Syntetisering som anonymiseringsmetod för kommundata, Helsingborgs stad

April 2023

Förord

Under hösten 2022 påbörjades en förstudie om syntetiska data i Helsingborgs stad. Förstudien utfördes som ett samarbete mellan tre förvaltningar, Stadsledningsförvaltningen (digitaliseringsavdelningen), Vård- och omsorgsförvaltningen samt Arbetsmarknadsförvaltningen.

En del av förstudien var att ta fram material för personer med ingen eller mindre teknisk kunskap som förklarar vad anonymisering och syntetisering är för något. Under studiens gång blev det tydligt att sådant material är väldigt viktigt och att det saknas kunskap om framför allt syntetisering vilket inte är förvånande då det är en relativt ny teknik. Denna rapport är en del av arbetet att sprida kunskap och rapporten riktar sig huvudsakligen till personer med mindre teknisk kunskap rörande anonymisering och syntetisering som anonymiseringsmetod.

Under studiens gång blev det också tydligt att juridiska expert borde ha varit mer inblandade då det är oklart hur vi skall förhålla oss till syntetisering som anonymiseringsmetod rent juridiskt. Denna rapport kommer även ta upp de juridiska problem vi har stött på. Observera dock att denna rapport inte är någon juridisk utvärdering och är heller inte skriven av någon jurist.

Innehåll

1 Tekniska termer

- Artificiell intelligens - AI: Löst definierat är AI ett system som försöker efterlikna människors eller djurs intelligens och på ett eller annat sätt utföra uppgifter till någon grad av självständighet.
- AI-modell: en uppsättning kod och algoritmer som kommer eller har tränats för att utföra en eller flera specifika uppgifter. Modellens förmåga är baserat på modellens arkitektur och den data som den tränats på.
- Attribut: en datapunkt som beskriver data i ett dataset. Anges ofta på första raden i ett dataset för att ge en förklaring till all data i tillhörande kolumn.
- Data: siffror, bokstäver eller andra tecken som håller någon form av information.
- Dataset: en samling med data av samma slag och som på något sätt relaterar till varandra.
- Maskininlärning (Machine Learning) - ML: Är en delkategori inom AI där maskiner eller system använder sig av data och algoritmer för att lära sig. Idag den mest använda tekniken inom AI.
- Syntetiserade data: Data som på ett eller annat sätt har blivit skapat, ofta baserat på annan data, med hjälp av bland annat AI.

2 Inledning

Sveriges kommuner står idag inför flera olika utmaningar. En sådan utmaning är en minskad befolkningsmängd i arbetsför ålder och flera äldre personer som kommer behöva hjälp av kommunen. Detta är en trend som sker i flera delar av världen och kommer att ha en stor påverkan på många samhällen i framtiden. Bland annat kommer den påverka all offentlig verksamhet i Sverige som kommer behöva utföra mer arbete fast med mindre resurser.

Vidare så vill vi inom offentlig verksamhet också göra bättre ifrån oss. Vi vill att de tjänster vi erbjuder imorgon är bättre än de tjänster vi har idag. Det ska vara enkelt i vardagen för våra medborgare och det ska finnas tillit till kommunen. Det betyder både att vi finns här när medborgare söker sig till oss, oavsett tid, och att vi löser problem redan innan de dyker upp.

För att klara av de två ovannämnda punkterna och alla de andra utmaningar som kommunsverige står inför idag, samt alla de utmaningar som kommer att uppstå framöver, behövs nya arbetssätt och nya tekniker kommer att behöva nyttjas.

En teknologi som har varit på uppgång under många år är artificiell intelligens (AI). AI används flitigt runt om i världen inom en mängd olika områden för att utföra uppgifter snabbare, säkrare och bättre, när som helst på dygnet. AI kommer inte vara lösningen på alla de utmaningar som finns, i alla fall inte ännu, men det kan vara lösningen på några av utmaningarna eller åtminstone en dellösning på många problem.

Huruvida, för att kunna nyttja AI krävs oftast stora mängder med data och i flera fall mer data än vad som finns inom någon enskild kommun. AI-modeller fungerar i nästan alla fall bättre ju mer data de har tillgång till att träna sig på. Det betyder att om kommuner skall kunna utveckla användbara och bra AI-modeller så kommer det behövas att data delas mellan kommuner, och ju fler kommuner desto bättre. Men behovet av att dela data leder till ett stort problem, hur kan kommuner dela data?

Att dela data i sig är nödvändigtvis inget stort problem men att dela data som innehåller personuppgifter, vilket är en stor del av den data kommuner har, är betydligt mer problematiskt. I normala fall används olika typer av anonymiseringsmetoder. Tyvärr brister ofta dessa metoder där ett avvägande måste göras mellan säkerhet, det vill säga att det inte är möjligt att identifiera någon enskild person, och informationsmängden, det vill säga den information som data bär med sig. Ju mindre information och ju mer generell information desto säkrare blir datasetet men ju mindre användbar blir den.

Att dela data som innehåller personuppgifter inom det offentliga Sverige för bättre tjänster är något som just nu är väldigt aktuellt inom Sveriges regioner med fokus på sjukvårdsdata där flera regioner har undersökt, eller håller på att undersöka frågan om datadelning och där syntetiserade data har lyfts som en möjlig lösning[1] [2] [3]. Syntetiska data kommer att vara i fokus även här men en bredare diskussion om anonymisering förekommer också.

[1] Emmy Fokine, Mattis Åhman, Hugo Mellvé, Samuel Bengtsson, Lina Andersson. 2022. *Syntetiska data i hälso- och sjukvården.* Sverige: Göteborgs Univeristet & AI Sweden.

[2] AI Sweden & Region Västerbotten. 2022. *Syntetiska data inom intensivvården.* Sverige. `https://www.ai.se/sites/default/files/content/bilder/1\_syntetisk\_data\_inom\_iva\_rapport.pdf`

[3] Centrum för Hälsodata, Region Stockholm. *Syntetiserade hälsodata för förbättrad vård.* `https://ssci.se/sv/aktuellt/syntetiserade-h-lsodata-f-r-f-rb-ttrad-v-rd`

3 Anonymisering

Det finns flera lagar som är till för, att på olika sätt, skydda personer och deras privata information som innefattas i data. Dessa lagar är dataskyddsförordningen också känd som ”general data protection regulation” (GDPR), registerlagar och offentlighets- och sekretesslag (OSL)[4]. Lagarna kommer att till viss del diskuteras senare i detta dokument.

Data som är direkt kopplat till individer eller som kan på olika sätt tillsammans med annan information användas för att identifiera individer räknas som personuppgifter. Data som innehåller personuppgifter får inte användas hur som helst och kräver ett större ansvar av den som handskas med denna typ av data. En stor del av den data som förekommer inom en kommun innehåller personuppgifter och vissa av dessa personuppgifter klassas även som särskilt känsliga. För att kunna arbete med data inom bland annat utveckling där fler personer, ibland externa aktörer, behöver tillgång till denna data så anonymiseras den för att upprätthålla skyddet av individer som förekommer i dataseten.

Att anonymisera data är att ta bort eller ändra delar av ett dataset med avsikten att skydda personlig integritet. Om data har blivit anonymiserat till en sådan grad att det ej längre går att identifiera individer räknas det inte längre som personuppgifter. Att anonymisera data kan göras av flera skäl men vanligt förekommande är vid delning av data till en annan part som inte är behörig eller som inte behöver ta del av personuppgifter. Helt anonyma data räknas som enbart data och inte personuppgifter och berörs därmed ej av lagar som GDPR!

Nedan följer en beskrivning av olika typer av anonymiseringsmetoder och exempel.

Följande påhittade dataset, Tabell 1, kommer att användas som exempel på ”originaldata”.

