skäl att avstå. Myndigheter ska använda sig av den på sikt, utifrån myndighetens bedömning, bästa teknik som finns att tillgå samt avstå från ny teknik som riskerar att äventyra dessa värden. Myndigheter behöver bedöma, och gärna kunna mäta, hur *enkel, snabb, kostnadseffektiv* och *rättssäker* handläggningen blir med hjälp av ny teknik jämfört med befintlig handläggning. På sikt skulle en sådan bedömning kunna innebära att maskininlärning i längden inte är en valbar metod utan ett måste för att efterleva nämnda principer.

Det finns stora ekonomiska vinster och välfärdsnyttor med artificiell intelligens (AI) i alla sektorer inom offentlig förvaltning, vilket har uppmärksammats av eSams medlemmar där allt fler påbörjar olika AI-initiativ.[3]

AI-utveckling kräver ytterligare förmågor än traditionell utveckling. Redan i utvecklingsfasen behöver myndigheten ha metoder och arbetssätt för att säkerställa att funktionella och icke-funktionella krav fångas upp och omhändertas på rätt sätt.[4]

Svenska myndigheter saknar i nuläget en generell process för utveckling av AI – maskininlärning (AI/ML), varför det kan finnas fördelar med att beskriva utvecklingsprocessen på ett mer standardiserat sätt.

2.1 Syfte och målsättning

Syftet med denna rapport är beskriva exempel på olika utvecklingsprocesser för AI/ML samt dess för- och nackdelar. Vidare beskrivs kortfattat roller som förekommer i en utvecklingsprocess för AI/ML. Syftet är också att beskriva en generell utvecklingsprocess för AI/ML på ett sätt som skapar transparens och förklarbarhet om hur data hanteras i de olika stegen. Rapporten syftar till att skapa en förståelse för utvecklingsprocessens olika steg, deras ordning och vilket resultat som skapas.

[1] 1 kap. 2 och 9 § regeringsformen.
[2] 9 § förvaltningslagen (2017:900).
[3] eSams rapport ES2022-03 Samverkan kring tillämpad AI, juni 2022.
[4] eSams rapport ES2022-03 Samverkan kring tillämpad AI, juni 2022.

Målsättningen är att beskrivningen av en generell utvecklingsprocess ska förenkla dialogen mellan de olika rollerna som är involverade i processen. Målsättningen är också att myndigheters ledande funktioner ska förstå processen och därmed behovet av datatillgång.

2.2 Avgränsning

Fokus i rapporten läggs främst på de olika utvecklingsstegen. Initiala överväganden och när modellen är i produktion berörs mer kortfattat. Det ska inte tolkas som att de mer kortfattat beskriva delarna är mindre viktiga, utan samtliga steg är av stor betydelse. Mer information om dessa finns bl.a. i *Förtroendemodellen* och *AI-guiden* som är publicerade på *Offentliga AI*.[5]

Rapporten är avgränsad till AI/ML-utveckling och till dess utvecklingsprocess och hantering av data. För ytterligare beskrivning av vad AI är, olika varianter av AI, vad AI kan användas till, nyttor och risker med AI, förmågor och hur AI ingår som en komponent i en systemmiljö, dvs. helheten, se eSams rapport *ES2022-03 Samverkan kring tillämpad AI*. För fördjupning om juridiska frågor hänvisas till eSams checklista *ES2023-06 Juridik vid användning av AI*.

2.3 Målgrupp

Målgruppen för rapporten är dels ledande funktioner inom myndigheten som har att besluta om AI-utveckling, dels andra som involveras i utveckling av AI, t.ex. verksamhets- och it-arkitekter samt jurister.

2.4 Medverkande

Arbetet med denna rapport har genomförts av en arbetsgrupp bestående av Anders Rydén (Lantmäteriet), Anders Enmark (Lantmäteriet), Anders Stenberg (Pensionsmyndigheten), Andreas Voxberg (Skatteverket), Boris Melvås (Pensionsmyndigheten), Jakob Engdahl (Statistiska centralbyrån), Katarina Lindh (Bolagsverket), Tina Chavoshi (E-hälsomyndigheten), Mikael Österlund (eSams kansli) och Linda Lindström (eSams kansli). Kvalitetssäkring har skett i eSams rättsliga expertgrupp, expertgruppen i säkerhet samt koordineringsgruppen för arkitektur. Beredning har skett via eSams samordningsgrupp.

[5] Se Sveriges Dataportal, Offentlig AI, https://beta.dataportal.se/offentligai/fortroendemodellen

2.5 Disposition

Avsnitt 1 ger en sammanfattning av innehållet i rapporten.

Avsnitt 2 ger en inledning till rapporten samt beskriver syfte, avgränsningar och målgrupp.

Avsnitt 3 ger en övergripande beskrivning av AI-system, maskininlärning och maskininlärningsprocessen.

Avsnitt 4 beskriver olika databegrepp som förekommer inom AI/ML-utveckling. Dessa begrepp är av vikt för förståelse av efterföljande avsnitt.

Avsnitt 5 beskriver olika utvecklingsprocesser och för- och nackdelar med dessa.

Avsnitt 6 beskriver en generell utvecklingsprocess på en *övergripande* nivå utifrån uppdelningen; initiala överväganden inför en AI/ML-utveckling, beskrivning av de olika utvecklingsstegen samt övervakning och livscykelhantering av AI/ML-system i produktion och förvaltning. Beskrivningen utgår från det gemensamma i de processer som beskrivs i avsnitt 5, men med fokus på datahantering. I avsnittet beskrivs också kortfattat exempel på roller som förekommer i en AI/ML-utveckling.

Avsnitt 7–11 ger en ytterligare beskrivning av utvecklingsstegen; identifiera och samla in data, förbereda data, välja algoritm och modell, träna, validera och testa samt utvärdera och finjustera, förbereda för produktion och förvaltning samt driftsätta. Det vill säga, avsnitten 7–11 ger en fördjupad beskrivning av utvecklingsstegen som beskrivs övergripande i avsnittet 6.

Avsnitt 12 innehåller reflektioner kring utmaningar och behov som kan föreligga för en myndighet vid AI/ML-utveckling.

3. AI-system och maskininlärning

3.1 AI-system

De finns ingen vedertagen definition av AI. Begreppet AI-system kommer att definieras inom ramen för EU:s arbete med den kommande förordningen om reglering av AI. Sannolikt omfattas system som baseras på maskininlärning och till viss del även kunskapsbaserade system (t.ex. expertmodeller).

I EU-kommissionens förslag till förordning om harmoniserade regler för AI (april 2021) föreslås en definition för AI-system.[6] Enligt denna är AI-system ett it-system eller en it-tjänst som innehåller programvara som utvecklats med en eller flera tekniker och metoder som, för en viss uppsättning människodefinierade mål, kan generera utdata såsom innehåll, förutsägelser, rekommendationer eller beslut som påverkar de miljöer som de samverkar med. De tekniker och metoder som åsyftas, och som kan komma att ändras över tid, är:

- maskininlärning, inbegripet övervakad, oövervakad och förstärkt inlärning, med hjälp av en mängd olika tillvägagångssätt, inklusive djupinlärning,
- logik- och kunskapsbaserade metoder, inklusive kunskapsrepresentation, induktiv (logisk) programmering, kunskapsbaser, inferens- och deduktionsmotorer, (symboliska) resonemang och expertsystem, eller
- statistiska metoder, bayesisk beräkning, sök- och optimeringsmetoder.

Europaparlamentet har i sitt reviderade förslag till AI-förordning (juni 2023) föreslagit följande definition: AI-system: ett maskinbaserat system som är utformat för att fungera med varierande grad av autonomi och som, för uttryckliga eller underförstådda mål, kan generera utdata såsom förutsägelser, rekommendationer eller beslut som påverkar fysiska eller virtuella miljöer.

Denna rapport fokuserar på utvecklingsprocessen för delområdet maskininlärning och därmed används uttrycket AI/ML genomgående i rapporten.

3.2 Maskininlärning

Maskininlärning (ML) är ett delområde inom AI och kan beskrivas som förmågan hos ett system att oberoende av mänsklig intelligens lära sig saker från exempel och tidigare

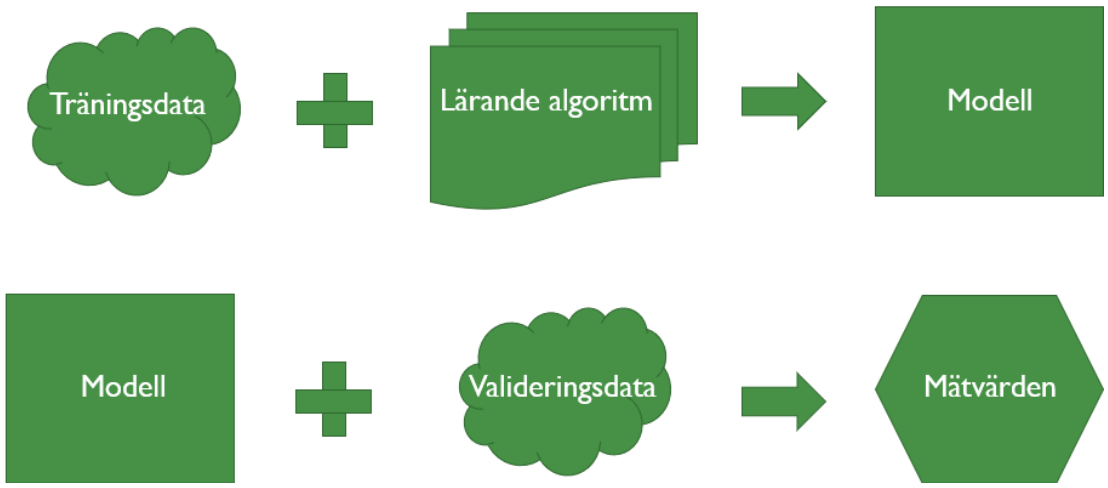
[6] Artikel 3.1, COM (2021) 206 Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

erfarenheter. Systemet korrigerar och anpassar beräkningsalgoritmerna utifrån utfall och behov.[7]

Maskininlärning syftar till att ge datorer förmågan att lära och anpassa sig till en uppgift utan att specifikt vara programmerad till det. Själva lärandet innebär att tillgängliga data kombineras med matematiska modeller för att i slutändan resultera i en modell som kan dra slutsatser om nya, okända data som inte ingått under själva träningen.

3.3 Översikt av maskininlärningsprocessen

Mycket översiktligt kan en maskininlärningsprocess sägas bestå av följande beståndsdelar:[8]



- **Träningsdata** är data som används för att träning av modellen.
- **Lärande algoritm(er)** är en serie instruktioner som beskriver hur, med vad och i vilken ordning en uppgift ska lösas, i det här fallet inlärningsprocessen.
- **En modell** är resultatet av träningsprocessen.
- **Valideringsdata** används för att utvärdera hur väl modellen uppnår förutbestämda mål och för att stämma av dess inlärningsprocess.
- Sådan utvärdering ger maskininlärningsprocessens **mätvärden** (metrics). Mätvärdena används för besluta om modellen uppfyller de förutbestämda målen tillräckligt väl.

Dessa delar beskrivs närmare i senare avsnitt. För beskrivning av ett AI/ML-system och AI/ML-komponenten som en del i helheten hänvisas till eSams rapport *Samverkan kring tillämpad AI*.

[7] eSams rapport ES2022-03 Samverkan kring tillämpad AI, juni 2022.
[8] Fredrik Heintz, professor, Institutionen för datavetenskap (IDA), Linköpings universitet.

4. Begrepp för data inom AI/ML-utveckling

Det förekommer flera olika databegrepp i samband med AI/ML-utveckling. Vid juridiska bedömningar används ofta begreppet ”testdata” som ett samlingsnamn för den mängd data som används vid AI/ML-utveckling. Ibland talas om produktionsdata, som en beskrivning för reella data från verksamheten. Ett annat sätt att beskriva data är att tala om strukturerade och ostrukturerade data.

Det är viktigt att vara medveten om att de olika beskrivningarna kan leda till missförstånd inom en arbetsgrupp med olika kompetenser. I avsnitt 6.2 samt 7–11 beskrivs olika utvecklingssteg (t.ex. träning, validering och testning) och det framkommer att det kan vara fråga om olika dataset i de olika stegen, vilket kan innebära skillnader i bedömningen av t.ex. de rättsliga förutsättningarna. Därför är det viktigt att ha en gemensam uppfattning om de begrepp som används. I detta avsnitt beskrivs några olika begrepp för data som förekommer inom AI/ML-utveckling. Ytterligare begrepp beskrivs på Sverige dataportal under *Offentlig AI*.[9]

4.1 Träningsdata

Vid framtagande av ett AI/ML-system måste en modell tränas med relevant data. Träningsdata är data som används för träning av modellen. Det sker genom att en lärande algoritm appliceras på data, vilket utmynnar i en tränad modell. Det är väsentligt att de data som används är representativa i förhållande till det man vill uppnå samt att snedvridningar eller partiskhet minimeras, se avsnitt 10.3 om bias. Fel vald eller sammansatt träningsdata ger felaktigheter i modellens resultat i produktion. Det är viktigt att säkerställa att de data som samlas in för ändamålet är tillförlitliga, korrekta och speglar de data som AI/ML-systemet ska behandla för att ge önskat resultat i produktion. Kvaliteten på, och kunskapen om, de data som används avgör hur bra AI/ML-systemet blir. Används felaktiga eller missvisande data resulterar det i felaktiga utfall eller förutsägelser som är irrelevanta. Det är viktigt att data inte är manipulerat eller på annat sätt innehåller felaktigheter.

Mängden data har också betydelse för resultatet. I vissa sammanhang kan mängden träningsdata vara avgörande för hur bra modellen presterar. Träningsdata kan utgöras av syntetiskt framställda data (avsnitt 4.5), men kan även komma från t.ex. myndighetens verksamhetssystem (se avsnitt 4.4 om produktionsdata). Det väsentliga blir således att

[9] https://beta.dataportal.se/offentligai/ai-i-praktiken-en-guide/datahantering

bedöma vad träningsdata består av. Om datamängden innehåller personuppgifter måste dataskyddsförordningens principer om bl.a. ändamålsbegränsning, uppgiftsminimering och proportionalitet beaktas, liksom myndighetsspecifika regler i registerförfattningar. Bedriver organisationen säkerhetskänslig verksamhet behöver detta beaktas särskilt.