Namn	Personnummer	Adress	Kontakt med förvaltning	Datum för kontakt
Anna Andersson	19810101-1645	Hemgatan 1, 111 11	Vård och omsorg	2022-06-29
Bengt Bertilsson	19920202-3471	Bortåtvägen 2, 222 22	Skola och fritid	2022-07-15
Cajsa Carlsson	20030303-2109	Mellangatan 3, 333 33	Social	2022-09-02

Tabell 1: Ett exempel på dataset som innehåller personuppgifter.

Pseudoanonymisering är ett sätt att anonymisera data men där det fortfarande ska gå att koppla data till en specifik individ. Detta görs exempelvis genom att byta ut uppgifter som är direkt kopplade till personen med en slumpmässigt vald unik id, och sedan spara personuppgifter-id kopplingen på en annan plats. Då kan enbart personen som sitter på både datasetet med användbara data och datasetet med personuppgifter-id kopplingen identifiera individer. Pseudoanonymiserade data räknas alltid som personuppgifter!

Namn	Personnummer	Adress	ID
Anna Andersson	19810101-1645	Hemgatan 1, 111 11	123123
Bengt Bertilsson	19920202-3471	Bortåtvägen 2, 222 22	456456
Cajsa Carlsson	20030303-2109	Mellangatan 3, 333 33	789789

Tabell 2: Dataset med Namn-id koppling.

ID	Kontakt med förvaltning	Datum för kontakt
123123	Vård och omsorg	2022-06-29
456456	Skola och fritid	2022-07-15
789789	Social	2022-09-02

Tabell 3: Dataset som delas för användning.

När en pseudoanonymisering utförs delas datasetet upp i två delar, ett dataset som innehåller data och en nyckel, se Tabell 3, och ett dataset som innehåller personuppgifter och nyckeln kopplad till resten av data, se Tabell 2. I detta fall blir inte det dataset som skall brukas direkt känsliga i sig och bara det ena datasetet innehåller direkta personuppgifter. Huruvida, då det finns en koppling mellan dataseten så kan en utomstående,

[4]För en kort summering över Sveriges dataskyddslagstiftning: `https://www.imy.se/verksamhet/dataskydd/sa-hanger-lagarna-ihop/`

om de kommer över båda dataseten, kunna koppla dem samman och därmed identifiera alla individer som finns i dataseten.

Attributborttagning är när ett attribut tas bort helt, som exempelvis individers namn. Attributborttagning kan tillämpas för att ta bort attribut som leder till direkt identifiering och som inte bär med sig någon nödvändig information samt för att ta bort attribut som hade varit bra för användningen men som med andra metoder inte har kunnat anonymiseras.

Kontakt med förvaltning	Datum för kontakt
Vård och omsorg	2022-06-29
Skola och fritid	2022-07-15
Social	2022-09-02

Tabell 4: Dataset som delas för användning.

Genom att ta bort alla attribut i Tabell 4 som kan kopplas till någon specifik individ, såsom namn, personnummer och adress är nu datasetet mer anonymt.

Maskning är en anonymiseringsmetod där vissa attribut helt eller delvis ”maskas”. Det kan exempelvis röra sig om en fritext där en individs namn förekommer och därmed måste maskas. Det kan också röra sig om exempelvis personnummer där de sista fyra siffrorna maskas för att ha kvar födelsedatum men ta bort en unik identifierare.

Namn	Personnummer	Adress	Kontakt med förvaltning	Datum för kontakt
Namn Efternamn	19810101-XXXX	Gata X, 111 11	Vård och omsorg	2022-06-29
Namn Efternamn	19920202-XXXX	Gata X, 222 22	Skola och fritid	2022-07-15
Namn Efternamn	20030303-XXXX	Gata X, 333 33	Social	2022-09-02

Tabell 5: Ett dataset där attribut som kan användas för att identifiera en individ har blivit maskerade.

Genom att maska namn, gatuadress och de fyra sista siffrorna i personnumret i Tabell 5, har det blivit betydligt svårare att identifiera någon enskild person. Namnet hade dock lika gärna kunnat tas bort helt men genom att maska delar av personnumret och adressen finns nu födelsedatum och postnummer kvar tillskillnad från attributborttagning där denna information försvann.

Generalisering är en metod där data generaliseras vilket leder till en lägre granularitet. Exempelvis så används område i stället för hemadress, i stället för exakt ålder så används ett åldersspann, i stället för exakta värden avrundas värdet till närmsta tiotal och så vidare.

Namn	Födelseår (årtionde)	Postnummer	Kontakt med förvaltning	Datum för kontakt
Namn Efternamn	1980	111 11	Vård och omsorg	2022-06
Namn Efternamn	1990	222 22	Skola och fritid	2022-07
Namn Efternamn	2000	333 33	Social	2022-09

Tabell 6: Ett dataset där attribut har blivit generaliserade vilket gör det svårare att identifiera enskilda individer.

Återigen bör namn tas bort då det både för med sig minimal information och kan användas för att direkt identifiera individen. Genom att använda exakt födelsedatum och postnummer är det relativt lätt att identifiera enskilda individer. Att generalisera födelsedatum i Tabell 6 till årtionde gör det svårare att identifiera personerna. Även den exakta dagen för kontakt generaliserar till månad och adressen till postnummer. Observera att granulariteten avgörs baserat på den data som finns. Exempelvis hjälper det inte att dela in födelseår i årtionden om det bara finns en personen född inom ett visst årtionde.

Andra metoder. Det finns även en rad olika metoder som ändrar, byter eller lägger till attribut eller datarader. Exempelvis för att göra det svårare att känna igen en specifik individ kan flera liknade individer läggas till i datasetet, det vill säga ett brus adderas. Andra möjliga metoder är att skyffla runt attribut bland olika individer eller helt enkelt ändra attribut som är för unika. Ju mer ett dataset förändras ju sämre representera det verkligheten och desto sämre blir användbarheten.

Genom att lägga till ett brus och skyffla om data i Tabell 7 blir det väldigt svårt att säga något om någon enskild individ. Det finns dock påhittade data i datasetet och flera samband har gått förlorade så denna data har

Namn	Personnummer	Adress	Kontakt med förvaltning	Datum för kontakt
Anna Andersson	20030303-2109	Nedåtvägen 1, 444 44	Stadsbyggnad	2022-08-21
Bengt Bertilsson	20140404-4534	Hemgatan 2, 111 11	Social	2022-07-15
Cajsa Carlsson	19810101-1645	Bortåtvägen 3, 222 22	Skola och fritid	2022-06-29
David Danielsson	19920202-3471	Mellangatan 4, 333 33	Vård och omsorg	2022-09-02

Tabell 7: Ett dataset som har skyfflats om och med ett adderat brus.

nu ett lägre användningsvärde. Observera att namn, fyra sista siffrorna i personnumret och gatuadress också borde ha plockats bort.

Statistik räknas också som en form av anonymisering men är även en bearbetning för nyttjande av data. Statistik säger något om datasetet men själva datasetet är fortfarande dolt. Därmed blir det svårare att utvinna annan information eller statistik ur datasetet än det som delas. Det är dock möjligt att ur statistik, om en för låg aggregering av värden har skett, att utvinna information om specifika individer.

Förvaltning	Antal kontakter juli	Antal kontakter augusti	Medelålder på person som sökt kontakt
Vård och omsorg	30	34	72
Skola och fritid	11	61	29
Social	14	13	41

Tabell 8: Påhittade statistik över antal personer som har varit i kontakt med olika förvaltningar på en kommun.

Under förutsättningen att originaldatasetet, Tabell 1, är mycket större skulle det gå att få ut statistik som går att finna i Tabell 8, vilket gör det svårt att identifiera någon enskild individ ur originaldatasetet. Men det finns alltid en risk. Om det finns kunskap om att en 92 åring har haft kontakt med kommunen så går det att använda ovanstående statistik för att härleda att personen troligtvis kontaktade vård- och omsorgsförvaltningen.

För en mer detaljerad beskrivning av anonymiseringsmetoder och hur de utförs se bland annat Singapores anonymiserinsguide[5].

Vanligtvis används inte en enskild anonymiseringsmetod utan ett flertal metoder används tillsammans beroende på vilken data som finns. Ett stort problem med nästan alla anonymiseringsmetoder är att data och relationer mellan data går förlorade när metoderna används. Därmed uppstår det en konflikt mellan användbarhet och säkerhet. Skall viktiga data och dess relationer sparas men med större risk för identifiering eller skall den data och dess relationer tas bort vilket leder till sämre resultat när den anonyma data används.

När det kommer till båda analyser och AI-modeller är det viktigt att ha med så mycket data som möjligt och att kvalitén på data är så hög som möjligt. Det kan därmed vara ett stort problem om det finns lite data eller om alla utstickande(extrema) fall har tagits bort då dessa kan vara av extra intresse. En AI-modell som tränas upp utan dessa unika fall kommer ha svårt att handskas med just sådana fall.

Det är här syntetiska data kommer in. Syntetisering av data kan också ses som en anonymiseringsmetod men om den utförs rätt kan den producera data som är både säkrare och av bättre kvalitet än data som utsätts för traditionella anonymiseringsmetoder.

[5]Personal Data Protection Commission. 2018. *Guide to basic data anonymization techniques*. Singapore.

4 Syntetisering

Att syntetisera data är att framställa data, oftast med hjälp av en AI-modell som tränats på riktiga data. Till skillnad från data som har anonymiserats med vanliga metoder ska den nya syntetiserade data ej innehålla någon originaldata. Syntetiserade data ses därför i teorin som anonym data men i praktiken finns det risker att den syntetiska data som tagits fram speglar originaldata så pass väl att det går att utvinna information om originaldata.

Syntetisering av data är inte bara en anonymiseringsmetod utan det finns flera andra anledningar till att syntetisera data förutom att skapa anonyma data. Syntetisering kan även utföras på flera olika typer av data. Exempelvis kan både nya röster och ansikten skapas utan koppling till någon specifik människa. Vidare kan syntetisering användas för att skapa mer data för bättre träning av AI-modeller och för att minska bias i den originaldata som finns.

Faktum är att syntetisering redan används mycket idag. Några exempel är applikationer som använder tal vilka ofta är syntetiskt framställda röster och företag som använder syntetiska bilder för att testa och träna bildigenkänning som finns i bland annat självkörande bildar. Nyligen har även flera modeller för att generera bilder och konst tagits fram. Ett annat bra typexempel är hemsidan `https://thispersondoesnotexist.com` som generar mänskliga ansiktet som inte finns i verkligheten.

Användningen av syntetiska data spås växa och kommer till om med gå om originaldata enligt konsultbyrån Gartner. Analytiker på Gartner förutspår att användningen av syntetiska data för att träna AI kommer öka ordentligt och vid 2024 kommer syntetiska data användas mer än riktigt data[6].

Även om de finns många möjligheter med syntetiska data så är denna studies fokus på att anonymisera tabulär data med anledning av att Helsingborg har mycket data av denna typ som innehåller personuppgifter. Vidare finns det en stor potentiell nytta med att kunna använda denna data för att skapa bättre tjänster inom kommunen, vilket idag är svårt att göra på grund utav att den data innehåller just personuppgifter.

Att ha bra och snabba metoder för att anonymisera data kommer göra det lättare att arbete med data både internt och extern. Internt inom förvaltningen och inom kommunen som idag kan vara svårt. Externt med privata aktörer men framför allt andra kommuner där det blir möjligt att dela data.

Några möjliga användningar av syntetiska data:

- Skapa öppna data för näringsliv och akademin så att de kan utforska och skapa nya lösningar.
- Dela data snabbt med analytiker som vanligtvis inte jobbar med den specifika data annars, och därmed inte har behörighet.
- Slå ihop data med andra aktörer för att skapa bättre AI-modeller.
- Spara data säkert över en längre tid.
- Skapa ett större skydd för individers integritet.
- Öka tillgången till data inom kommunen.
- Använda data för gemensam utveckling av AI-modeller eller tjänster.

En sista viktig punkt som till viss grad redan nämnts är att den höga kvalitét och säkerhet som syntetiska data ska ha enbart är i teorin. I praktiken däremot, finns det inga garantier för att syntetiska data har samma kvalitét som originaldata eller att den saknar personuppgifter utan det beror på vad för originaldata som finns, vad för data som skapas och hur den skapas. Syntetisering i sig är därmed inte den bästa anonymiseringsmetoden men den kan vara det om den utförs korrekt. Också nämnvärt är att syntetisering som anonymiseringsmetod är något relativt nytt och många framsteg sker just nu vilket betyder att situationen kan och kommer troligtvis att förändras framöver.

Nedan i Tabell 9 följer en jämförelse mellan originaldata, anonymiserade data och syntetiserade data.

Den anonymiserade data i Tabell 9 gör det svårare att direkt få reda på vilken individ det handlar om. Det skulle dock vara möjligt att härleda vilken individ det handlar om genom att använda födelsedatum och postnummer eller datum för kontakt tillsammans med annan information. Med andra ord är den anonymiserade data nödvändigtvis inte anonym. Den syntetiserade data däremot har andra värden då den inte längre refererar till en specifik individ utan det har i stället framställts en ny individ. Det är dock svårt att i en bild som denna visa på skillnader då syntetisering av data bara fungerar för en större mängd data, det finns med andra ord ingen direkt koppling mellan raden originaldata och raden syntetiserad, till skillnad från originaldata och anonymiserade data. Men det är just det som är hela poängen med syntetiska data!

[6] Andrew White, Gartner. *By 2024, 60% of data used for the development of AI and analytics projects will be synthetically generated.* 2021. `https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/`

Data	Namn	Personnummer	Adress	Kontaktat	Datum
Original	Anna Andersson	19810101-1645	Hemgatan 1, 111 11	Vård och omsorg	2022-06-29
Anonymiserad	Person 1	19810101-xxxx	111 11	Vård och omsorg	2022-06-29
Syntetiserad	Frida Davidsson	19820104	112 42	Vård och omsorg	2022-06-13

Tabell 9: En jämförelse mellan data som har anonymiserats och syntetiserats.

4.1 Teknisk beskrivning

En förenklad förklaring av syntetisering är följande: från originaldata tas olika statiska värden fram som sedan används för att generera data. Exempelvis följande tabell med åldrar:

Data:	Statistik:	Syntetiska data:
25	Minimum = 25	26
27	Maximum = 51	29
33	Antal värden = 7	30
42	Medelvärde = 38	42
44	Median = 42	43
45	-	46
51	-	50

Tabell 10: Ett simpelt exempel där första kolumnen innehåller data, andra kolumnen statistik över den data och den tredje kolumnen innehåller genererade data baserat på statistiken.

I Tabell 10 användes enbart enkel statistik. Den nya data håller sig inom minimum- och maximumvärden samt har samma medelvärde och median. I praktiken finns det betydligt fler statistiska värden som är av intresse och det kommer vara flera attribut vilket ökar komplexiteten samt kräver att även relationer mellan attribut finns kvar.

Följande exempel visar på hur vissa regler och samband måste sparas.

Kön:	Ålder:	Antal barn:
M	28	1
M	31	0
F	7	0
M	6	0
F	18	0
M	44	3
F	21	1

Tabell 11: Ett påhittat exempel över ett dataset som innehåller personers kön, ålder och antal barn.