4.2 Valideringsdata

Valideringsdata är data som används för att utvärdera hur väl modellen uppfyller de krav som ställts på modellen, exempelvis analysera och kontrollera noggrannheten. Vanligen sker träningssteget och valideringssteget iterativt inom utvecklingsprocessen, med en viss upprepning. Träningsdata och valideringsdata är vanligen samma dataset som kan delas i 80 % träningsdata och 20% valideringsdata.

Ibland används begreppet verifieringsdata synonymt med valideringsdata. Verifieringsdata används också ibland synonymt med testdata (se avsnitt 4.3). Detta begrepp kan därmed leda till missförstånd om vad som avses. I rapporten används därför begreppet valideringsdata.

4.3 Testdata

Testdata är data som används för att tillhandahålla en oberoende utvärdering av den tränade och validerade modellen för att bekräfta hur väl modellen presterar inför produktionssättning, dvs. *efter* att den har tränats färdigt och ska användas på nya data som inte använts i träningen. Testdata är en separat datamängd som inte använts under träning eller validering. Genom att använda testdata kan man bedöma modellens generalisering och hur väl den kan göra förutsägelser på nya, otestade exempel. Det är viktigt att testdata är representativt för de exempel modellen kommer att ställas inför i verkliga tillämpningar. Testdata kan således bestå av produktionsdata (se avsnitt 4.4). Ofta kommer testdata därmed att innehålla personuppgifter varvid myndighetens registerlagstiftning och dataskyddsförordningen måste beaktas.

4.4 Produktionsdata

När myndigheter ska ta fram ett AI/ML-system som ska kunna användas i verksamheten behöver modellen ibland tränas på verksamhetsdata. Sådan data kallas också produktionsdata eller skarpa data. Ibland används också begreppet historiska data, som avser information från avslutade ärenden eller fastslagen information som inte längre uppdateras. Ofta handlar det om information ur myndighetens verksamhetssystem, vilken ofta innefattar personuppgifter och dataskyddsreglerna därmed blir tillämpliga.

4.5 Syntetiska data

Syntetiska data är data som är framställda, dvs. fabricerade istället för verkliga data. Syntetiska data används ofta för träning och test när verkliga data är otillgängliga, bristfälliga, eller för att undvika integritets- och sekretessproblem. Syntetiska data kan skapas för att efterlikna egenskaperna och variationerna i verkliga data.

4.6 Annotering

Om det inte redan finns historiska data som kan utgöra ett facit behöver ett facit istället skapas. Denna process kallas annotering, där träningsdata tillförs ett facit med korrekta svar (se även avsnitt 10.2.1). Facit tillsammans med annan data för inlärning utgör ett komplett dataset för utveckling av maskininlärning.

4.7 Metadata

Metadata är data som beskriver egenskaper eller information om andra data som används i AI/ML-processen. Detta kan inkludera information om datasetets struktur, såsom antal datapunkter, antal och typer av egenskapsvariabler, ursprung och skapelsedatum för data, samt information om hur data har behandlats eller transformeras under förbearbetningen. Metadata används för att underlätta förståelsen, hanteringen och spårningen av data under AI/ML-processen.

4.8 Strukturerade data och ostrukturerade data

Data kan delas in i strukturerade respektive ostrukturerade data och kan även vara semi-strukturerade. Strukturerade data är organiserade i definierade fält och kolumner, t.ex. i databastabeller eller kalkylblad. Ostrukturerade data, såsom text, bilder, video eller ljud har inte en fast struktur eller format. Ostrukturerade data kräver ofta mer avancerade analysmetoder till skillnad från strukturerade data. Tekniker för att omvandla ostrukturerade data utvecklas i snabb takt.

4.9 Tidsbaserade data och geospatiala data

Tidsbaserade data är data som samlas in över tid, t.ex. loggfiler, data från sensorer, sekvenser eller tidserier. Geospatiala data har en geografisk aspekt, t.ex. koordinater (kartor).

4.10 Egenskapsvariabler (features)

Egenskapsvariabler (features), även kallat egenskaper, funktioner eller attribut, används för att beskriva varje dataexempel. De representerar den information som modellen använder för att lära sig mönster och göra förutsägelser eller beslut. Egenskapsvariabler kan vara av olika slag, såsom numeriska (t.ex. ålder, temperatur), kategoriska (t.ex. kön, färg) eller binära (t.ex. närvaro/avsaknad av ett visst attribut).

4.11 Målvaribler (targets)

Målvariabler är de variabler som AI/ML-modellen försöker förutsäga eller klassificera. I övervakad inlärning (se avsnitt 10.2.1) är målvariabler de kända utgångarna som modellen tränas att replikera. Vid oövervakad inlärning är det inte bestämt vilket samband som man letar efter och därmed finns inte heller målvariabler. Målvariabler representeras också i dataset, ofta som en separat kolumn, där varje rad motsvarar målvärdet för motsvarande egenskapsvariabler.

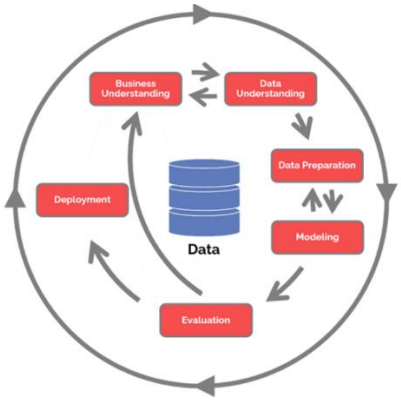
4.12 Hyperparametrar

Hyperparametrar är parametrar för den lärande algoritmen, som definierar strukturen och inställningarna för hur träningen av en AI/ML-modell ska gå till. Hyperparametrar ställs in innan träningen påbörjas och justeras ofta iterativt under träning- och valideringsfasen, genom att prova olika kombinationer av hyperparametrar för att hitta den version av modell som uppfyller uppsatta krav, se avsnitt 10.5. Exempel på en hyperparameter är antalet dolda lager (hidden layers) i ett neuralt nätverk.

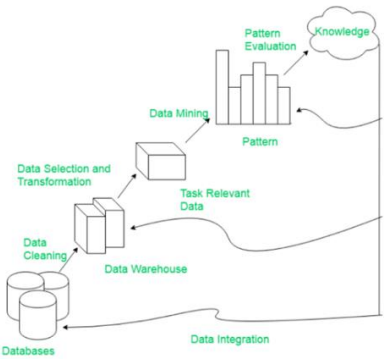
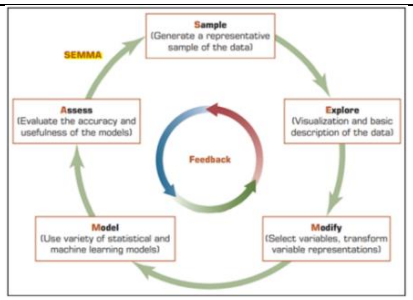
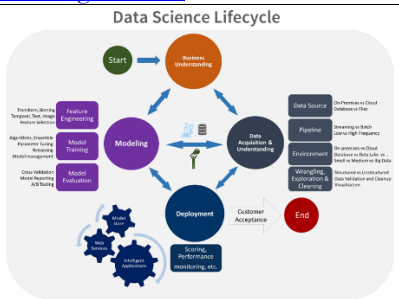
5. Exempel på befintliga utvecklingsprocesser

Som huvudprincip behöver en myndighet en särskild process för AI/ML-utveckling. Många gånger kommer inte befintlig process för traditionell utveckling av verksamheten vara tillräcklig, utan denna behöver utökas eller se ut på ett annat sätt. Riskerna är annars att det uppstår både effekts- och tidsförluster då väsentliga steg hoppas över eller att det uppstår brister i kommunikation och förståelse. Arbetet med utveckling av AI/ML handlar till stor del om en god förmåga till kommunikation, bl.a. att formulera gemensamma mål och skapa en delad förståelse av dem. En utvecklingsprocess (benämns även arbetssätt eller designmetod) för AI/ML ger stöd till att skapa förståelse mellan de olika roller som ingår i utvecklingsarbetet. Processen underlättar kommunikationen genom en strukturerad beskrivning, ofta visuell, av olika steg där begrepp definieras och används på ett gemensamt sätt. En gemensam och väl definierad process är också en förutsättning för att skala upp utvecklingen av AI/ML, t.ex. att utöka med fler utvecklingsteam inom organisationen.

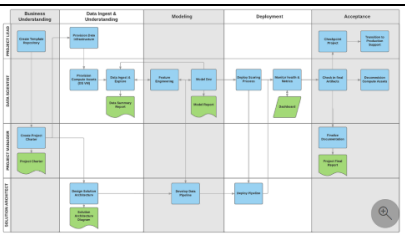
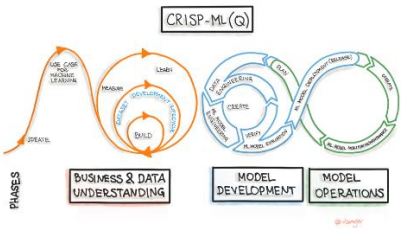
I detta avsnitt beskrivs några exempel på befintliga utvecklingsprocesser. Fokus för de beskrivna utvecklingsprocesserna ligger i första hand på arbetet som är kopplat till AI/ML och till mindre del annat utvecklingsarbete. Processerna ska ses som just exemplifieringar som kan ge inspiration och läsaren kan med fördel fördjupa sig i de beskrivna, eller andra, processer för att undersöka om och hur processerna kan stödja den egna organisationens utvecklingsarbete. Processerna beskrivs kortfattat med för- och nackdelar.

Namn	Bild	Fördelar	Nackdelar
CRISP DM (Cross Industry Standard Process for Data Mining) En allmänt använd process för utveckling av AI/ML. Består av sex steg: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, och Deployment.	 https://www.datascience-pm.com/crisp-dm-2/	Anses täcka alla aspekter av ett utvecklingsprojekt, från att förstå affärs-problem till att använda den slutliga modellen.[10] En styrka är fokus på starten med verksamhetsförståelse och att processen ger stöd i den delen. Många roller i utvecklings-arbetet kommer känna igen sig i CRISP-DM och dess steg.	Processen anses vara dokumentationstung och sakna ramverk för samarbete samt sakna steg för arbetet efter att modellen satts i produktion -”post-deployment”. Beskrivs ibland som omodern och att processen inte tillräckligt väl kopplar till ett agilt arbetssätt (jfr med CRISP-ML(Q)).

[10] Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical application of knowledge discovery and data mining (pp. 29-39).

KDD (Knowledge Discovery in Databases) En process för att identifiera användbar kunskap från en samling av data. Kombinerar tekniker från statistik, maskininlärning, databaser och domän-kunskap. Fokuserar mycket på det förutsättnings-skapande arbetet såsom förberedelser av data. Utgångspunkten för processen är data snarare än verksamhetsbehov jfr med CRISP-DM och CRISP-ML(Q). KDD fokuserar mer på upptäckt av kunskap än utveckling av AI/ML modeller.[11]	 https://www.geeksforgeeks.org/kdd-process-in-data-mining/	Förbättrat beslutsfattande, ökad effektivitet genom hantering av data för analys, förbättrad kundservice, identifiering av risker och prognostisering.[12]	Då processen har ett datacentriskt angreppssätt kan risker uppstå kopplat till hantering av personuppgifter. Hanteras processen fel och domän-kunskapen inte är med längs hela vägen kan det innebära att kunskapen, dvs. det som ska undersökas eller lösas, definieras sent i processen. Det kan också göra det svårt att avgränsa vilket syfte som den specifika analysen avser och således hur hantering av personuppgifter kan avgränsas.
SEMMA (SAS Enterprise Miner Methodology) En process för datautvinning och datavetenskapsprojekt som utvecklats av SAS. Den består av fem steg: Sample, Explore, Modify, Model och Assess.[13]	 https://sis.binus.ac.id/2021/09/30/data-mining-semma/	Är generell och kan användas för att stödja utveckling kopplat till olika typer av verksamhetsproblem.	Saknas fokus på idégenerering och verksamhets-förståelse. Användningen av processen kan minska över tid (författaren till artikeln menar dock att underlagen ska tolkas med viss försiktighet).[14]
TDSP (Team Data Science Process) Utvecklad av Microsoft för datavetenskapliga projekt (kopplad till Azure). Den består av sex steg: Business Understanding, Data Understanding, Data Preparation, Modeling, Deployment, och	 https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview	En mer komplett process jämfört med t.ex. OSEMN, SEMMA och KDD. Ansluter till ett agilt och teambaserat arbetssätt. Här definieras också roller som behövs kopplat till ett utvecklingsarbete och var rollerna kommer in i processen, något som saknas i de andra exemplifierade processerna.	Lägger lite för stor vikt på leverans och produktifiering och mindre på utforskande och forskningsinriktade projekt.

[11] Ref: Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.
[12] https://www.geeksforgeeks.org/kdd-process-in-data-mining/
[13] https://www.datascience-pm.com/semma/
[14] https://www.datascience-pm.com/semma/

Customer Acceptance. TDSP fokuserar på teamaspekten av data science och hjälper teammedlemmar att bli medvetna om affärsproblem och hur data kommer att användas för att lösa det.[15]	 https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview		
CRISP-ML(Q) (Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance methodology) Bygger på CRISP-DM, men där steg som idégenerering och övervakning av modeller lagts till. Processen innehåller också tydligare steg för arbete med data vilka syftar till att säkerställa datas kvalité inför modellering. Processen knyter också an till mer moderna koncept som ML Ops.[16]	 https://ml-ops.org/content/crisp-ml	När man arbetar med data som inte är välstrukturerade bidrar processens steg för hantering av dataset. Det här är något som också bidrar in i nästa steg kring modell-utveckling. Det cirkulära flödet visar tydligt att processen är iterativ och att stegen kan behöva göras mer än en gång. Arbetet kopplat till AI/ML och dess ökande användning gör att behovet av ML-Ops är något som snabbt växer. Förvaltning av AI/ML modeller skiljer sig från traditionell mjukvaruutveckling där ML-Ops för samman dessa discipliner.	Beskriver inte de olika rollerna inom processen lika väl som t.ex. TDSP.
OSEMN (Obtain, Scrub, Explore, Model and iNterpret) Är en akronym för de fem steg som ingår i datavetenskaps-projekt. Dessa steg hjälper till att vägleda processen för datavetenskapliga projekt, från rådata till de slutliga resultaten.	 https://www.datascience-pm.com/osemn/	Enkel och lätthanterlig process som kan användas i alla datavetenskapliga projekt. Lätt att begripa och förbättrar snabbt kommunikationen mellan rollerna i utvecklingsprocessen. Ett alternativ för nybörjare eller vid mindre utvecklings-arbeten.[17]	Är i många fall för lättviktig, saknar djup och viktiga delar som verksamhets-förståelse, implementation och monitorering av modeller, vilket gör den mindre lämpad för att hantera ett komplett utvecklingsarbete. Saknar också koppling till ett mer agilt arbetssätt och aspekter kring roller och teamarbete.