Från de första två attributen, kön och ålder i Tabell 11 finns ett samband att kvinnor i datasetet har lägre ålder än män, något som kan vara viktigt att få med i det syntetiska datasetet. Det tredje attributet, antal barn, behöver ännu fler regler och har fler samband och relationer. En person kan inte få barn innan puberteten så en åldersgräns måste införas. Vidare kommer det troligtvis finnas samband både mellan kön och ålder gällande hur många barn personen har. Bara från ett enkelt dataset med tre attribut börjar det dyka upp flera regler och samband som blir viktiga att få med. Regler är något som sätts under skapandet av modellen och definierar grundstrukturen samt hur kolumner förhåller sig till varandra. Samband och relationer mellan attribut är något som måste följa med från originaldata under uppträningsfasen av modellen.

4.2 Användning av syntetiska data

I Tabell 9, ovan, gjordes ett försök till att visa på skillnader mellan originaldata och syntetiska data. Det nämndes även kort att det inte blir en rättvis jämförelse för att det inte finns någon direkt koppling mellan specifika data i originaldatasetet och det nya datasetet. Denna avsaknad av koppling medför viss begränsning av användandet av syntetiska data. Det går exempelvis inte att titta på ett specifikt fall hos det nya syntetiska

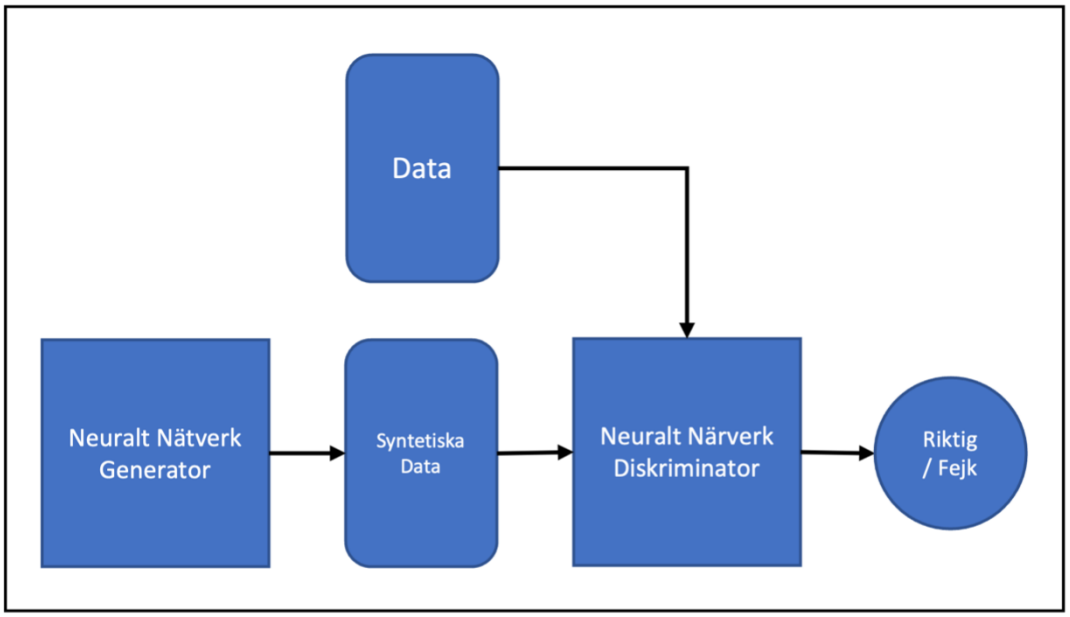
datasetet och säga något om verkligheten då dessa inte korrelerar. Exempelvis kommer denna rapport gå in på matsvinnsdata. Den syntetiska data går därmed inte att använda för att säga hur mycket matsvinn en specifik skola hade på en specifik dag. Men ovanstående är heller inte så konstigt, det är ju nämligen det som är poängen med syntetiska data, att det inte ska gå att härleda originaldata från den syntetiska data. Värdet ligger i stället hos mängden data och de samband som går att återfinna.

4.3 Kort teknisk förklaring om framställandet och kontroller av syntetiska data

Men hur skapas syntetiska data? Som nämnts tidigare används olika typer av AI-modeller som tränas på originaldata för att generera ny syntetiska data. Vilken modell som passar bäst beror på vilken data som finns och vilken data som skall skapas. Det är heller inte alltid klart vilken modell som skulle fungera bäst och i vissa fall bör ett flertal modeller testas och jämföras.

Två vanliga arkitekturer för AI-modeller är Genarative Adversarial Network (GAN) och Variational Auto-encoders (VAEs). Dessa arkitekturer kan användas för att skapa AI-modeller för att lösa flera olika uppgifter, bland annat syntetisering.

GAN är en arkitektur som innehåller två artificiella neurala nätverk som är sammankopplade med varandra och tävlar mot varandra. Det ena nätverket genererar en representation av data medan det andra nätverket diskriminerar, det vill säga försöker urskilja verkliga data mot generade data. Genom att kontinuerligt utföra denna process blir båda nätverken bättre på sin respektive uppgift och i slutändan finns en AI modell som är mycket bra på att generera nya data som liknar den data som användes vid träning.



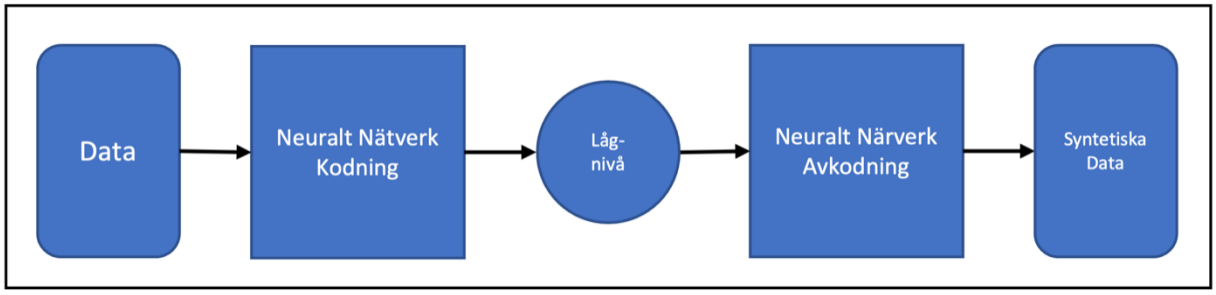
Figur 1: En simpel överblick över GAN-arkitekturen.

VAEs är också en arkitektur som består av två artificiella neurala nätverk som är sammankopplade med varandra. Men i stället för att tävla mot varandra så försöker det ena nätverket att koda ner så mycket information som möjligt till en lägre dimensionell representation och det andra nätverket försöker återskapa så mycket information som möjligt från en lägre dimensionell representation. På så sätt samarbetar nätverken genom att bli bättre på att konstruera och rekonstruera data från denna representation.

Det finns ännu fler arkitekturer än dessa för syntetisering och en specifik arkitektur kan anpassas på flera olika sätt. Vad som fungerar bäst beror oftast på den data som finns och hur den data är strukturerad.

Utöver att skapa data med hjälp av en AI-modell behöver också kontroller utföras så den nya syntetiska data håller tillräckligt hög kvalitet och att den är anonym. Det finns en del metoder för att göra dessa tester men ingen metod kan idag garantera att den nya syntetiska data bibehåller den generella kvalitén av originaldata och är helt fri från personuppgifter. Det finns heller ingen metod som kan garantera att data är anonym genom andra anonymiseringsmetoder heller.