[15] Microsoft. (2016). Team Data Science Process.
[16] https://ml-ops.org/content/crisp-ml
[17] https://www.datascience-pm.com/osemn/

Sammanfattningsvis anses CRISP-DM vara en omfattande metod som täcker alla aspekter av ett utvecklingsprojekt kopplat till AI/ML. KDD är en datacentrisk process som fokuserar på upptäckt av kunskap med ledning av data. SEMMA liknar CRISP-DM men har ett lite annat tillvägagångssätt och ordning på stegen, mindre fokus på idégenerering och verksamhetsförståelse. TDSP beskriver roller på ett bra sätt och fokuserar på teamaspekten av utvecklingsarbetet. OSEMN är en enkel och lättviktig process vilket gör att den är enkel att förstå och ta till sig. CRISP-ML(Q) bygger på CRISP-DM, men där steg som idégenerering och övervakning av modeller lagts till. Processen knyter också an till mer moderna koncept som ML Ops.[18] I processen beskrivs också stegen för kvalitetssäkring av data inför modellering.

En organisation som befinner sig i någon slags uppstartsfas eller vid mindre utvecklingsarbeten skulle mycket väl kunna använda sig av en enklare process, t.ex. OSEMN. Processen passar dock inte bra till större organisationer eller mer komplexa projekt och insikten om att organisationen ”växt ur” processen kommer uppstå ganska snabbt. Här behövs istället en mer heltäckande process och kanske en process som kopplar till ML Ops, där förvaltning av såväl modell som andra komponenter beskrivs – här är CRISP-ML(Q) och TDSP bättre exempel.

[18] ML Ops syftar på processer för att leverera och förvalta AI/ML modeller på ett säkert och effektivt sätt. ML Ops kan beskrivas som skärningspunkten mellan ren systemutveckling, utveckling av AI/ML och hantering av data. På samma sätt som med Dev Ops som är motsvarigheten inom systemutveckling så syftar tillvägagångssättet att öka automationen kring leverans av modeller, öka kvaliteten och samtidigt ta hänsyn till verksamhetsmässiga och regulatoriska krav. MLOps - Wikipedia

6. Generell utvecklingsprocess för AI/ML med fokus på datahantering

Som konstaterats i avsnitt 5 behöver en myndighet som huvudprincip en process för AI/ML-utveckling och en sådan process ger stöd till att skapa förståelse och underlätta kommunikationen. De utvecklingsprocesser som beskrivs i avsnitt 5 har mycket gemensamt, men har också sina olika styrkor och svagheter.

I detta avsnitt (6) beskrivs en generell utvecklingsprocess som utgår från det gemensamma i de processer som beskrivs i avsnitt 5, men med fokus på datahantering, eftersom data och behandling av data är centralt i offentlig verksamhet.

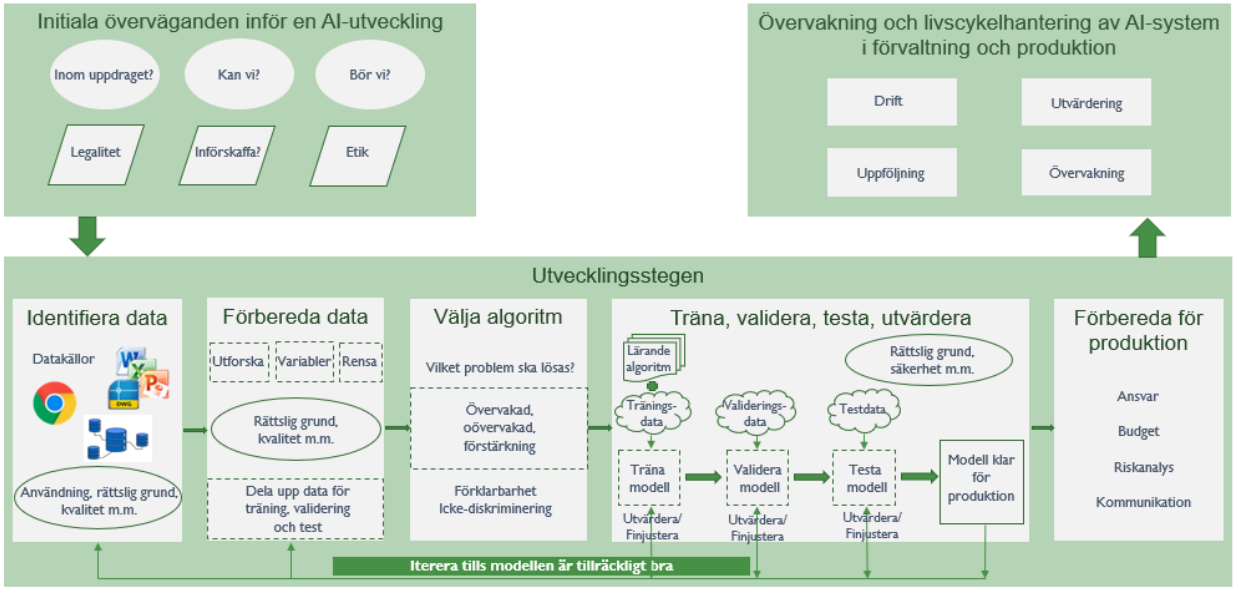
Vi har valt följande uppdelning av utvecklingsprocessen:

- Initiala överväganden inför en AI/ML-utveckling
- Beskrivning av de olika utvecklingsstegen
- Övervakning och livscykelhantering av AI/ML-system i produktion och förvaltning

Fokus ligger på en generell beskrivning av utvecklingsstegen och hur datahantering sker i de olika stegen. Initiala överväganden (avsnitt 6.1), övervakning och livscykelhantering då modellen är i produktion och förvaltning (avsnitt 6.3) samt olika roller som involveras vid en AI/ML-utveckling (avsnitt 6.4) beskrivs endast mycket kortfattat. Utvecklingsstegen beskrivs först övergripande i avsnitt 6.2 och en fördjupad beskrivning av dessa steg görs därefter i avsnitt 7–11.

En generell process för AI/ML-utveckling med datahantering i fokus kan övergripande illustreras enligt bilden nedan. Det är fråga om en iterativ process, dvs. där stegen återupprepas. Processtegen komma i olika ordning, t.ex. kan val av algoritm ske före eller efter identifiering av data.

En mer detaljerad beskrivning av processen framgår av nedanstående bild. Bildens olika rutor och dess beståndsdelar beskrivs närmare i avsnitt 6–11.



Bilden har fokus på utvecklingsprocessen för ett AI/ML-system. Övriga processer som verksamhetsutvecklingsprocessen, projektprocessen, arkitekturprocessen, eventuella förvaltningsprocesser m.m. redovisas inte i bilden. Processen är iterativ, om ett steg inte uppfyller förväntat resultat kan man gå tillbaka för att åtgärda eller förbättra. Stegen kan också förekomma i annan ordning. Vissa steg, exempelvis "Träna modell" är också iterativa i sig. Om det behövs uppdateringar eller förändringar efter att systemet satts i produktion startar man utvecklingsprocessen på nytt.

6.1 Initiala överväganden inför en AI/ML-utveckling

Tre grundläggande frågor[19] myndigheten bör ställa sig innan den påbörjar en AI/ML-utveckling är:

- Ligger detta inom vårt uppdrag?
- Kan vi göra detta?
- Bör vi göra detta?

All utveckling behöver ha sin grund i något slags verksamhetsbehov. Behovet måste vara kopplat till myndighetens uppdrag och således också till den rättsliga grund som myndigheten har för att utöva och bedriva sin verksamhet (legalitetsprincipen). Initialt behöver myndigheten ställa sig frågan om den har laglig grund för det som är tänkt att göras. Nyttor och effekter med utvecklingen behöver bedömas i förhållande till de grundläggande principer som myndigheten ska efterleva, se avsnitt 1. Det är viktigt med en tydlig målformulering om vilket problem AI/ML-systemet är tänkt att lösa. Man behöver även ställa sig frågan om AI/ML är rätt lösning för utvecklingen, dvs. vid utvecklingen bedöma för- och nackdelar med AI/ML-lösning jämfört med traditionell it-lösning eller manuell lösning. Myndigheten behöver göra en första risk- och konsekvensanalys.

Myndigheten behöver ta ställning till om den kan genomföra utvecklingen tekniskt. Har myndigheten kompetens för utvecklingen och finns det andra förutsättningar för ett genomförande? Det finns också många gånger ett beroende till öppen källkod (open source) vid AI/ML-utveckling som verksamheten behöver ha insikt om. Myndigheten behöver överväga egenutveckling eller om det istället finns anledning att införskaffa tjänsten. Förutsättningar för införskaffande beskrivs bl.a. i eSams rapport *Samverkan kring tillämpad AI*.[20]

Därtill behöver också frågan ställas om myndigheten *bör* genomföra utvecklingen. I det sammanhanget behöver myndigheten även ta hänsyn till etiska aspekter, se t.ex. frågor i eSams checklista *Juridik vid användning av AI*[21] om fri- och rättigheter och diskriminering. Vidare behöver man överväga om AI/ML-system som ska utvecklas i produktion kan påverka förtroendet hos allmänheten eller tvärtom, påverkas förtroendet av att inte använda AI:s möjligheter?

[19] Dessa övergripande frågor är ofta lika relevanta vid annan verksamhetsutveckling.
[20] ES2022-03 Rapport Samverkan kring tillämpad AI
[21] ES2022-08 Checklista Juridik vid användning av AI

Om myndigheten kommer fram till att den själv ska genomföra AI-utvecklingen behöver man också bedöma vilken it-miljö som bör användas för utvecklingen.[22] I detta behöver även de säkerhetsmässiga förutsättningarna bedömas.

6.2 Övergripande beskrivning av utvecklingsstegen

En utveckling av AI/ML innehåller ofta olika steg. Vanligen sker *itererin*g av stegen, dvs. flera av stegen återupprepas och en del steg är iterativa i sig. I denna rapport har vi valt en övergripande beskrivning med följande steg.[23]

- Identifiera och samla in data
- Förbereda data
- Välja algoritm, modell
- Träna, validera och testa samt utvärdera och finjustera
- Förbereda för produktion och förvaltning samt driftsätta

Data behandlas på olika sätt i de olika stegen. Vilken datahantering som sker i de olika stegen varierar mellan myndigheter och även mellan olika projekt inom samma myndighet. Utvecklingsstegen beskrivs närmare i avsnitt 7–11.

6.3 Övervakning och livscykelhantering av AI/ML-system i produktion och förvaltning

Med förvaltning menas att på ett flexibelt och förändringsvänligt sätt se till att det är en balans mellan produktens operativa förmåga och en iterativ utveckling av densamma.

Som beskrivs i avsnitt 11 om förberedelse inför produktion är det viktigt att ansvaret och befogenheterna för AI/ML-systemet är tydligt definierade och att det finns fastställda rutiner för hantering, dokumentation m.m. Bevarandefrågor, dvs. frågor kring uppkomst av allmänna handlingar och arkivhantering måste beaktas.

En driftsatt AI/ML-modell utvecklas iterativt hela tiden. Modellen behöver underhållas och övervakas för att fungera korrekt i förhållande till det planerade målområdet. Modellens funktion och beteende behöver följas upp över en längre tid och utvärderas och finjusteras kontinuerligt. Vid utvärderingen är det av stor vikt att ha med verksamhetsexpertis.

[22] Se t.ex. Sveriges dataportal, Offentlig AI, AI-guiden Infrastruktur och arkitektur: https://beta.dataportal.se/offentligai/ai-i-praktiken-en-guide/infrastruktur-och-arkitektur
[23] Jämför även beskrivning på Sveriges dataportal, Offentlig AI. https://beta.dataportal.se/offentligai/ai-i-praktiken-en-guide/kritiska-fragor-i-den-inledande-fasen

Generellt är det viktigt att de som ansvarar för AI/ML-modellen har tydliga samarbetsvägar med övriga organisationen eftersom arbetet med AI/ML-modellen är en ständig, iterativ process där flera delar av verksamheten ofta behöver konsulteras.

6.4 Roller i AI/ML-utveckling

Vid en AI/ML-utveckling involveras ett antal olika roller inom organisationen. Det finns ofta en uppdragsledare eller projektledare som driver uppdraget. Juristen bistår med bedömning om rättsliga förutsättningar för utvecklingen och användning av tänkt datamängd. En data scientist säkerställer att organisationen väljer rätt lösning för problemet, såsom val av algoritm, kontrollerar bias (se avsnitt 10.3) kopplat till verksamhetens krav och andra avvägningar. En data engineer står för bl.a. för informationsförsörjningen och systemutvecklaren bygger själva systemet där AI-komponenten enbart är en del av hela systemet. Arkivarien utreder informationens långsiktiga bevarandevärde och åtgärder för bevarande och gallring. Detta är bara några exempel.

I bilden nedan ges en övergripnade beskrivning om vilka roller som skulle kunna vara aktuella vid en AI/ML-utveckling. Bilden ska inte uppfattas som ett facit, vissa roller blir inte alltid aktuella och andra roller kan tillkomma.

Roller						
Initiala överväganden inför en AI/ML-utveckling	Övergripande beskrivning av utvecklingsstegen					Övervakning och livscykelhantering i produktion och förvaltning
Verksamhetsbehov Legalitet och etiska aspekter Införskaffa/egenutveckla, It-miljö och säkerhet	Identifiera och samla in data	Förbereda data	Välja algoritm/ modell	Träna, validera, testa, utvärdera, finjustera	Förbereda för produktion och driftsätta	Drift Uppföljning Utvärdering Övervakning
Verksamhetexpert Innovation Produktägare Verksamhetsarkitekt It-arkitekt It-infrastruktur Utvecklingsteam Data scientist Statistiker Jurist Inköpare Arkivarie	Verksamhetsexpert Verksamhetsarkitekt It-arkitekt Utvecklingsteam It-infrastruktur Informationssäker-hetsansvarig Informationsägare Data scientist Data engineer Statistiker Jurist	Verksamhetsexpert Verksamhetsarkitekt It-arkitekt Utvecklingsteam It-infrastruktur Informationssäker-hetsansvarig Informationsägare Data scientist Data engineer Statistiker Jurist	Data scientist Data engineer Utvecklingsteam	Verksamhetsexpert Utvecklingsteam It-infrastruktur Data scientist Data engineer Jurist	Förvaltning Verksamhetsexpert Utvecklingsteam Användare Arkivarie	Förvaltning Verksamhetsexpert Utvecklingsteam Data scientist Statistiker Användare

7. Identifiera och samla in data

För att kunna genomföra en AI-utveckling behöver myndigheten identifiera vilka data som finns tillgängliga, dvs. vilka olika datakällor som finns och vilken datamängd som kan erhållas från dessa datakällor. Dessutom behöver man göra en bedömning av om identifierade data får användas till det avsedda ändamålet, dvs. avsedd AI-utveckling. I vissa fall finns det tillgång till data som redan förprocessats, se avsnitt 8 om förberedelse av data. I andra fall handlar det om att samla in rådata från olika källor, såsom databaser, filer, sensorer, webbskrapning eller API:er, som efter insamling ofta behöver förberedas ytterligare innan datamängden kan användas för träning. Vid valet av data behöver myndigheten beakta vad AI/ML-systemet förväntas göra, dvs. vilket problem ska lösas. Exempelvis kan AI/ML-systemet lösa ett klassificeringsproblem (avgöra om det är en katt på bilden) eller förutsäga priset på en fastighet (regression), se mer om olika typer av maskininlärning i avsnitt 10.2. Även valet av typ av algoritm kan behöva beaktas, liksom vilket urval som kan behöva göras vid insamlingen eller om det kan hanteras senare vid förberedelse av data.[24]

Detta avsnitt tar upp några aspekter som behöver beaktas i samband med val av data.

7.1 Rättsliga, säkerhetsmässiga och etiska perspektiv

Myndigheten behöver säkerställa att vald data får användas på det sättet som är tänkt, utifrån rättsliga, säkerhetsmässiga och etiska perspektiv. Dessa frågor behöver således komma in i detta tidiga skede i processen. Datamängd och innehåll kan behöva anpassas för den tänkta AI-användningen.

Datamängd och datakälla behöver bedömas utifrån syftet med användningen och om datamängden får användas för det ändamålet. Är det fråga om personuppgifter behöver det finnas en rättslig grund för att få behandla datamängden (inklusive en rättslig grund för insamlingen i sig), de grundläggande principerna, såsom principerna om uppgiftsminimering och lagringsminimering, liksom kravet på korrekthet, behöver beaktas. Om det förekommer sekretessuppgifter behöver myndigheten säkerställa en korrekt hantering av dessa uppgifter. Det kan finnas begränsningar i avtal om hur datamängden får användas liksom upphovsrättsliga frågor. I vissa myndigheters registerförfattningar finns även sökbegränsningar som myndigheten behöver ta hänsyn till vid val av variabler. Bedriver organisationen säkerhetskänslig verksamhet behöver

[24] Används t.ex. neuralt faltningsnätverk (CNN (Convolutional Neural Network)) sker valet av variabler automatiskt i modellen och inget urval behövs inför användning av modellen. Urval kan dock ändå behövas i dessa fall pga. andra skäl t.ex. att vissa data inte får behandlas enligt dataskyddsbestämmelser.

detta särskilt beaktas. Därutöver behöver myndigheten ta hänsyn till krav på hantering av allmänna handlingar, åtgärder för långsiktigt bevarande respektive gallring.

Ofta är det en fördel om dessa överväganden kan ske för hela utvecklingsprocessen i samband med valet av data, dvs. att bedömningen inte bara avser insamlingssteget utan även senare lagring, behandling och bearbetning vid förberedelse- och träningssteget, se mer om dessa steg i avsnitt 8–11. I stegen tillkommer också frågor om transparens och förklarbarhet samt icke-diskriminering, se avsnitt 9.3.

En bra utgångspunkt är att använda eSams vägledningar och checklistor, exempelvis checklistan *Juridik för användning av AI*[25] och checklistan *Rättsliga förutsättningar i utvecklingsinsatser m.m.* som stöd.[26] Ytterligare stöd finns t.ex. i eSams promemoria *Stordataanalyser och datasjöar*[27] eller den guide som är publicerad på Offentlig AI, *AI-guiden Juridik och etik*.[28]

7.2 Relevant data

Ett AI/ML-system lär sig av de data som systemet tränas på. Vid träning av AI/ML-system är det därför viktigt att den datamängd som samlas in för ändamålet är tillförlitlig, korrekt och som så bra som möjligt representerar den datamängd datamodellen sen ska användas till. Det behövs för att modellen så korrekt som möjligt ska kunna identifiera, hitta mönster och förutspå resultatet avseende den datamängd den är tränad på. Om syftet är att träna en modell som ska känna igen bilder på katter behöver datamängden bestå av ett urval som avspeglar en sådan ”population” på ett representativt sätt. Ytterst handlar relevant data om rättssäkerhet, dvs. att datamängden (både rådata och från rådata generade data) som används är korrekt så att resultatet blir rättssäkert (t.ex. om resultatet ska användas som underlag för beslut).

7.3 Kvalitet på data

Kvaliteten på de data som används vid träning avgör hur bra modellen blir. Felaktiga eller missvisande data kan leda till felaktiga utfall eller förutsägelser. Datakällorna behöver vara av hög kvalitet, med noggranna, korrekta och uppdaterade data.

[25] ES2022-08 Checklista Juridik vid användning av AI.
[26] Checklista för jurister Introduktion i rättsliga förutsättningar i utvecklingsinsatser, version 2.0, eSam juni 2019.
[27] ES2023-2 PM Stordataanalyser och datasjöar begrepp och rättsliga förutsättningar.
[28] https://beta.dataportal.se/offentligai/ai-i-praktiken-en-guide/juridik-och-etik

7.4 Val av datakälla

Beroende på vilka data som behövs och vad de ska användas till finns det olika datakällor och de kan övergripande delas in i:

- Kommersiell datakälla, data som köps från en extern part, t.ex. ett företag
- Öppen datakälla, data som är offentlig även om immateriella rättigheten kan tillhöra t.ex. en juridisk person
- Inomstatlig datakälla, data som levereras av en eller flera myndigheter eller ett bolag under statlig kontroll

I de fall en förtränad modell köps in (se avsnitt 9.2) eller ett AI-system köps in färdigt, kan träning ha gjorts med data från en kommersiell datakälla där datamängden då ofta är en affärshemlighet och inte tillgänglig för köpande myndighet. Det kan också bli fråga om en federerad (samordnad) datakälla, där viss del av träningen är utförd på en okända träningsdata från kommersiell datakälla och en del på kända träningsdata.

Ofta kan det vara en fördel att använda *flera* datakällor för att undvika bias (se avsnitt 10.3) och förbättra AI/ML-systemets förmåga att hantera olika situationer och ämnen. En chattbot kan t.ex. använda en kommersiell eller öppen datakälla för generell språkförmåga som kompletteras med en inomstatlig datakälla med myndighetsspecifika språkdata. Detta inkluderar att använda data från olika geografiska områden, kulturer och perspektiv. Det sätt man väljer att förbereda data på, kan också påverka valet av datakälla, se avsnitt 8.

Det är viktigt att bedöma hur *tillförlitlig* en datakälla är, eftersom datamängden har en direkt påverkan på resultatet och hur korrekt modellen kommer prestera. Datamängden behöver undersökas så att den inte är manipulerad eller på annat sätt innehåller felaktigheter. Även källans rykte och auktoritet behöver kontrolleras inom sitt område. Vanligtvis har källor som regeringar, forskningsinstitutioner och branschorganisationer hög tillförlitlighet, men behöver bedömas i varje fall. Det kan också vara betydelsefullt att undersöka om källan förvaltas och i så fall hur. Om datakällan inte är möjlig att kontrollera kan det innebära att den inte kan användas.

De datakällor som används i utvecklingsprocessen bör dokumenteras (i vissa fall ett krav)[29] så att de är transparenta och spårbara. Detta bidrar till att fördela ansvar och bygga förtroende för funktion och resultat.

[29] Jfr dokumentationskrav i dataskyddsförordningen.

8. Förbereda data

När myndigheten har identifierat datakällorna och samlat in data för träning och validering behöver insamlade data oftast förberedas. Vilken förberedelse som krävs beror på vilken AI/ML-utveckling som är aktuell.

I förberedelsearbetet identifieras de variabler som ska användas, se avsnitt 4.10 och 4.11 om variabler. I vissa fall krävs förbehandling eller transformation av rådata för att extrahera meningsfulla variabler. I vissa sammanhang behöver man förbereda och bearbeta variablerna innan modellträningen kan göras (tidskrävande, men leder till högre förklarbarhet). I andra fall sker denna typ av bearbetning inuti själva modellen (djupinlärningsmodeller med lägre förklarbarhet). Förberedelsen kan vara manuell eller automatisk. Vid automatisk förberedelse hittar programmet själv hitta lämpliga egenskapsvariabler att använda i maskininlärningen. Automatisk förberedelse kräver stora datamängder och balanserade data, dvs. data som ger ett representativt snitt av verkligheten men som är objektiv, givet att metoden är pålitlig. Manuell förberedelse innebär att man väljer ut en egenskapsmängd och applicerar metoder för att välja egenskapsvariabler. Manuell förberedelse riskerar att vara mer subjektiv eftersom det sker ett urval av en egenskapsmängd som man sedan selekterar och extraherar från. Å andra sidan har man kontroll över de egenskaper som används.

Val av algoritm och modell kan också påverka hur förberedelsen behöver genomföras. Förberedelsesteget och steget att välja algoritm och modell är därmed beroende av varandra. Vilken förberedelse som krävs kan även påverkas av val av och urval i datakällorna.

För att kunna förbereda data krävs ofta en mycket bra verksamhetsförståelse för datamängdens innebörd. Liknande begrepp kan ha olika innebörd vilket kan leda till felaktigheter, så det är viktigt att säkerställa tillgången till verksamhetskompetens som kan tolka innebörden.

8.1 Rättsliga förutsättningar

Innan myndigheten påbörjar förberedelsen av data behöver den undersöka att det finns stöd och syfte för hanteringen. Många gånger har analysen kunnat göras i samband med insamling av datamängden och val av datakälla och kan då återanvändas i detta steg, se avsnitt 7 om identifiera data och påbörja datainsamling. Om datamängden till exempel innehåller personuppgifter kommer de olika förberedelserna av datamängden att utgöra behandlingar enligt dataskyddsbestämmelserna. I sådana fall behöver man göra en bedömning av om det finns stöd för sådana behandlingar. Om förberedelsen innebär att

nya dataset skapas t.ex. genom ett urval så kan man behöva fundera kring uppkomst av allmänna handlingar och förutsättningar för arkivering eller gallring av dessa efter avslutad utveckling. Se t.ex. eSams checklista *Juridik för användning av AI*[30]om vilka rättsliga frågor som behöver analyseras.

8.2 Dataförståelse och utforskande analys, identifiera variabler

Utforskande analyser syftar till att förstå datamängdens egenskaper, såsom fördelningar och kvalitet på variabler. Det omfattar att undersöka statistik och visualisera data för att t.ex. identifiera trender, anomalier och korrelationer mellan variabler. Det kan handla om att identifiera min- och maxvärden eller andelen saknade värden för enskilda variabler, klusteranalyser i syfte att organisera eller upptäcka grupperingar av datapunkter, statistiska tester eller regressionsanalyser för att hitta trender i data. Genom att undersöka data skapas en förståelse för vilka begränsningar data har (kvalitet, kompletthet, relevans, osv.) och vilka åtgärder som krävs inför modellträning. Med en utforskande analys kan man t.ex. upptäcka att vissa data saknas och behöver samlas in, att delar av data inte är relevant och behöver rensas eller att vissa variabler behöver transformeras eller på annat sätt bearbetas.

Ett vanligt exempel inom maskininlärning är en modell som ska förutspå marknadsvärdet på en fastighet. En sådan modell kan skapas med hjälp av data som beskriver historiska försäljningar av fastigheter. Försäljningspriset utgör då utfallsvariabeln *y* och oberoende variabler kan vara egenskaper som beskriver fastighetens storlek (boyta, biyta, tomtstorlek), belägenhet (kommun, avstånd till strand, tätort/landsbygd, osv.) och ålder (byggnadsår). Genom utforskande analyser kan man t.ex.:

1. **Undersöka samband eller mönster i data.** Ex: Försäljningspriser är (som förväntat) högre vid lägen nära strand och i storstadsområden jämfört med priser vid icke-strandnära lägen och på landsbygd.
2. **Upptäcka vilken databearbetning som krävs.** Ex: Försäljningspriser är skevt fördelade (en stor andel fastigheter säljs för låga priser och enbart ett fåtal säljs för höga priser) och kräver därför sannolikt någon typ av transformering (exempelvis logaritmisk transformering) om en linjär modell ska användas.
3. **Få idéer om nya/alternativa variabler att experimentera med.** Ex: Antalet försäljningar i vissa kommuner är mycket få och variabeln kommun kommer

[30] ES2022-08 Checklista Juridik vid användning av AI

därför att kräva någon typ av bearbetning (feature engineering), alternativt ersättas av en annan variabel (t.ex. län).