Det bästa sättet att testa kvalitén av den syntetiska data är att använda båda originaldata och den syntetiska data i ett riktigt ändamål och sedan jämföra resultatet. Exempelvis om en AI-modell har tagits fram som skall tränas kan både originaldata och den syntetiska data användas för att testa modellen. Om modellen fungerar



Figur 2: En simpel överblick över VAE-arkitekturen.

lika bra tränad på den syntetiska data som på originaldata är den syntetiska data likvärda originaldata. Det är nu möjligt att dela den syntetiska data och tillsamman med syntetiska data från flera källor träna en gemensam, mycket större och bättre AI-modell.

Det finns naturligtvis flera andra tester som kan utföras för att ge en någorlunda förståelse över kvalitén på den syntetiska data som genererats, bland annat statistiska tester. Ett problem som finns idag är att det inte går att säga något om den ”allmän kvalitén” på den syntetiserade data. Med andra ord, om ett tjugotal statistiska tester utförs med bra resultat så har det syntetiska datasetet bra kvalitet med avseende på det som har testats men det finns inga garantier att ett samband som inte har testats för finns representerar i det syntetiska datasetet.

När det kommer till att testa om datasetet är anonymt finns det också en del metoder som kan användas men det tillkommer att vissa bedömningar behöver göras efteråt. Bland annat går det att automatiskt se om någon syntetisk individ korrelerar exakt med originaldata för att sedan ändra eller ta bort denna individ. Men om data måste ändras eller raderas så finns det risk att kvalitén minskar.

En risk som kan finnas med syntetisering av data är när det tränas på extremfall. Då kan liknande extremfall återskapas och en koppling kan uppstå även om den inte är ett-till-ett. Det kan därför, beroende på data, vara nödvändigt att kontrollera hur väl extremfallen som skapats korresponderar med extremfall i originaldata. Observera att dessa extremfall kan vara väldigt viktiga i möjliga analyser eller för att träna AI-modeller och därmed är det viktigt att ha kvar dessa fall. Metoder för att kontrollera huruvida ett dataset har data som sticker utfinns det gott om och används ofta för att mäta anonymitet.

Syntetisering är en väldigt bred kategori där det finns många metoder för framställning och för att kontrollera den nyproducerade data. Utvecklingen av nya metoder är även i full fart just nu. Det är därför svårt att ge en fördjupning förståelse av syntetiseringsmetoder utan ovanstående exempel ämnar skapa en generell förståelse och nämner enbart ett fåtal metoder.

4.4 Kort marknadsanalys

Men hur sätter jag i gång med att syntetisera data? Det finns just nu flera bolag som erbjuder olika typer av tjänster för syntetisering av data. Dessa tjänster är mjukvara i form av en plattform eller ramverk för syntetisering men även expertis hjälp. De innehåller också oftast lösningar för hela processen av syntetisering, från datastädning till framtagning av modell och till slutkontroller för att säkerställa kvalitén samt kontroller så att inga uppgifter har läckt från originaldatasetet till det syntetiska datasetet.

Vilken tjänst som är bäst beror på behov. Finns kunskap inom organisation behövs kanske bara någon enklare integrationsplattform, saknas kunskap behövs troligtvis ett ramverk där den externa parten sköter hela processen.

Ett annat alternativ är mjukvara med öppen källkod vilka är gratis men kräver utvecklar med kunskaper inom syntetisering och tid till att bygga egna lösningar. En del öppen källkod är dock relativt lätt att arbete med och det går snabbt att komma i gång. Men beroende på komplexiteten hos den data som skall syntetiseras kan det gå från snabbt och enkelt till långsamt och krångligt.

Vilket redskap eller tjänst som passar bäst beror på situationen men något som är positivt med öppen källkod är att det går snabbt och sätta i gång och testa, speciellt för kommuner som ofta kan ha långa upphandlingstider. Men blir det aktuellt att i framtiden syntetisera större mängder data mer kontinuerligt kan det vara lättare att köpa in färdiga professionella lösningar så att varje enskild kommun inte behöver anställa teams av data scientists för att syntetisera all sin data.

Ett exempel på öppen källkod som även användes under denna förstudie är Synthetic Data Vault (SDV). SDV beskriver sig som ett ekosystem för generering av syntetiska data. SDV har flera färdiga bibliotek som lätt

kan användas med bland annat programmeringsspråket Python. Dessa bibliotek innehåller en mängd olika AI modeller för flera olika typer av data samt färdiga tester för att mäta kvalitet på den nyproducerade data.

När det kommer till privata bolag finns det en stor mängd företag, varav några går att finna i denna lista `https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784`. Två bolag som vi har varit i kontakt med är Sogeti och Syndata. Syntetisering av data är en av många tjänster som Sogeti erbjuder medan Syndata fokuserar just på syntetisering av data. Att många privata aktörer väljer att arbeta med just syntetisering, tyder på att behovet finns och att tekniken börjar mogna.

4.5 Exempel syntetiska data (matsvinn)

För att lära oss samt kunna visa upp och dela ett exempel gjordes ett val att syntetisera ett dataset som inte är känsligt. Det dataset som valdes är matsvinnsdata hos förskolor, skolor och gymnasieskolor i Helsingborgs stad över tidsperioden 2017–2022. Observera att detta exempel huvudsakligen är riktat mot personer utan någon större kunskap rörande syntetisering. Exemplet är därför relativt enkelt och bör ses som en simpel syntetisering.

Datasetet innehåller följande attribut:

Namn	Beskrivning	Värde
Index	Index	Unikt positivt heltal
Datum	Datum som rapporteras	ÅÅÅÅ-MM-DD
EnhetNamn	Namn på enheten	Fritext
Nodtyp	Om enhet är en förskola, skola eller gymnasieskola	Kategorisk(förskola, skola, gymnasieskola)
EnhetId	En intern id för varje enhet	Positivt heltal, unikt för varje enhet
Maltid	Vilken måltid som rapporteras	Kategorisk(enbart lunch används just nu)
AntalbestalldaPortioner	Antalet portioner som enheten har beställt	Positivt heltal, kan vara 0
AntalKoptaPortiner	Antalet portioner som köps in från en annan enhet	Positivt heltal, kan vara 0
AntalAtande	Antalet personer som äter vid den dagen	Positivt heltal, alltid mer än 0
AntalSaldaPortioner	Antalet portioner som köket säljer till andra enheter	Positivt heltal, kan vara 0
KoksSvinnGram	Mängden svinn som uppstår i köket	Positivt heltal, kan vara 0, gram
ServeringsSvinnGram	Mängden svinn som uppstår vid servering, exempelvis överbliven mat	Positivt heltal, kan vara 0, gram
TallriksSvinnGram	Mängden svinn som kastas från tallrikar	Positivt heltal, kan vara 0, gram
DWSource	Var data inrapporteras ifrån	Kategorisk(Historisk eller tjänst)

Tabell 12: Beskrivning av matsvinnsdatasetet som användes under denna förstudie.

Datumet delades upp i år, månad, vecka och veckodag för att lättare kunna syntetisera den data som fanns och gör att göra det lättare att utvärdera resultaten.

Steg 1. Datarensning

Här ”städas” data genom att först analysera den data som finns och sedan göra förändringar. De förändringar som görs beror på vad problemet är och hur problemet har uppstått. Den senare användningen har också en påverkan på hur denna städning går till.

- Tar bort rader som saknar viktiga värden eller lägger till ett värde, till exempel medelvärdet för hela kolumnen.
- Tar bort eller justerar felaktiga värden.
- Tar bort irrelevant data som inte kommer ha någon påverkan, exempelvis namn.
- Tar bort dubbletter.
- Åtgärdar fel, exempelvis felstavningar.
- Ändrar struktur så all data följer samma format.

- Och så vidare.