Att undersöka förväntade samband (punkt 1) är viktigt för att säkerställa att data är korrekt och relevant för den frågeställning modellen syftar till att besvara. Ofta upptäcker man oväntade mönster i data, som man då behöver förstå orsaken till. Kan det ha gått fel i insamlingen av data? Har felaktiga antaganden gjorts om vilken information datamängden innehåller eller finns det en naturlig förklaring till de mönster man ser? En tät dialog med verksamhetsexperter behövs för att reda ut de frågor som uppkommer och säkerställa att data är korrekt och relevant för ändamålet.

Ett annat syfte med utforskande analyser är att förstå egenskaper hos de data som modellen ska tränas och utvärderas på (punkt 2 och 3). Till exempel kan man upptäcka att vissa variabler innehåller en stor andel saknade värden eller avvikande datapunkter, att det finns linjära samband mellan vissa variabler (pris vs boyta) och icke-linjära samband mellan andra variabler (pris vs närhet till strand). Sådana egenskaper är viktiga att känna till, då de bland annat kan innebära att data behöver bearbetas på ett visst sätt eller att vissa typer av algoritmer är mer eller mindre lämpliga att använda. Utforskande analyser kan också hjälpa till att ge idéer till hur data bäst förbereds inför modellträning (så kallad feature engineering). Istället för att använda två variabler där den ena anger storleken på fastigheten och den andra den genomsnittliga storleken på fastigheter i området, så kanske en ny variabel som anger differensen mellan dessa två (dvs. hur mycket större fastigheten är jämfört med den genomsnittliga storleken i området) är ännu bättre? Genom att analysera och förstå egenskaper hos data kan man få idéer på alternativa sätt att förbereda data och testa dem i efterföljande modellexperiment (under modellutvecklingsfasen). Det bidrar i slutändan till en bättre produkt i termer av högre precision på modellen, snabbare träning, minimalt med underhåll (t.ex. genom att hålla nere antalet variabler), ökad förklarbarhet, eller annat.

8.3 Rensa data

Det är viktigt att säkerställa att data håller hög kvalitet och korrekthet för att undvika att modellen dras med felaktig information. Därför kan data behöva ”rensas” genom att korrigera felaktiga värden, hantera saknade värden och ta bort dubbletter. Exempelvis kan data för bildanalys innehålla bilder med oskärpa eller som har dålig färg och därför kan behöva tas bort. Det kan också handla om irrelevant data, t.ex. vissa kolumner i en tabell.

En överdriven rensning kan leda till att datasetet inte ligger tillräckligt nära verkliga förhållanden och att modellen inte klarar av detta när den sedan utsätts för verkliga data.

Exempelvis finns en risk att modellen inte klarar av att hantera text med stavfel om alla stavfel rättas upp. Det behöver vara en balans i hur mycket som rensas bort.

8.4 Integrera data

Om data kommer från olika källor kan de behöva kombineras och integreras i en gemensam struktur. Det kan innebära att sammanfoga data från olika tabeller, matcha kolumner och hantera inkonsekvenser mellan datakällorna. Om datamängder kommer från t.ex. två olika myndigheter behöver det finnas gemensamma nycklar för att förstå hur dessa datamängder förhåller sig till varandra. Till exempel kan en fastighetsbeteckning vara en sådan nyckel. Det behöver också säkerställas att datamängderna uttrycker nyckeln i samma format, till exempel om det finns med en sekelsiffra i en datamängd men inte i den andra så måste det läggas till.

8.5 Transformera data

För att förbättra datasetets användbarhet för maskininlärningsmodeller kan datasetet behöva transformeras. Det kan omfatta att normalisera eller skala data, dvs. sätta data på samma skala. Exempelvis om en variabel har ett spann på 1–1 miljon och en annan variabel har ja–nej (dvs. 0–1), behöver de anpassas till samma skala, annars uppstår en skevhet. Det kan också innebära att konvertera kategoriska variabler till numeriska, till exempel om det finns tre svarsalternativ så görs de om till 0,1,2. Man kan även behöva skapa nya variabler och utföra dimensionell reduktion, dvs. minska komplexiteten i modellen och öka stabiliteten.[31]

8.6 Hantera obalanserade data

Obalanserade data kan vara ett problem då modellen blir bättre på att predicera vissa grupper jämfört med andra (dvs. risk för bias). Om data är obalanserade, med vissa klasser som är över- eller underrepresenterade, kan det vara nödvändigt att använda tekniker som oversampling, undersampling eller generering av syntetiska data[32] för att balansera klassfördelningen. Exempelvis om det finns en obalans mellan män och kvinnor i datamängden kan detta behöva balanseras för att vara mer representativt. Ibland talas om hål (missing values) i datamängderna som behöver hanteras, t.ex. genom att ta bort kolumner som innehåller dessa värden eller ersätta hål med skattningar (det senare bör göras med försiktighet för att undvika bias).

[31] Detta kan t.ex. göras med PCA (Principal component analysis).
[32] Detta kan t.ex. göras med SMOTE (Synthetic Minority Oversampling Technique).

8.7 Pseudoymisering och anonymisering

Om data innehåller personuppgifter, behöver de behandlas på ett sätt som skyddar individernas integritet, i enlighet med gällande dataskyddslagstiftning (till skillnad från syntetiska data som inte har detta krav, se avsnitt 4.5). En åtgärd kan t.ex. vara att pseudonymisera eller anonymisera uppgiften. Sådan behandling kräver i sig en rättslig grund enligt dataskyddsförordningen, vilket behöver analyseras innan åtgärden genomförs. Läs mer om förutsättningar för pseudonymisering och anonymisering i eSams vägledning *Pseudonymisering av personuppgifter*.[33]

8.8 Dela upp data

När bedömningen är att valda data innehåller en tillräckligt bra mängd, struktur och kvalitet delas data in i tränings-och valideringsdata respektive testdata. Träningsdata och valideringsdata är vanligen samma dataset som kan delas i 80 % träningsdata och 20% valideringsdata. Testdata är ett separat dataset som i möjligaste mån ska återspegla de data som modellen ska appliceras på vid tillämpningen av modellen i produktion, dvs. är mer anpassat till verkliga förhållanden. Se mer om de olika dataseten i avsnitt 4.

Uppdelningen görs för att säkerställa att modellen tränas, valideras och testas på separata data, vilket gör det möjligt att mäta modellens förmåga att generalisera till nya, osedda exempel. Genom att iterativt variera vilka datapunkter som ingår i träningsdata och valideringsdata, kan man säkerställa att modellen ger ett stabilt resultat (robusthet).

[33] ES2022-01 Vägledning Pseudonymisering av personuppgifter.

9. Välja algoritm och modell

9.1 Val av algoritm

Lärande algoritm(er) (även kallat *otränad modell*) är en serie instruktioner som beskriver hur, med vad och i vilken ordning en uppgift ska lösas, i det här fallet inlärningsprocessen. Det handlar om att träna ett system med data för att upptäcka och lära sig regler för att lösa en uppgift, utan att systemet har programmerats på traditionellt sätt (villkorssatser/if-then-else), med regler för just den uppgiften. Kort kan en lärande algoritm sägas bestå av

- en *representerande funktion* (t.ex. en linjär funktion som anger den typ av modell som används),
- en *definition av korrekthet* (t.ex. att en låg genomsnittlig skillnad mellan de värden som genereras av modellen och de värden som framgår av träningsdata ger en högre korrekthet) och
- en *optimeringsmetod* (dvs. en metod för att justera modellen så att graden av korrekthet ökar).[34]

Det finns ett flertal algoritmer som är bra på att utföra olika saker som taligenkänning, analys av bilder, analys av numeriska data, prognoser, förutsägelser m.m.[35]

Övergripande kan man säga att valet av algoritm görs utifrån vilket problem man vill lösa. Det är viktigt att det finns en tydlig problemformulering. Ofta talas om tre huvudtyper av maskininlärning:

- Övervakad inlärning
- Oövervakad inlärning
- Förstärkningsinlärning[36]

Gränsdragningarna dem emellan kan vara något oklara, t.ex. förekommer semiövervakad inlärning som är en kombination av övervakad och oövervakad maskininlärning. I avsnitt 10.2 ges en närmare förklaring av de olika typerna.

Vid val av algoritm(er) behöver generellt olika bedömningar ske för att säkerställa att algoritmen är lämplig för uppgiften och fyller organisationens behov. Förutom

[34] Guttag, J.V. (2016) Introduction to Computation and Programming Using Python: With Application to Understanding. MIT.
[35] https://learn.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet?view=azureml-api-2#download-machine-learning-algorithm-cheat-sheet
[36] Benämns ibland även återkopplingsinlärning.

nedanstående centrala aspekter att överväga kan organisationen ha en lokal policy eller andra förutsättningar som behöver tas i beaktande.

- **Mål och syfte:** Förstå organisationens mål med AI/ML-systemet och vilka problem det ska lösa. Detta hjälper till att identifiera vilka algoritmer som är bäst lämpade för ändamålet.
- **Typ av data:** Överväg vilka typer av data algoritmen ska användas för, såsom text, bilder, ljud eller numeriska data. Olika algoritmer är specialiserade på att hantera olika typer av data.
- **Datakvalitet och kvantitet:** Bedöm storleken och kvaliteten på tillgängliga träningsdata, valideringsdata och testdata. Algoritmens prestanda och korrekthet kan påverkas av datamängd och kvalitet.
- **Träningstid:** Utvärdera hur lång tid det tar att träna algoritmen och om det är acceptabelt för organisationen.
- **Prestanda:** Bedöm vilken prestanda som krävs av algoritmen, exempelvis hur snabbt den ska kunna analysera stora datamängder och ge resultat.
- **Variabler:** Ta hänsyn till antalet och komplexiteten av de variabler som behöver extraheras och analyseras från data.
- **Mätning av korrekthet:** Bestäm hur korrektheten i algoritmens bedömningar ska mätas och vilka krav som ställs på korrektheten i produktion.
- **Förklarbarhet:** Överväg om det är viktigt att eventuella beslut av AI/ML-systemet är förklarbara och lättförståeliga för användare och intressenter (se avsnitt 9.3 om förklarbarhet).
- **Begränsningar och utmaningar:** Identifiera eventuella tekniska, juridiska eller etiska begränsningar som kan påverka valet av algoritm, såsom dataskydd, integritet och säkerhet.
- **Erfarenhet och resurser:** Ta hänsyn till organisationens erfarenhet och tillgång till teknisk expertis samt tillgången på resurser för att utveckla, implementera och underhålla AI/ML-systemet.

Genom att överväga dessa aspekter kan organisationen fatta informerade beslut vid val av algoritmer och säkerställa att AI/ML-systemet är anpassat till organisationens specifika behov och förutsättningar.

9.2 Modell

En modell är resultatet av träningsprocessen (även kallad *tränad modell*). En modell kan sägas bestå av en prediktiv funktion. Den tränade modellens förmåga avgör resultatet när den används på en specifik datamängd i produktion. Den modell som blir resultatet av

träningen och som har tillräcklig prestanda och kvalitet är den som sedan används i produktion.

Det är inte alltid som organisationen själv tränar upp modellen. Man kan köpa eller på annat sätt erhålla en modell som är förtränad, t.ex. modeller med öppen källkod (open source). Modeller kan ingå i ett köpt AI-system och då antingen tränas eller inte tränas vidare. Vid förtränade modeller är det viktigt att bedöma om träningen är tillräckligt bra för modellens ändamål, att den inte innehåller bias eller andra utmaningar av juridisk eller säkerhetsmässig karaktär m.m. Ett exempel på sådan modell är t.ex. KB-BERT som förtränats av Kungliga biblioteket och som kan tränas vidare[37] av t.ex. en myndighet (myndighetsspecifik domänkunskap) och sedan användas av exempelvis kommuner, länsstyrelser eller andra aktörer.

9.3 Förklarbarhet, tillförlitlighet och icke-diskriminering

9.3.1 Förklarbarhet

Beroende på vad modellen ska användas till kan det finnas olika nivåer för krav på förklarbarhet av hur modellen kommer fram till sitt resultat, se t.ex. eSams checklista *Juridik för användning av AI*[38] om krav enligt dataskyddsförordningen, förvaltningslagen och i förslag till AI-förordning. Exempelvis finns krav vid förekomsten av automatiserat beslutsfattande, inbegripet profilering, i artikel 13.2 (f) dataskyddsförordningen. I de fallen ska meningsfull information lämnas, till den vars personuppgifter har behandlats, om logiken bakom samt betydelsen och de förutsedda följderna av sådan behandling för den registrerade. Ett annat exempel är 32 § förvaltningslagen som ställer krav på att myndigheten klart och tydligt kan redogöra på vilka grunder ett beslut har fattats.

Vissa typer av algoritmer är enklare att förklara (t.ex. linjär regression och beslutsträd), medan andra mer komplexa algoritmer (t.ex. djupa neurala nätverk) är svårare att förstå och förklara. Komplexa modeller med låg förklarhet resulterar ofta i noggrannare förutsägelser, varför valet av algoritmen ofta innebär en avvägning mellan hög förklarbarhet och hög träffsäkerhet.

Ibland talas det om generell förklarbarhet respektive individuell förklarbarhet.[39]

Generell förklarbarhet avser att man förklarar hur modellen fungerar. Ett sätt att åstadkomma förklarbarhet är att välja en lärande algoritm som är mindre tillförlitlig men mer förklarbar. Detta är ett alternativ när kravet på förklarbarhet är väldigt högt. Linjär

[37] Kallas ibland fine tuning.
[38] ES2022-08 Checklista Juridik vid användning av AI.
[39] Bibal, A., Lognoul, M., de Streel, A. et al. Legal requirements on explainability in machine learning. Artif Intell Law 29, 149–169 (2021).

regression och Random Forest Classifiers är exempel på sådana typer av algoritmer som generar modeller som är matematiskt lättare att förstå. Vanligen försöker man dock först använda mindre ingripande metoder, t.ex. att justera algoritmen för att reducera antalet egenskapsvariabler, t.ex. med hjälp av den s.k. Lasso-funktionen.[40] Detta innebär dock att man offrar viss träffsäkerhet. För att inte behöva ändra i algoritmen kan man istället använda metoder där man, utan att justera algoritmen, återger hur modellen fungerar. Detta är metoder som t.ex. kan användas för att uppfylla kravet enligt artikel 13.2 (f) dataskyddsförordningen. En metod är att skapa en *approximation* av den typ av modell som algoritmen genererar som är matematiskt enklare att förstå och med liknande prediktiv förmåga.[41] En annan metod är att ange alla egenskapsvariabler eller de som har störst betydelse för förutsägelserna, utan att ange hur modellen i övrigt ser ut. Som ett tredje alternativ kan nämnas *visualiseringar* eller andra beskrivningar som inte uttrycks på ett matematiskt sätt.

Individuell förklarbarhet[42] betyder att man inte förklarar modellen för hela tillämpningen utan endast hur den fungerar för en mindre mängd indata eller för en specifik indata. Denna förmåga till förklarbarhet kan användas för att uppfylla kravet på motivering av beslut enligt 32 § förvaltningslagen. LIME[43] är t.ex. en metod för detta. Andra metoder är att antingen *extrahera alla egenskapsvariabler* som har haft betydelse för tillämpningen för en enskild datapunkt eller, för de fall att de är för många, endast de som är av *störst betydelse,* t.ex. genom s.k. *saliency maps* eller *Grad-CAM* som används för att analysera neurala faltningsnätverk (CNN (Convolutional Neural Networks)).

9.3.2 Tillförlitlighet och icke-diskriminering

Myndigheten behöver även bedöma *tillförlitlighet* och *icke-diskriminering*. Med *tillförlitlighet* menas att egenskapsvariablerna t.ex. ger en hög andel korrekta klassificeringar och en låg andel felaktiga klassificeringar för *en* grupp utan jämförelse med vad som gäller för en annan grupp. Med *icke-diskriminering* (fairness) menas istället att varken *direkt* eller *indirekt* diskriminering förekommer, jämför 1 kap. 4 § punkterna 1–2 diskrimineringslagen (2008:567), även om grupperna som omfattas inte ska ses som begränsade till de som räknas upp i den bestämmelsen. Att undvika direkt diskriminering kan vara att avstå från att använda sig av t.ex. hudfärg som en egenskapsvariabel, medan indirekt diskriminering undviks genom att man sorterar ut egenskapsvariabler som dels saknar någon berättigad funktion, dels är partiska.

[40] Lasso-funktionen används ofta för att undvika överinlärning för att förbättra modellens prediktiva förmåga, se t.ex. Cady F. (2017) The Data Science Handbook. Wiley. I detta sammanhang används dock Lasso-funktionens andra fördel, att öka förklarbarheten. Eftersom överträning alltid ska undvikas talar vi här om att öka förklarbarheten *på bekostnad* av träffsäkerheten.
[41] T.ex. med SLIM (Supersparse LinearIntegerModels).
[42] Guidelines on Automated individual decision-making and Profiling, p. 25.
[43] Local InterpretableModel-AgnosticExplanations.

Ett problem är att icke-diskriminering kan definieras på olika sätt, vilket kan resultera i olika svar på frågan om det förekommer diskriminering. Att definiera icke-diskriminering utifrån flera mått som samtliga ska vara uppfyllda har visat sig omöjligt av matematiska skäl då det kommer att ge ett skevt resultat förutom i mycket speciella undantagsfall där resultatet är helt jämt fördelat.[44]

Så var fallet med programmet COMPAS (Correctional Offender Managemant Profiling for Alternative Sanctions), som användes för att klassificera personer som högrisk respektive lågrisk för att återfalla i brott. Denna bedömning används i Wisconsin i USA, avseende ett brott som begicks 2013, som en faktor för att avgöra vilket straff den skyldige skulle få. Här resonerades kring tre mått för icke-diskriminering, att icke-diskriminering avseende två kategorier av personer kunde definieras som

- kalibrerad modell (sanna klassificerade): Andelen som har klassificerats som högrisk och som faktiskt återfallit i brott är ungefär densamma oavsett om man studerar personer från den ena eller den andra kategorin.
- modell som ger samma andel falska positiva: Andelen som felaktigt klassificeras som högrisk är ungefär densamma för de två kategorierna.
- modell som ger samma andel falska negativa: Andelen som felaktigt klassificeras som lågrisk är ungefär densamma för de två kategorierna.

(Se avsnitt 10.5 om förklaring av positiva och negativa.)

Programmet hade andelen sanna klassificerade i 63 % av fallen avseende mörkhyade och i 59 % av fallen avseende ljushyade (dvs. ungefär lika fördelat). Andelen icke-återfallande mörkhyade som felaktigt klassificerades som återfallande var 45 % medan motsvarande siffra för ljushyade endast var 23 %. Samma skevhet till mörkhyades nackdel gällde för andelen mörkhyade respektive ljushyade ur kategorin återfallande som klassificerats som ej återfallande.[45] Det vill säga, det första måttet gav att programmet inte var diskriminerade, medan de två andra måtten påvisade förekomst av diskriminering.

Exemplet visar att när man mäter förekomst att diskriminering måste det tas i beaktande att det kan finnas flera intuitivt giltiga definitioner av icke-diskriminering och att definitionerna inte kommer att peka åt samma håll. Det vill säga, det behöver övervägas vilken definition som ska vara avgörande för att bedöma förekomst av diskriminering. Det matematiska faktum som beskrivs ovan har alltid funnits men visar sig inte förrän man börjar ersätta intuitiva bedömningar av icke-diskriminering med matematiska definitioner.

[44] Kleinberg/Mullainathan/Raghavan (2016) ”Inherent Trade-Offs in the Fair Determination of Risk Scores”. https://arxiv.org/abs/1609.05807
[45] Häggström, Olle (2021) Tänkande maskiner – Den artificiella intelligensens genombrott. Fri tanke.

Ett annat problem är att även om man tar bort vissa data t.ex. uppgift om ålder, hudfärg eller religion för att undvika diskriminering (jfr fairness throught unawareness), så är detta inte någon garanti för att diskriminering inte uppstår. Det kan snarare vara så att en modell som inte har tränats på sådana data riskerar att inte kunna hantera denna typ av data när den sedan ska användas i produktion, vilket i sin tur kan leda till diskriminerande utfall. Exempelvis kan en persons betyg från utbildningssystemet ha ett samband med personens socioekonomiska bakgrund, med följd att en statistisk analys som exkluderar sådana faktorer ger en felaktig eller ofullständig bild av vissa människors förmåga att uppnå höga betyg.[46]

9.3.3 Förtroendemodellen och andra initiativ

Det pågår ett flertal initiativ inom området förklarbarhet, tillförlitlighet och icke-diskriminering, t.ex. är förtroendemodellen ett sådant initiativ.

Förtroendemodellen är ett verktyg för självutvärdering av användningen av AI hos aktörer inom offentlig sektor. Det främsta syftet är att bevara, och till och med öka, den öppenhet och transparens som finns inom svensk offentlig förvaltning. En ifylld förtroendemodell ska som huvudregel kunna visas upp för tillsynsmyndigheter och intresserad allmänhet. Genom att använda förtroendemodellen får den som vill använda AI i sin verksamhet en ökad trygghet om de regler och förhållningssätt som behöver beaktas vid användningen av AI. Därmed kan förtroendemodellen även vara till nytta i den interna verksamhetsutvecklingen.[47]

Utöver förtroendemodellen finns också andra insatser både inom EU och på nationell nivå, exempelvis:

- Etiska riktlinjer för tillförlitlig AI[48]
- High level expert group on Artificial intelligence[49]
- Rapporter från Diskrimineringsombudsmannen, t.ex. Transparens, träning och data, Myndigheters användning av AI och automatiserat beslutsfattande samt kunskap om risker för diskriminering

[46] Deldjoo, Y., Jannach, D., Bellogin, A. et al. Fairness in recommender systems: research landscape and future directions. User Model User-Adap Inter (2023).
[47] https://beta.dataportal.se/offentligai/fortroendemodellen
[48] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[49] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

10. Träna, validera, testa, utvärdera och finjustera

10.1 Rättsliga, etiska och säkerhetsmässiga bedömningar

Liksom vid insamlings- och förberedelsestegen behöver myndigheten göra en rättslig bedömning av databehandlingen i detta steg (träning, validering, test). Det är en fördel, se avsnitt 7, om bedömningen för hela utvecklingsprocessen kan göras i samband med insamlingen. Är det inte möjligt eller om det uppkommit nya syften och ändamål under utvecklingens gång måste man göra nya bedömningar t.ex. avseende frågor om personuppgiftsbehandling, sekretess, arkiv och gallring, möjlighet till transparens och förklarbarhet, ansvarsförhållanden. Även informationsklassningen kan behöva uppdateras. Stöd för de rättsliga frågorna finns t.ex. i eSams vägledningar och checklistor, exempelvis checklistan *Juridik för användning av AI,*[50] checklistan *Rättsliga förutsättningar i utvecklingsinsatser m.m.*,[51] eSams promemoria *Stordataanalyser och datasjöar*[52] eller den guide som är publicerad på offentliga AI, *AI-guiden Juridik och etik*.[53]

10.2 Olika typer av maskininlärning

Som beskrivits i avsnitt 9.1 kan maskininlärning delas in i tre huvudtyper; övervakad inlärning, oövervakad inlärning och förstärkningsinlärning. Det finns också kombinationer av dessa t.ex. semiövervakad inlärning. Bilden nedan är en översikt av exempel på algoritmer inom respektive typ av maskininlärning, typ av data (märkta eller omärkta) och exempel på användningsområden. I efterföljande avsnitt (10.2.1–10.2.4) beskrivs innebörden av huvudtyperna lite närmare.

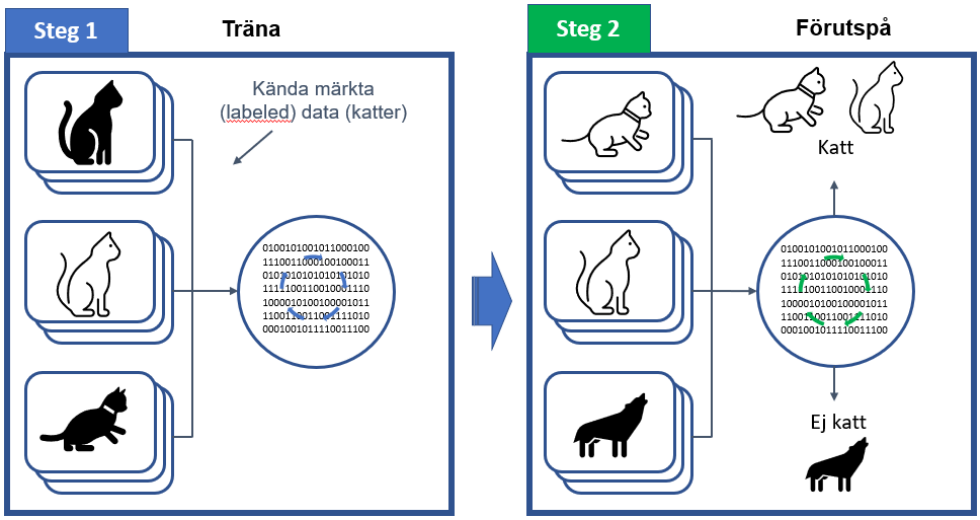
	Exempel på algoritmer[54]	Typ av data	Exempel på användningsområden
Övervakad inlärning	Linear Regression Logistic Regression Support Vector Machines (SVM) K-Nearest Neighbors (KNN) Naive Bayes Decision Trees Random Forests Neurala nätverk (t.ex. Convolutional Neural Networks, Recurrent Neural Networks)	Alla data är märkta (labeled)	Redan klassificerade data – text, bilder filmer m.m. Förutspå resultat Prognoser (väder) Diagnoser Tillväxt (population) Skräppostfiltrering Identifiera bedrägerier

[50] ES2022-08 Checklista Juridik vid användning av AI.
[51] Checklista för jurister Introduktion i rättsliga förutsättningar i utvecklingsinsatser, version 2.0, eSam juni 2019.
[52] ES2023-2 PM Stordataanalyser och datasjöar begrepp och rättsliga förutsättningar.
[53] https://beta.dataportal.se/offentligai/ai-i-praktiken-en-guide/juridik-och-etik
[54] https://learn.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet?view=azureml-api-2#download-machine-learning-algorithm-cheat-sheet

Semiövervakad inlärning	Label Spreading Label Propagation Self-Training Co-Training	Mindre märkt datamängd Större (oftast) omärkt (unlabeled) datamängd	När tiden för märkta data blir för kostsam Självträning
Oövervakad inlärning	K-means klustering Hierarkisk klustering DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Principal Component Analysis (PCA) T-distributed Stochastic Neighbor Embedding (t-SNE) Autoencoders Generative Adversarial Networks (GAN)	Alla data är omärkta	Klassificera ostrukturerade data Hitta gömda mönster Förminska (zip) data Hitta relationer i data Rekommendationer Riktad marknadsföring Kundsegmentering Hitta olika strukturer
Förstärknings-inlärning	Q-Learning Deep Q-Learning (DQN) State-Action-Reward-State-Action (SARSA) Proximal Policy Optimization (PPO) Asynchronous Advantage Actor-Critic (A3C) Monte Carlo Tree Search (MCTS)		Simultantolkning Summering av texter Fråga–Svar-system Spelutveckling Resurshantering Självkörande bilar Styrning av robotars rörelse Aktiehandel Identifiera datavirus Realtidsbeslut

10.2.1 Övervakad inlärning

Övervakad inlärning (supervised learning), även kallad lärarledd träning, innebär träning där modellen lär sig att göra förutsägelser baserat på en känd kategoriserad datamängd. Man kan också beskriva det som en förutsägelse av vilket utfall en viss indata ger utifrån exempel på indata och det förväntade utfallet. Ibland består ”facit” av historiska data, ibland tillförs facit genom en s.k. annoteringsprocess, se avsnitt 4.6. Träningen utgår således från märkta (labeled) data och man kan därmed förutspå resultatet. I exemplet nedan sker träning på bilder av katter där AI/ML-systemet lär sig identifiera katter, men inga andra djur.