För matsvinnsdata utfördes följande:

- Det kontrollerades att inga värden är negativa.
- Rader där allt svinn är NaN togs bort.
- NaN-värden för svinn ändrades till 0 (så länge det finns annat svinn rapporterat).
- Uppenbara fel togs bort (som uppstått pga. felrapportering. En typ av extremvärden som tittades på utöver varje enskilt attribut var totalt matsvinn per ätande).
- Kolumnen EnhetNamn togs bort (Id finns redan så namn för inte med något värde).
- Veckodagar ändrades till nummer (måndag = 0 ... fredag = 4).
- Nodtyp ändrades till nummer (förskola = 0, skola = 1, gymnasium = 2).
- Source ändrades till nummer (historik = 0, web = 1, api = 2, external = 3).

Kvar fanns 14 attribut som alla är numeriska eller kategoriska. För att det skulle gå snabbare och bli lättare att jämföra gjordes valet att enbart använda data från 2022 vilket var 8502 rader data (matsvinnsrapporter).

Steg 2. Statistisk analys

När datasetet har städats kan en statistisk analys utföras. Det finns flera anledningar till att genomföra en analys. Först och främst är det viktigt att förstå den data som arbetas med. För det andra så kan vissa av dessa värden användas som villkor för modellen när den skapas. För det tredje så kan statistska värden användas för att kontrollera så den syntetiska data som skapats är lik originaldata.

För numeriska värden och enskilda kolumner kan minimum, maximum, medel, median, kvartil och standardavvikelse beräknas. Ändra eventuell kategoriska data till numeriska värden så kan relationer mellan kolumner och datadistributioner undersökas för alla numeriska och kategoriska värden. Även simpla värden för kategorier kan vara intressant så som antal.

Korrelationsmatriser är ett bra verktyg för att visualisera den data som finns och de korrelationer som finns mellan de olika attributen.

För matsvinnsdata gjordes en analys varav en del av detta kommer visas i steg 8 då originaldata och syntetiska data jämförs.

Steg 3. Bestäm regler

Den data som skall syntetiseras innehåller troligtvis en del underförstådda regler och relationer som inte får brytas. En underförstådd regel skulle kunna vara att en person inte kan ha en negativ ålder. Därmed sätts strikta minimum- och maximumvärden. Andra underförstådda regler kan finnas mellan kolumner, exempelvis kan en person med manligt kön inte bli gravid. Det finns även relationer mellan kolumner som inte får brytas som exempelvis land och stad, eller att ett startdatum inte kan vara efter ett slutdatum.

Alla dessa regler måste specificeras och sättas upp innan träningen påbörjas.

För matsvinnsdata användes fördefinierade regler i SDV för alla numeriska värden som sattes att hamna inom min = 0 och max = maxvärde för respektive attribut. Vidare sattes en regel för Nodtyp och EnhetId då det var viktigt att bevara vilken nodtyp varje enhet tillhörde. Denna regel blev dock överflödig då i ett senare skede så delades datasetet upp ytterligare baserat på just Nodtyp för att uppnå ett bättre resultat. En begränsning definierades som följande i Python med paketet SDV: $koksvinn = ScalarRange(column\_name = 'KoksSvinnGram', low\_value = 0, high\_value = 33000, strict\_boundaries = False)$.

Steg 4. Ta fram och träna olika modeller

När alla regler är satta kan en modell tränas. Då det innan kan vara svårt att avgöra exakt vilken modell som passar bäst för den data som finns är det rekommenderat att testa olika modeller och olika inställningar. Det finns en stor mängd modeller som går att använda och det är även möjligt att skapa egna modeller för att kunna syntetisera data.

Modellen tränas på all data, det vill säga att datasetet behöver inte delas upp i träning och validering.

För matsvinnsdata testades följande modeller som finns i SDV:

- Tabular Preset
- GaussianCopula
- CTGan

- CopulaGan
- TVAE

Steg 5. Generera syntetiska data från varje modell

Efter att modellen är skapad så kan nya syntetiska data genereras. Oftast går det att välja hur mycket data som genereras. Det är dock viktigt att tänka på att den data som skapas är baserad på originaldata och därmed tillkommer inte ny information även om mer syntetiska data genereras än vad det finns originaldata. Det kan också vara lättare att jämföra två dataset som är lika stora, så för en senare kvalitetskontroll kan det vara bra att inte generera mer data. Däremot skulle det kunna öka säkerheten hos den syntetiska data om det finns mer data då det gör det svårare att urskilja unika datapunkter. Det kan också möjliggöra bättre träningen av AI-modeller med ett större dataset.

För matsvinnsdata genererades 5 olika dataset baserat på ovannämnda modeller.

Steg 6. Utvärdera de olika modellerna

När den syntetiska data har genererats kan den utvärderas och jämföras med originaldata. Det är bra att ha automatiska verktyg som gör detta men det kan också vara bra att manuellt jämföra korrelationsmatris för att få en visuell jämförelse och på så sätt förstå vad som fungerar eller inte fungerar.

För matsvinnsdata gav alla modeller någorlunda resultat men konstigheter kunde uppstå, exempelvis så uppstod ofta ojämna matsvinnsfördelningar baserat på Nodtyp. Bland annat blev det flera rapporter från skolor och färre rapporter från förskolor.

Steg 7. Val av modell och förbättring

När de olika modellerna har jämförts kan den modell som fungerar bäst plockas ut och arbetas vidare med. Det kan behövas göra många små ändringar för att få den syntetiska data till att bli så bra som möjligt, kvalitetsmässigt och anonymt.

För matsvinnsdata var de tre modeller som visade på bäst resultat; CTGan, GaussianCopula och TVAE. Utöver detta delades datasetet upp i tre delar baserat på Nodtyp då detta var en enkel lösning på en del av de problem som uppstod. Efter denna förändring fick TVAE bäst resultat och därmed valdes denna modell för slutgiltig utvärdering.

Steg 8. Utvärdera och manuell finslipa

Gör en sista utvärdering av den slutgiltiga syntetiska data som tagits fram. Om den syntetiska data inte uppfyller de krav som finns på anonymitet kan en sista manuell finslipning göras där någon data tas bort eller ändras.

En utvärdering gjordes av data men ingen manuell förändring gjordes då det inte fanns ett behov av detta. Följande är ett par utvalda bilder som visar på likheter och skillnader.

I Figur 3 och Figur 4 nedan går det att se en korrelationsmatris för originaldata respektive syntetiska data

Över lag så sparas korrelationerna för det syntetiska datasetet men det blir självklart små variationer. Vissa variationer kan vara större än andra och hade det funnits ett slutmål med användningen av den syntetiska data hade det varit viktigt att se till att de korrelationer som är nödvändiga är så nära originalet som möjligt.

Nedan finns även en exempelbild, Figur 5, som visar fördelningen av data inom en viss kategori, i detta fallet antalet ätande.

I Figur 5 går det att se hur den nya syntetiska data liknar originaldata men med små skillnader vilket är vad vi förväntar oss.

Steg 9. Färdiga syntetiska data

Nu finns ett nytt dataset med högre anonymitet än originaldata och därmed högre säkerhet för de individer som ingår i datasetet.

Det nya syntetiska datasetet kan nu användas till förbestämda användningsfall.