Övervakad inlärning avser vanligen klassificering eller regression (t.ex. linjär eller logistisk) som lite förenklat kan beskrivas enligt följande:

- *Klassificering* innebär att indata används för att förutsäga vilken av en mängd klasser som aktuell indata tillhör, t.ex. om en bild ska klassificeras som en bild av en katt.[55]
- *Regression* innebär att indata används för att förutsäga ett värde från en kontinuerlig skala, t.ex. vad en fastighet borde värderas till.[56]

Skillnaden dem emellan är att klassificeringsmetoderna alltid väljer svar ur en begränsad uppsättning alternativ (till exempel katt/inte katt, eller 0, 1…9). Regressionsmetoderna producerar däremot ett svar i form av ett tal som inte behöver vara ett heltal.

10.2.2 Oövervakad inlärning

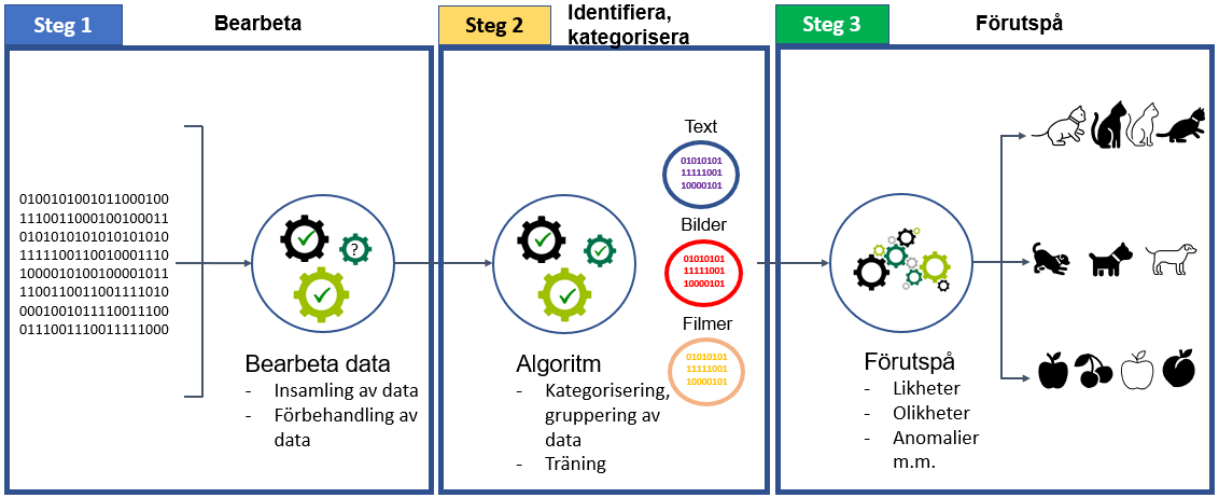
Oövervakad inlärning (unsupervised learning) innebär att en algoritm tränas utan att använda några fördefinierade märkta data eller kända svar på träningsdata, dvs. här finns inget tillgängligt facit till skillnad från vid övervakad inlärning. I stället för att försöka förutsäga ett specifikt svar, syftar oövervakad inlärning till att identifiera och synliggöra mönster, strukturer eller relationer inom data, som t.ex. kluster.

Klusteranalys innebär att algoritmer appliceras på träningsdata för att lära sig att hitta grupper av liknande data baserat på deras attribut. Algoritmen lär sig själv vilka attribut som är viktiga för att skilja mellan olika grupper och hur data är organiserade. Eftersom algoritmen inte får någon explicit återkoppling om huruvida dess grupperingar är korrekta, är träningsprocessen mindre styrd än i övervakad inlärning. Målet är i stället att

[55] Bishop. Christopher M. (2006) Pattern Recognition and Machine Learning (Springer).
[56] Bishop. Christopher M. (2006) Pattern Recognition and Machine Learning (Springer).

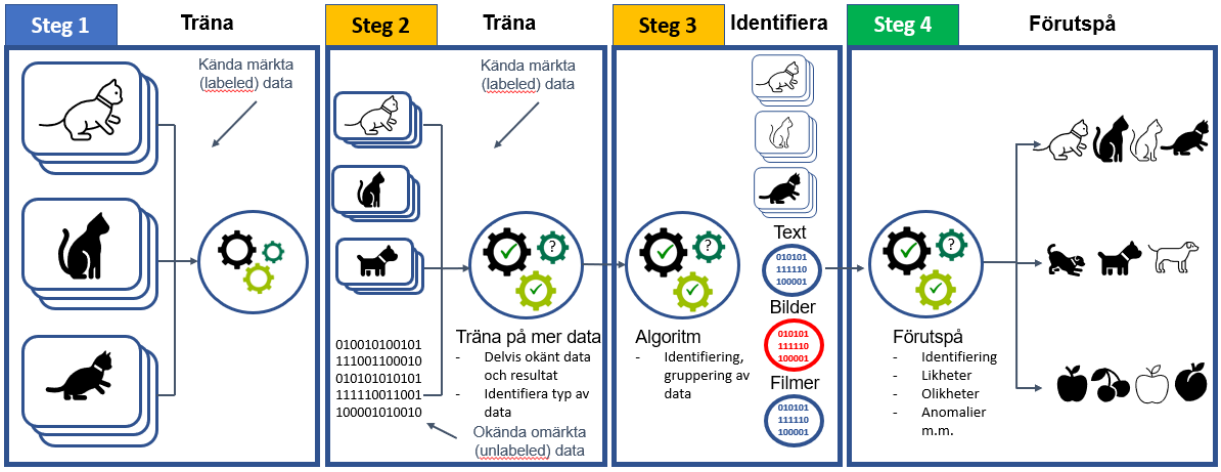
representera data på nya sätt, t.ex. för att utvinna ny kunskap från data, genom att låta algoritmen upptäcka och presentera mönster och strukturer som kan vara svåra att upptäcka manuellt.

Oövervakad inlärning används vanligen för att identifiera och kategorisera olika typer av data. Tidigare ej kända mönster och kopplingar i data kan identifieras och därigenom skapa ökad förståelse för dessa data. Vid träning kan oövervakad inlärning vara enklare, snabbare och mindre kostsam än övervakad inlärning. Oövervakad inlärning kan även användas vid mycket stora datamängder där det är svårt eller omöjligt att göra motsvarande mänsklig analys. Ett annat ganska vanligt sätt att använda oövervakad inlärning är att blanda olika typer av information och sedan låta algoritmen identifiera och klassificera data. Förväntningen är att dra slutsatser som inte har varit möjliga att göra tidigare på grund av mänskliga begränsningar eller att datamängden är för stor, varierad och komplex.



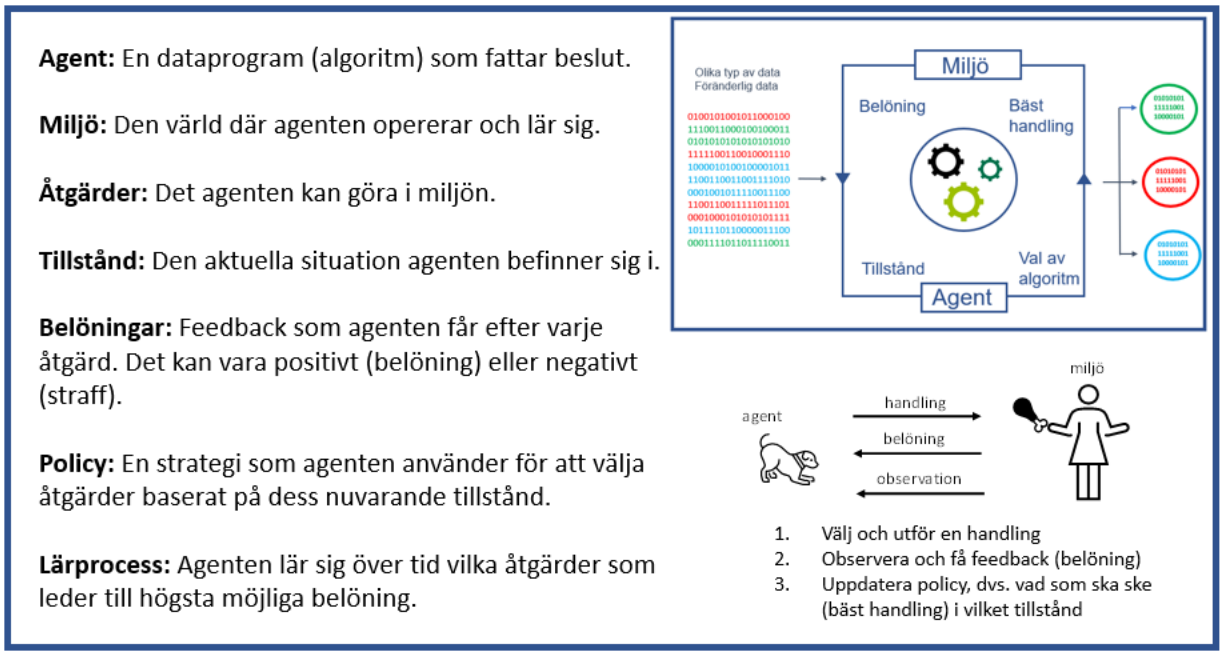
10.2.3 Semiövervakad inlärning

Semiövervakad inlärning är en kombination av övervakad och oövervakad inlärning och innebär att algoritmen använder en mindre mängd märkta (labeled) data tillsammans med en större mängd omärkta data (unlabeled) för att lära sig och kunna förutspå resultatet. Enkelt uttryckt används den märkta datamängden för att lära algoritmen vad den ska identifiera och förutspå i den större omärkta datamängden. Det innebär att man kan träna algoritmen på en mindre datamängd och sedan successivt använda allt större mängder nya omärkta data där algoritmen förutspår resultatet.



10.2.4 Förstärkningsinlärning

Förstärkningsinlärning (reinforcement learning) avser att träna och lära in det optimala beteendet i en miljö för att uppnå förväntade totala (inklusive framtida) belöningar.



Algoritmen/-erna skapar olika resultat och är tränade på att välja rätt handling beroende på specifika variabler eller situationer. Agenterna[57] lär sig utföra uppgifter genom upprepade försök och misstag i en föränderlig miljö eller på ett föränderligt scenario.

Ett exempel är en självkörande bil – en agent – som försöker lära sig att köra på ett säkert och effektivt sätt. Bilens miljö är vägen, med alla andra fordon, fotgängare och hinder som finns där. Bilens mål är att nå sin destination utan att orsaka olyckor eller

[57] Detta är en mycket förenklad beskrivning av agent, som även kan användas till andra saker t.ex. att upptäcka vissa beteenden och sedan presentera användaren för individanpassat innehåll eller anpassade tjänster.

förseningar. Bilens tillstånd är den aktuella situationen på vägen. Det inkluderar bilens position, hastighet, riktning, och avstånd till andra fordon och hinder. Bilens handlingar är det som bilen kan göra för att påverka sin miljö, t.ex. att starta och stanna, accelerera, bromsa, svänga och ändra fil m.m. Bilens belöning är en funktion av hur säkert och effektivt den kan köra.

10.3 Bias

Bias används som begrepp för systematisk snedvridning eller skevhet i bedömningsprocesser som missgynnar vissa grupper. Bias i AI/ML kan uppstå på många sätt och påverkar modellernas rättvisa, noggrannhet och tillförlitlighet. Bias är inte specifikt för AI/ML-system utan förekommer även i övriga it-system och i slutsatser eller beslut gjorda av människor. Exempel på några vanliga typer av bias är:

- **Bias i data:** När data som används för att träna eller testa modellen inte är representativt för det som AI/ML-systemet ska användas till, exempelvis utifrån mångfaldsperspektivet.
- **Algoritmisk bias:** Uppstår när själva algoritmen genererar fördomsfulla resultat, oavsett datakvaliteten. Det kan bero på fel i algoritmdesignen eller begränsningar i den matematiska modellen.
- **Sampling bias:** Uppstår när processen för att välja data för träning eller validering introducerar systematiska fel på grund av ett icke-slumpmässigt eller ett icke representativt urval. Det kan leda till felaktiga förutsägelser eller skeva resultat.
- **Kulturell bias:** Uppstår när AI/ML-modeller och deras träningsdata omedvetet reflekterar, förstärker eller favoriserar vissa kulturella normer, värderingar eller perspektiv.

Bias kan alltså uppträda både i tränings- och valideringsdata samt vid användning av ett tränat AI/ML-system på produktionsdata. Risk för bias behöver bedömas såväl vid val av data som vid tillsättning av olika roller i stegen i utvecklingsprocessen. Det är viktigt att identifiera potentiella källor till bias. Det är omöjligt att helt undvika bias. Ansatsen behöver dock vara att minimera sådan typ av bias som anses vara oacceptabel. Läs mer om bias och risker i eSams rapport *Samverkan kring tillämpad AI*.[58]

[58] ES2022-3 Rapport Samverkan kring tillämpad AI.

10.4 Överanpassning

Resultatet vid maskininlärning kan skilja sig avsevärt mellan det som en modell förutspår och det som tränings- eller testdata visar beroende på s.k. överanpassning.[59]

Maskininlärningsmetoder kan vara känsliga för överanpassning, vilket innebär att man har anpassat en modell för mycket till de mönster som finns i träningsdata att modellen fungerar sämre på verkliga data. Det är ett av skälen till varför valideringsdata och testdata behövs, dvs. för att säkerställa att man inte överanpassat modellen till träningsdata. Det gäller särskilt metoder som skapar flexibla modeller, som kan anpassas till nästan vilket fenomen i datamängden som helst. Dessa modeller orsakar lätt överanpassning om inte mängden träningsdata är tillräckligt stor.

10.5 Träning av modell

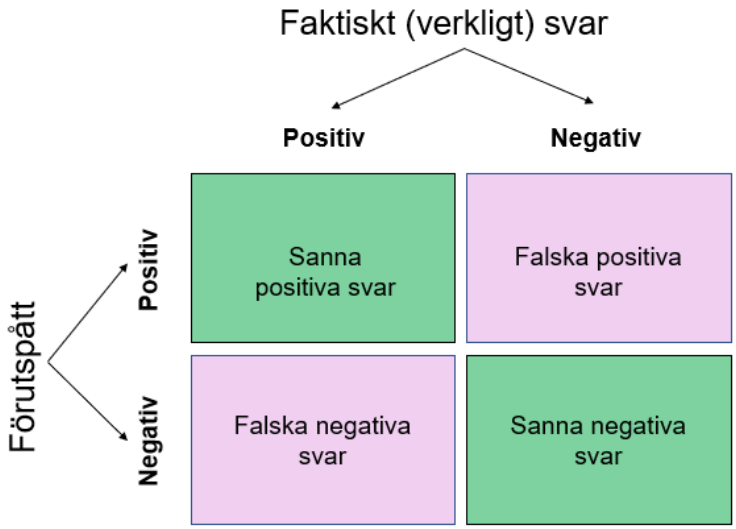
Träning av en modell är en iterativ process.

Vid träning av modellen är det viktigt att identifiera hyperparametrar (se avsnitt 4.12) som används för att styra inlärningsprocessen. Hyperparametrar kan vara ett antal dolda lager och antal parametrar i neurala nätverk eller att mäta antal fel, exempelvis vid översättning av äldre svenska från analoga akter (papper). Beroende på typ av och vald algoritm finns det olika sätt att mäta och därmed olika hyperparametrar som kan användas för att optimera träningen och därmed också resultatet. Det är vanligt att man mäter följande:

- Hur stor andel av modellens samtliga utfall både positiva och negativa som är korrekta (accuracy)
- Hur stor andel av modellens positiva utfall som är korrekta (precision)
- Hur stor andel av de faktiskt positiva fallen som modellen korrekt identifierar som positiva (recall)

Nedanstående bild visar vad som menas med positiva och negativa svar och hur de kan vara sanna eller falska (true or false).

[59] I sammanhanget kan nämnas att även begreppet underanpassning förekommer, vilket då avser att modellen inte lär sig tillräckligt från data.



När träningen är klar har modellen lärt sig att göra förutsägelser. Modellen anses vara bra om den har hög bedömd kvalitet, uppfyller fastställda mätvärden och inte medför oacceptabla bias eller andra utmaningar.

10.6 Validering av modell

Validering är när ett urval av data används för att ge en utvärdering av hur väl en modell presterar på aktuellt träningsdata. Datamängd för validering är skild från datamängd för träning, se avsnitt 4 om träningsdata och valideringsdata. Även hyperparametrarna kan justeras beroende på valideringen. Eftersom valideringsdata har ett redan känt facit kan de ställda kraven på modellen analyseras baserat på initialt fastställda mätvärden. Antal teckenfel vid transkribering av handskriven text eller antal korrekt identifierade cancerfläckar är exempel på det.

10.7 Test av modell

För att utvärdera den tränade och validerade modellen testas den på nya data som inte tidigare använts i träningen och som är representativa för de exempel som modellen kommer att ställas inför i verkliga tillämpningar. Ett exempel är analys av förändringar på byggnader i Sverige, där landskapsskillnader, höjdskillnader, vad som täcker markytan (såsom skog, bebyggelse, våtmark, hav, m.m.) kan påverka AI/ML-systemets förmåga. Den tränings- och valideringsdata som skapats motsvarar endast en liten del av Sveriges yta[60] och täcker inte alla varianter som kan förekomma. Virkesupplag, skugga, flygplatsrakor eller innergårdar kan vara särskilda utmaningar. För att öka korrektheten testas därför också AI-modellen på produktionsdata (se avsnitt 4.4) på utvalda områden.

[60] Idag har Lantmäteriet 2400 flygbilder i storlek 128x128 meter, som utgör tränings och valideringsdata. De täcker 8,5 promille av Sveriges yta (39,24 km² av 455 000 km²).

Ett annat exempel är när en organisation köper in eller på annat sätt får tillgång till en förtränad modell. Vidare träning av modellen kommer inte att göras men man vill fortfarande testa hur väl modellen presterar på egna data innan den sätts i produktion.

10.8 Utvärdering och finjustering

Utvärdering och finjustering[61] av hyperparametrar sker iterativt under tränings-, validerings- och testfasen, se även bild i avsnitt 6, till dess att modellen bedöms prestera tillräckligt bra.[62] Modellen är då färdig att användas i produktion.

[61] Med finjustering avses här justering av hyperparametrar, dvs. avser inte ”fine tuning” som ibland används för att beskriva vidareträning.
[62] T.ex. kan detta göras med en ”baseline model”.

11. Förbereda för produktion och förvaltning samt driftsätta

Steget från utveckling av AI/ML-system till löpande drift kan innebära påfrestningar för en organisation. Bästa sättet att motverka det är att påbörja implementeringen så tidigt som möjligt. Det bör klargöras tidigt i processen var produkten ska förvaltas. Ofta behövs en annan form av förvaltning än för traditionella förvaltningsmodeller. Eftersom den iterativa utvecklingen sker löpande med tillkommande nya data behöver hanteringen vara mer lik de agila modellerna. Sannolikt kommer fler delar av organisationen än it-funktionen att vara inblandade i användningen av AI/ML-systemet. Därför är det en fördel om de delarna av verksamheten kan vara delaktiga i förflyttningen från utveckling till någon form av löpande användning av produkten.

Den AI-drivna produkten eller produktserien som ska sättas i produktion måste ägas av någon del av organisationen. Det är fördelaktigt om det finns tydliga kopplingar mellan relevant verksamhet och utvecklingsverksamheten, eftersom en AI-modell satt i produktion utvecklas iterativt över tid. En annan anledning är att modellen ska underhållas och övervakas för att fungera korrekt i förhållande till det planerade målområdet.[63]

Det är en rekommendation att göra en risk- och konsekvensanalys av hur modellen kommer påverka verksamheten och identifiera presumtiva risker innan en AI-modell ska sättas i produktion. Analysen bör även ha ett säkerhetsperspektiv.

När ett AI-projekt sätts i produktion behöver det finnas en ansvarsstruktur för AI-modellens resultat. Vilka roller som tilldelas ansvaret är olika för varje organisation och för varje AI-projekt. Det är en stor fördel om resurserna är definierade före produktionssättningen. I förvaltningen ingår att både dokumentera och att hantera dokumentationen för produkten. I förvaltningen ingår även att samla in, organisera, förbereda och rensa data från verksamhetsområdet. Det behöver även finnas resurser som hanterar bevarandefrågorna. Det behöver vara tydligt vem som har ansvaret vid eventuella fel och vem som hanterar dem om de uppstår. Det är viktigt att tydliga rutiner, regler och policyer för det löpande arbetet är på plats före produktionssättningen.

En AI-modell i produktion behöver budgeteras. Kostnader för förvaltningen måste vara omhändertagna innan modellen går i produktion. Om myndigheten köper in en färdig

[63] I detta sammanhang kan nämnas Human-in-the-loop (HITL) som är en metod inom AI/ML där människor aktivt deltar i tränings-, övervaknings- eller beslutsprocesser.

modell med licenskostnader bör kostnaderna vara budgeterade. Resurserna i organisationen för AI-modellen ska vara klarlagda, oavsett om det är interna eller externa resurser. Det kan t.ex. finnas behov av att uppgradera befintlig hård- och/eller mjukvara, vilket bör vara uppmärksammat före produktionssättning, liksom estimerade framtida kostnader för uppdateringar av t.ex. serverhallar. Sammantaget måste det finnas utrymme i budgeten för eventuella merkostnader

I förvaltningsuppdraget av AI-modellen ingår även att ansvara för kommunikationen och utbildningsinsatserna, både internt och externt. Ansvaret för detta varierar i organisationerna, men oavsett vilka rutiner som finns är det viktigt att ha ett nära samarbete mellan organisationen för AI-modellen och kommunikationsfunktionen.

12. Reflektioner

I diskussioner och rapportering om AI är det vanligt att det lyfts fram olika slags risker och det görs en skillnad mellan traditionell utveckling jämfört med maskininlärning. Hur stor är egentligen skillnaden och vad består den av? Handlar det om data, om utvecklingsmetoder, om resultatet och/eller om riskerna? Det kan konstateras att det finns många likheter. Både traditionell utveckling och maskininlärning tar sin utgångspunkt i data för att formulera krav, att testa och utvärdera kod, att träna en modell etc. Mängden data och datakvalitet i relation till det som ska lösas är viktigt i båda metoderna. Vid traditionell utveckling tar man fram krav och utifrån dessa utvecklar en människa en kod för att lösa ett problem. Vid maskininlärning utvecklar maskinen logik och kod givet ställda krav för det lärande som ska ske, dvs. ändamålet och data för träning.

I det stora perspektivet kommer det alltid att finnas risker vid verksamhetsutveckling. De risker som ofta diskuteras i samband med AI, såsom bias, är varken nya eller unika för AI/ML-utveckling, de finns även vid mer traditionell utveckling. Vid maskininlärning är det fråga om förutsägelser. Det innebär en skillnad i hur resultaten uttrycks där resultaten från en modell utvecklad med stöd av maskininlärning ofta uttrycks i sannolikheter till skillnad från resultaten från en modell utvecklad med traditionell utveckling som oftare uttrycks som ett tvärsäkert ja eller nej. I det senare fallet blir eventuella underliggande osäkerheter inte lika tydliga. Maskininlärning bygger på att hitta samband, inte orsaker eller orsaksförhållanden.

Det kan utifrån detta finnas skäl att tala om risker i förhållande till konsekvens samt frekvens i förhållande till allvarlighet. Ett fel i ett system för e-postsortering kommer sällan ha samma konsekvens som ett fel som uppstår vid urval för utbetalning. Detta bör avspeglas i kravställningen för systemet. Här behöver myndigheten göra en samlad bedömning utifrån faktorerna: Vilken är risken? Vad blir konsekvensen om den uppstår? Hur ofta uppstår risken och hur allvarlig är den? Det behöver sedan vägas mot vinsterna med utvecklingen. Det finns också anledning att fundera på vilka krav som ställs på ett AI-system jämfört med de manuella processer som systemet ska ersätta. Kan en viss felprocent vara acceptabel så länge utfallet ändå är bättre än det manuella utfallet? Här kan man debattera att det är en skillnad då systemet kommer att ge systematiska felaktigheter jämfört med de manuella, men emot detta kan argumenteras att fördelen med systematiska felaktigheter är att de går att upptäcka och rätta.

Andra utmaningar är att få tillgång till tillräcklig mängd data för att genomföra en AI/ML-utveckling. En stor del av lagstiftningen är utformad för en analog värld vilket

gör det svårt att omsätta till variabler som AI-systemet ska förhålla sig till. Det kan också vara svårt att genomföra träning och validering. Många gånger kräver en AI/ML-utveckling stora manuella insatser i form av kunskaps- och insiktshöjande åtgärder, framtagning av tränings- och valideringsdata, träning, datahantering, tolkningar, juridiska bedömningar, integritets- och säkerhetsbedömningar m.m.

I ett större perspektiv behöver man ha i åtanke att ett AI/ML-system inte äger människans förmåga till intuition, sunt förnuft, omdöme, fantasi, moral, generalisering, innovation, medkännande m.m. Det kan i sig generera negativa känslor kring användningen. Vanliga invändningar är ”detta känns inte bra”, ”jobben kommer försvinna”, ”varför kan vi inte göra som vi alltid har gjort”. Ytterst kan det leda till ett minskat förtroende för offentlig förvaltning om AI-systemet genererar fel, innehåller bias, inte kan förklaras, inte kan rättfärdigas eller inte följer lagar eller praxis m.m. Därför är det viktigt att AI/ML-utveckling, på samma sätt som en traditionell utveckling, är väl förankrad och följer en viss process.

Som angetts i inledningen är syftet med denna rapport att skapa en gemensam bild och förståelse för hur en utvecklingsprocess för AI/ML-utveckling kan se ut, vilka överväganden som myndigheten behöver göra och vilka roller som behöver finnas och involveras inom organisationen. Förhoppningen är att denna rapport kan hjälpa till att konkretisera hur en verksamhet kan ta sig an en AI/ML-utveckling, dvs. att utgå från det specifika användningsfallet, identifiera vilka krav som måste ställas på systemet utifrån användning och se det som ett verktyg i verksamhetsutvecklingen som likt all annan verksamhetsutveckling bör följa en process. Genom samverkan och gemensam kunskapshöjning kan förhoppningsvis flera verksamheter komma igång med sin AI/ML-utveckling och använda AI/ML för att möta de grundläggande principerna i lagstiftningen om enkelhet, snabbhet, kostnadseffektivitet och rättssäkerhet.

eSam är ett medlemsdrivet program för samverkan mellan myndigheter för att underlätta och påskynda digitaliseringen inom det offentliga. eSam bildades 2015 som en frivillig fortsättning på E-delegationen. En viktig uppgift för eSam är att ta fram stöd och vägledningar som ger förutsättningar för att öka den digitala samverkan inom offentlig förvaltning.

Alla stöddokument finns på esamverka.se

I eSam ingår Arbetsförmedlingen, Arbetsmiljöverket, Bolagsverket, Boverket, Centrala Studiestödsnämnden, Domstolsverket, E-hälsomyndigheten, Ekonomistyrningsverket, Finansinspektionen, Folkhälsomyndigheten, Försäkringskassan, Havs- och vattenmyndigheten, Inspektionen för vård och omsorg, Jordbruksverket, Kemikalieinspektionen, Kriminalvården, Kronofogdemyndigheten, Kustbevakningen, Lantmäteriet, Länsstyrelserna, Migrationsverket, Naturvårdsverket, Patent- och Registreringsverket, Pensionsmyndigheten, Riksarkivet, Rättsmedicinalverket, Sida, Skatteverket, Skolverket, Statens institutionsstyrelse, Statens servicecenter, Statens tjänstepensionsverk, Statistiska centralbyrån, Tillväxtverket, Trafikverket, Transportstyrelsen, Tullverket och Universitets- och högskolerådet (okt 2023).