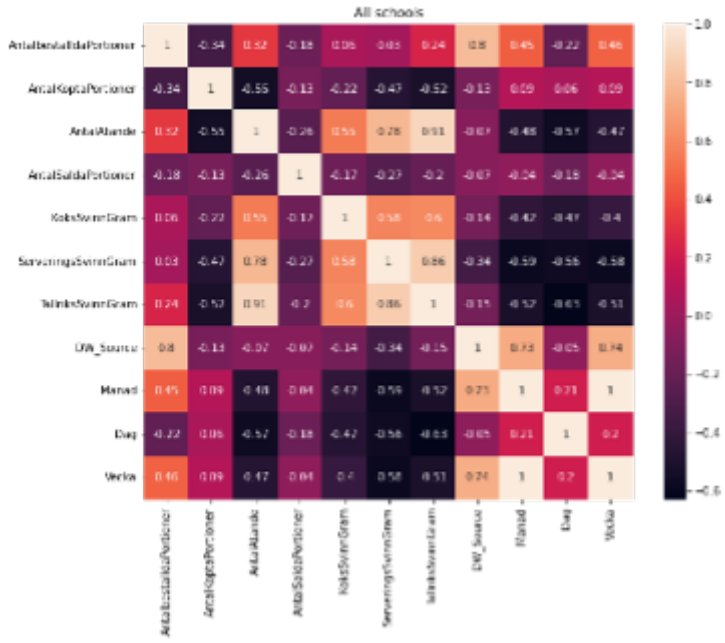
Slutkommentar syntetiska matsvinnsdata

Det finns en hel del förbättringar som hade kunnat göras för att skapa ännu bättre kvalitet på den syntetiska matsvinnsdata. Exempelvis skulle begränsningar kunna sättas på hur många rapporter en enhet kan ha på ett år, och så vidare. Målet i denna förstudie var dock inte att skapa helt användbara och tillförlitliga syntetiska data utan att pröva på tekniken.

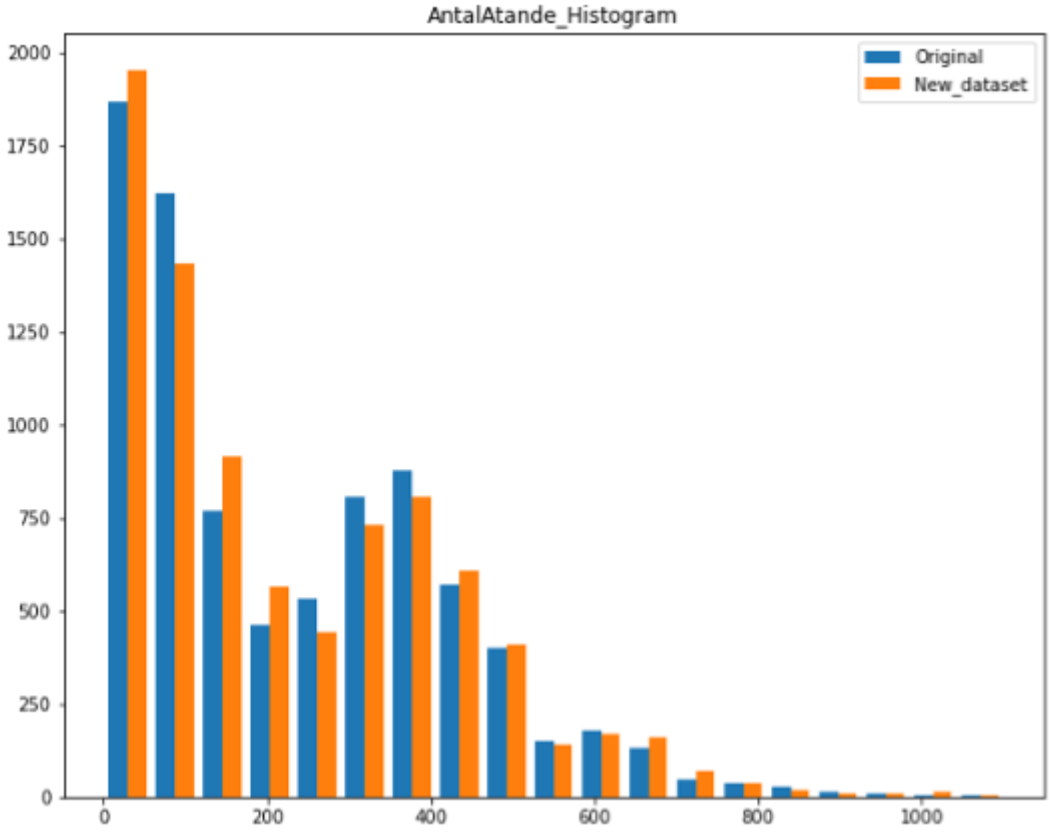
Vad som har beskrivits ovan bör inte ses som en absolut guide utan mer en vägledning eller möjligt tillvägagångssätt.

Allt jobb utfördes på en modern laptop och gick relativt snabbt. De modeller som tränades och användes var med andra ord inte speciellt krävande hårdvarumässigt, men det var även en mindre mängd data.

Figur 3: Korrelationsmatris över originaldata.



Figur 4: Korrelationsmatris över syntetiska data.



Figur 5: Histogram över antalet rapporter (y-axeln) som innehåller antalet ätande (x-axeln).

5 Vad är anonym data?

Det finns flera metoder för att mäta anonymitet och beroende på metod går det att till en viss grad att påvisa anonymitet men aldrig fullt ut. I denna rapport har sådana metoder diskuterats ytters lite, inte för att sådana metoder inte är viktiga utan för att problemet med kvalitét mot anonymitet kvarstår. Vidare verkar det finnas flera problem med hur anonymitet diskuteras vilket kommer att diskuteras i denna del av rapporten.

Ett missförstånd som verkar uppstå i diskussion kring anonyma data är att den direkta betydelsen och den praktiska implementeringen inte är densamma. Anonyma data betyder att det inte under några omständigheter går att få ut någon information om individer på något sätt. I praktiken när data anonymiseras handlar det dock mer om en skala där personer går från direkt identifierbara mot att bli anonyma men data blir oftast aldrig helt anonyma för då skulle det inte längre finnas något värde i den data.

Två exempel på "attacker" som kan utföras vilket förtydligar hur en person skulle kunna utvinna information ur annars relativt anonyma data.

Det första exemplet är relativt enkelt. Anta att ett dataset har med bland annat födelsedatum. I detta dataset förkommer en person som är 105 år gammal. Problemet är att det finns så få personer som är 105 år gamla inom Sverige, speciellt på kommunnivå, att det direkt går att identifierad individen. Detta trots att namn, adress, kön, et cetera har tagits bort från datasetet. Läggs område eller postnummer till, kan en person trots att den har ett relativt vanligt födelseår lätt identifieras beroende på vilka åldrar andra personer har inom detta område eller postnummer. Vidare kanske kön förekommer och helt plötsligt är det möjligt att identifiera en stor mängd människor. Med andra ord kan det vara relativt lätt att identifiera individer med relativt få attribut. Attribut som också kan vara väldigt viktiga och därmed inte bör tas bort.

Det andra exemplet lyder som följande. Anta ett dataset som innehåller datum en individ har besökt en viss förvaltning samt en kategorisk anledning. En person som vet att deras partner har besökt en kommun vid två tillfällen kommer relativt lätt att kunna härleda ur datasetet vilken den kategoriska anledning är då det är relativt få individer som troligtvis besökt kommunen på exakt samma dagar även om de har besökt kommunen vid flera tillfällen. Likväl skulle personen kunna komma över positionsdata ur sin partners mobiltelefon och på sådant sätt tillsamman med information om var förvaltningsbyggnaden ligger härleda den kategoriska anledningen till partners besök. Med andra ord kan det finnas annan data som är okänd för de som delar vad de tror är anonyma data men som sedan visar sig går att använda tillsammans med annan data för att utvinna information om individer.

För att undkomma ovanstående problem behövs ytterligare anonymiseringsmetoder där granulliteten minskar och/eller där data ändras, tas bort eller läggs till. Något som är svårt att göra utan att påverka kvalitén.

När syntetiska data genereras så finns statistiska förhållanden och relationer med som gör att kvalitén påverkas till en mindre grad. Men det som är mest intressant är att även om det går att hitta kopplingar mellan originaldata och den syntetiska data så är det inget ett-till-ett förhållande. Detta skapar en säkerhet som är svår att uppnå med traditionella metoder. Även om det skulle gå att relatera en individ från det syntetiska datasetet till verkligheten så har troligtvis alla attribut mer eller mindre ändringar vilket gör det svårt att ur det syntetiska datasetet utvinna exakt information om individen. Det går inte med säkerhet att säga exakt till vilken grad attribut stämmer med verkligheten. Detta räknas dock idag fortfarande som personuppgifter, då det går att identifiera en individ trots stora osäkerheter vad som stämmer eller inte stämmer överens med det syntetiska datasetet.

Avslutningsvis skall det nämnas att även syntetisering kan lida av kvalitét och säkerhetsdilemmat. Beroende på vad användningen är och till vilken grad som originaldata måste vara representerad i den syntetiska data uppstår problemet med kvalitén och säkerheten. I det övergripande exemplet om matsvinn skulle det kunna vara viktigt att den syntetiska data inte innehåller exakta antalet rapporter från varje enhet samt har en viss variation över antalet ätande och matsvinnet. Det går exempelvis att räkna antalet ätande en viss dag på en viss enhet för att lista ut vilka syntetiska data som skapats baserat på den enheten och sedan skulle det vara möjligt att eventuellt utvinna information om enheten(individen) ur syntetiska data.

6 Juridisk diskussion

När detta arbete startades fanns en förhoppning att syntetiska data skulle räknas helt som anonym data men så är inte fallet. Dels insåg vi under detta arbete att så troligtvis inte var fallet, dels gick Integritetsskyddsmyndigheten (IMY) ut i mitten av november med följande uttalande "Det pågår mycket försök med datorgenerade data, så kallade syntetiska data, som skulle kunna bli ett alternativ till anonymisering. Om syntetiska data skapas av personuppgifter så gäller GDPR"[7].

Det är dock inte helt klart vad IMY menar då uttalande kan tolkas på flera sätt. Det är också märkligt att de särskiljer traditionella anonymiseringsmetoder och syntetisering som anonymiseringsmetoder. All data som har anonymiserats, oavsett metod, måste självfallet kontrolleras så att sagd data är anonym. Det vore konstigt om IMY menar att anonym data skall räknas som personuppgifter bara för att den data har skapats genom syntetisering baserat på personuppgifter. Grunden till denna inkonsistenta bedömning ligger troligtvis i att det är skillnad på den definition av anonymitet som används juridiskt och hur anonymisering går till i verkligheten.

Trots att detta uttalande från IMY till viss del hindrar den potentiella användningen av syntetiska data just nu finns det ett stort intresse för syntetisering som anonymiseringsmetod och förändringar kan komma att ske. En utredning från Europakommissionens gemensamma forskningscentral (JRC) så kommer syntetiska data vara en viktig del i att möjliggöra AI i Europa[8]. JRCs rapport tillsamman med ett stort intresse och ett behov av att kunna nyttja data bland både offentlig och privat sektor kommer troligtvis leda till förändringar framöver.

Det finns så klart andra intressanta juridiska frågor rörande syntetisering som anonymiseringsmetod. En sådan fråga är om originaldata får användas till att skapa syntetiska data. Då den data som används för att skapa den syntetiska data är personuppgifter faller den under GDPR. En av huvudpelarna i GDPR är att användning av data är begränsat till det eller de specifika syften som har kommunicerats i samband med insamlingen av data. På så sätt är skapandet av syntetiska data begränsat av GPDR. Dock får kommuner nyttja sin data till att utveckla sina tjänster och skulle syntetisering vara ett steg i denna process bör detta falla inom ramen av ett uttalat användningsområde. Därmed bör GDPR inte påverka en kommuns skapande av syntetiska data från ett dataset som innehåller personuppgifter. Men då den skapade data också räknas som personuppgifter och faller under GDPR blir det väldigt viktigt att de lagkrav som ställ efterlevs för den nya syntetiska data.

Vidare finns det en annan intressant frågeställning som bör utredas. Det finns en möjlighet att AI-modell som genererat syntetiska data, eller kunskap om den modellen, används till att återskapa originaldata eller delar av originaldata. Om en modell ses som ännan"information leder det då till att den syntetiska data som framställts bör ses som pseudoanonymiserat till dess att AI-modellen är borttagen?

Ett sista eventuellt juridiskt problem som uppstår på en kommun är om någon skulle kunna begära ut denna AI-modell som en offentlig handling. Kommunen måste kunna hålla modeller hemliga för att säkerställa anonymiteten i den data som skapats.

[7] IMY. 2022-11-18. https://www.imy.se/verksamhet/dataskydd/innovationsportalen/vanliga-fragor/vi-hanterar-bara-anonymiserade-personuppgifter-da-kan-vi-val-bortse-fran-gdpr/

[8] EU's JRC. Multipurpose synthetic population for policy applications. 2022. https://publications.jrc.ec.europa.eu/repository/handle/JRC128595

7 Etik

Syntetiserade data, även om den innehåller personuppgifter, har en större säkerhet än originaldata. Bara kunskapen om att datasetet är syntetiskt skapar en viss ovisshet om hur ett specifikt attribut skulle kunna korrelera med originaldata, oavsett om det gör det eller ej. I fall där data delas kan det finnas en fördel att syntetisera den data som delas för att öka säkerheten. Det bör alltså övervägas att nyttja syntetisering i de fall där det inte är nödvändigt för användningen att anonymisera den data som finns men där det ändå kan skapa ett bättre integritetsskydd för de individer som ingår denna data.

Som nämns tidigare i denna studie, så finns det även en otrolig potential med att kunna dela data och där kan syntetisering vara en möjliggörare, speciellt med tanke på att tekniken utvecklas och förutsatt att juridiken blir klarare. Offentliga verksamheter bör därmed arbeta med att röja de hinder som finns idag för att kunna nyttja syntetisering.

Utöver detta kan syntetiseringsmetoder användas för att minska bias och på så sätt skapa bättre och mer jämlika AI-modeller. Det finns alltså flera stora potential samhällsvinster med syntetiska data. Men syntetisering bör utövas försiktigt, då om syntetisering utförs felaktigt så kan bias öka i stället. Med andra ord måste denna teknik likt alla andra tekniker utövas ansvarsfullt.

8 Sammanfattning

Det går inte att garantera att syntetisering som anonymiseringsmetod skapar helt anonyma data men metoden, om utförd korrekt, är oftast bättre än andra anonymiseringsmetoder på så sätt att den syntetiska data är mer anonym och mer användbar. Är det lite oklart rent juridiskt hur syntetisering kan användas som anonymiseringsmetod. Denna oklarhet medför en begränsning hos syntetisering som en lösning på de datadelningsproblem kommuner står inför idag.

Det finns dock redan idag vissa fördelar med syntetisering. Bland annat kan metoden öka skyddet av personliga data och metoden kan även användas för att minska bias i den data som finns för att skapa mer rättvisa AI-modeller. En nackdel är dock att syntetisering är en mer komplicerad metod jämfört med andra anonymiseringsmetoder och därmed kräver mer resurser.

Just nu sker det mycket utveckling när det kommer till syntetisering, både i allmänhet och som anonymiseringsmetod. Användningen av syntetiska data spås öka ordentligt och det finns en enorm potential. Tekniken kommer med stor sannolikhet bara bli bättre framöver och det juridiska spelrummet kan komma att ändras.

Sveriges kommuner och regioner behöver ställa högre krav på regeringen i allmänhet när det kommer till att möjliggöra nyttjandet av invånares data till invånares nytta. Syntetisering som anonymiseringsmetod är fortfarande en potentiell lösning som bör lyftas fram och diskuteras mer i detta sammanhang, men även andra lösningar är välkomna.