

Nya metoder för text- och dataanalys inom IAF:s tillsynsarbete

Tillsynsforum 22 okt, 2021



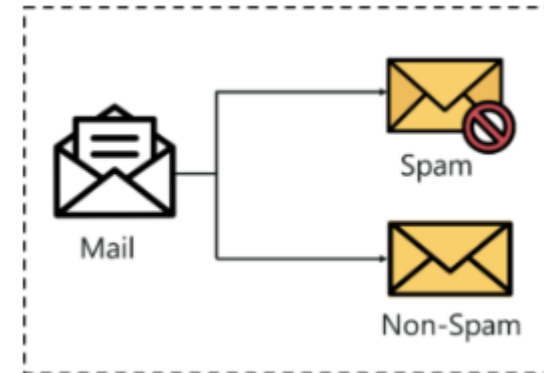
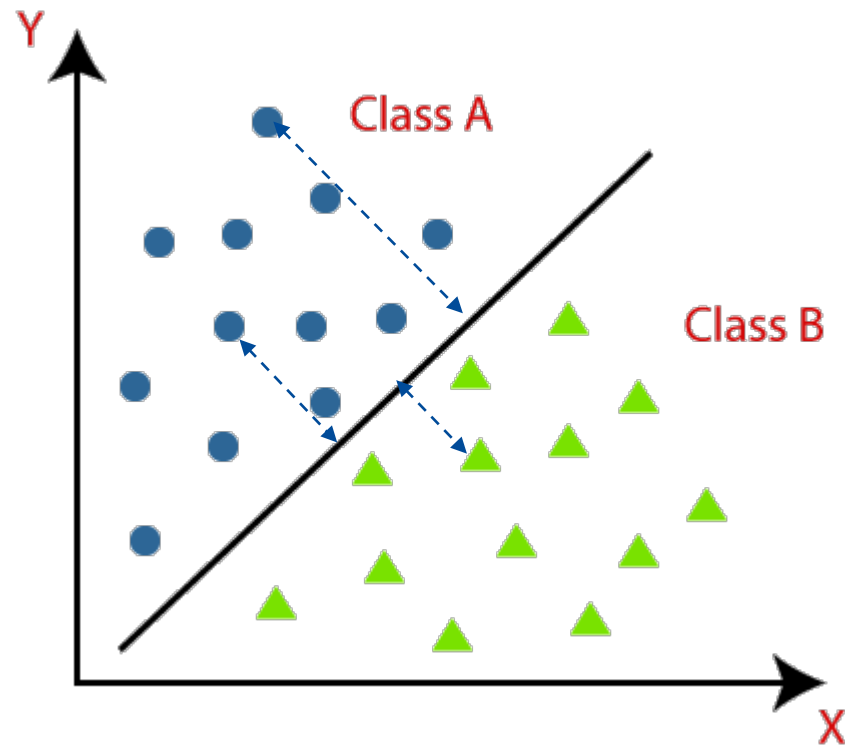
Behovet av artificiell intelligens för text- och dataanalys

- Mängden data IAF (och andra myndigheter) förfogar över växer ständigt
- Manuella granskningar kan inte täcka alla behov
- Ökat behov av automatisering av (delar av) ärendegranskningar
- Artificiell intelligens (AI) möjliggör stora automatiska kvalitativa undersökningar av tillgängligt data:
 - Textdata (fritext och/eller textdata som kan “kategoriseras”)
 - Numeriskt data
- En metod är dataklassificering via AI baserat på resultaten från manuella granskningar

Artificiell intelligens för text- och dataanalys

- Klassificering av data sker via träning av olika matematiska modeller
- Träningen görs med träningsdata där utfallet är känt, s.k. *vägled maskininlärning*
- Den bästa modellen för ett visst problem kan hittas genom uttömmande sökning och sofistikerad testning
- Den färdigtränade modellen kan sedan göra *förutsägelser* för data där utfallet är okänt
- För IAF:s räkning så kan data med tillräckligt säker förutsägelse undantas från manuell granskning, vilket ökar effektiviteten

Artificiell intelligens för text- och dataanalys



IAF:s eget analysverktyg

- Webb-gränssnitt (C#) för enkel användning
- Klassificering av data görs i Python med standardbibliotek (sklearn, m.m.)
- Textmaterial krypteras vid behov (ex. vid personuppgifter) och kan filtreras på onödiga ord

IAF Automatisk Dataklassificering (beta)

Projektnamn:

Specifika databastabell att klassificera:

Host:

SQL-användare:

SQL-lösenord:

Välj ODBC drivrutin:

Databas:

Tabell:

Välj kolumner:

Klasskolumner: ☐ bed_dat ☐ hplid ☐ hplStat ☐ statDat ☐ TextArbmarkn ☐ TextArbmRelevant

Dataskolumner: ☐ bed_dat ☒ hplid ☐ hplStat ☐ statDat ☐ TextArbmarkn ☐ TextArbmRelevant

Välj 1 klasskolumn, 2 dataskolumn(er)

Klass distribution:

- Klassificering fel förekommer 25 ggr
- Klassificering ja förekommer 255 ggr
- Klassificering "" / NULL förekommer 22 ggr (Klassvärdet med tomma strängen eller NULL anses båda klassificerad)
- Totala antalet rader i tabellen är: 302

Ange en unik nyckel: ☒ hplid ☐ TextForut

Ange om valda kolumner är text eller numeriska:

Text: ☒ TextForutRelevant ☐ hplid ☒ TextForut

Numeriska: ☐ TextForutRelevant ☒ hplid ☐ TextForut

Ange följande:

Max antal använda rader i databastabell:

Max antal iterationer:

Storlek på testdata (%):

Algoritm & förbehandling:

Bedömningsmetod:

Behandling av textdata: ☐ Ingen ☒ Kryptering ☐ Stoppord

Verbose (on/off): ☒

Variabelreducing: ☐ Ingen ☐ RFE ☒ PCA

Antal variabler:

Start

Träningsdata: Noggrannhetstabell

	precision	recall	f1-score	support
Ja	0.96	1.00	0.98	52
Nej	1.00	0.50	0.67	4
accuracy	0.98	0.75	0.82	56
macro avg	0.97	0.96	0.96	56

Classification report for SGD1-NON with nfeatures: 381

Utvärdera

Nya klassificeringer:

Hämta resultatet:

SQL snippet: `SELECT TOP(22) A.[TextForutRelevant] A.[hplid] A.[TextForut] B.[class_name] B.[class_time] B.[class_algorithm] FROM [Arbetsdatabas].[aterkommande_automat].[aterkommande_hpl_bearb] A INNER JOIN [Arbetsdatabas].[aterkommande_automat].[AutoKlassificering] B ON A.hplid = B.[unique_key] WHERE B.[class_user] = 'autoclass_user' AND B.[table_name] = 'aterkommande_automat_aterkommande_hpl_bearb' ORDER BY B.[class_time] DESC`

Utiliter: [results.csv](#) [output.txt](#) [error.txt](#)

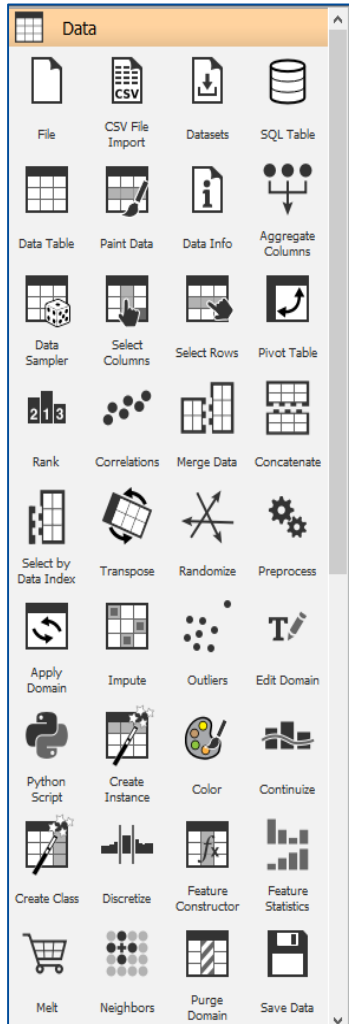
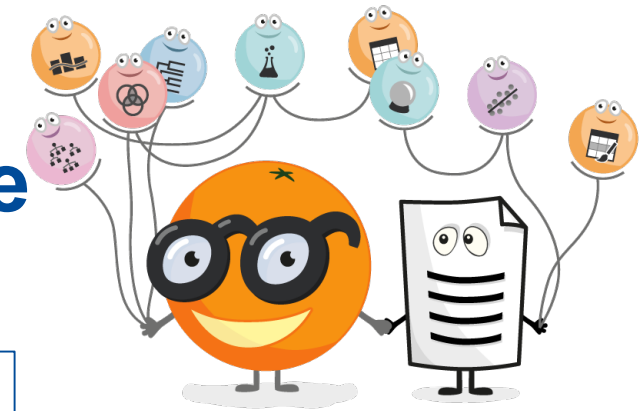
Totalt: 96.3% (22 st.), Ja: 96.3% (22 st.)

Förvirringsmatris

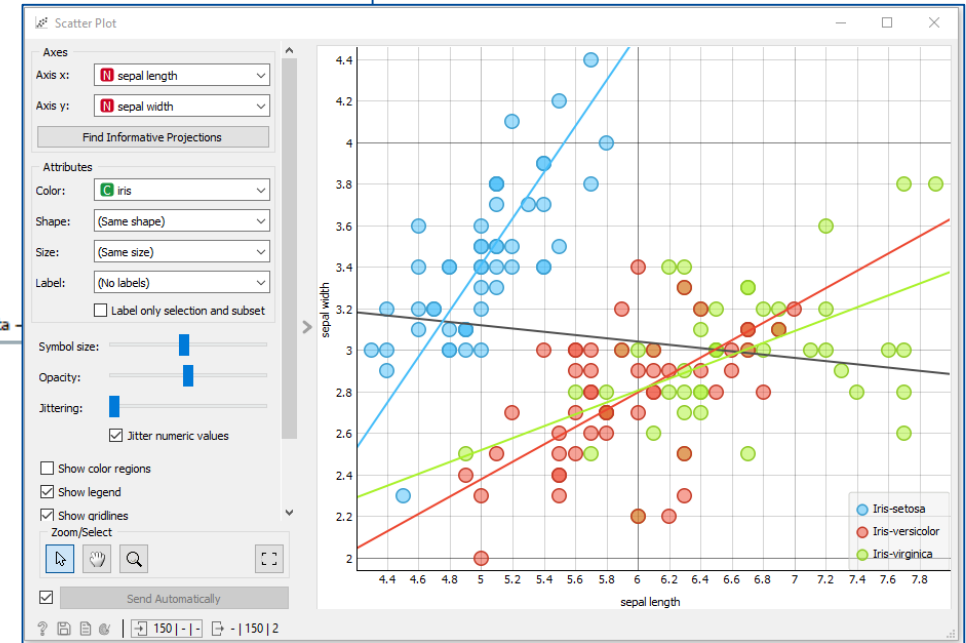
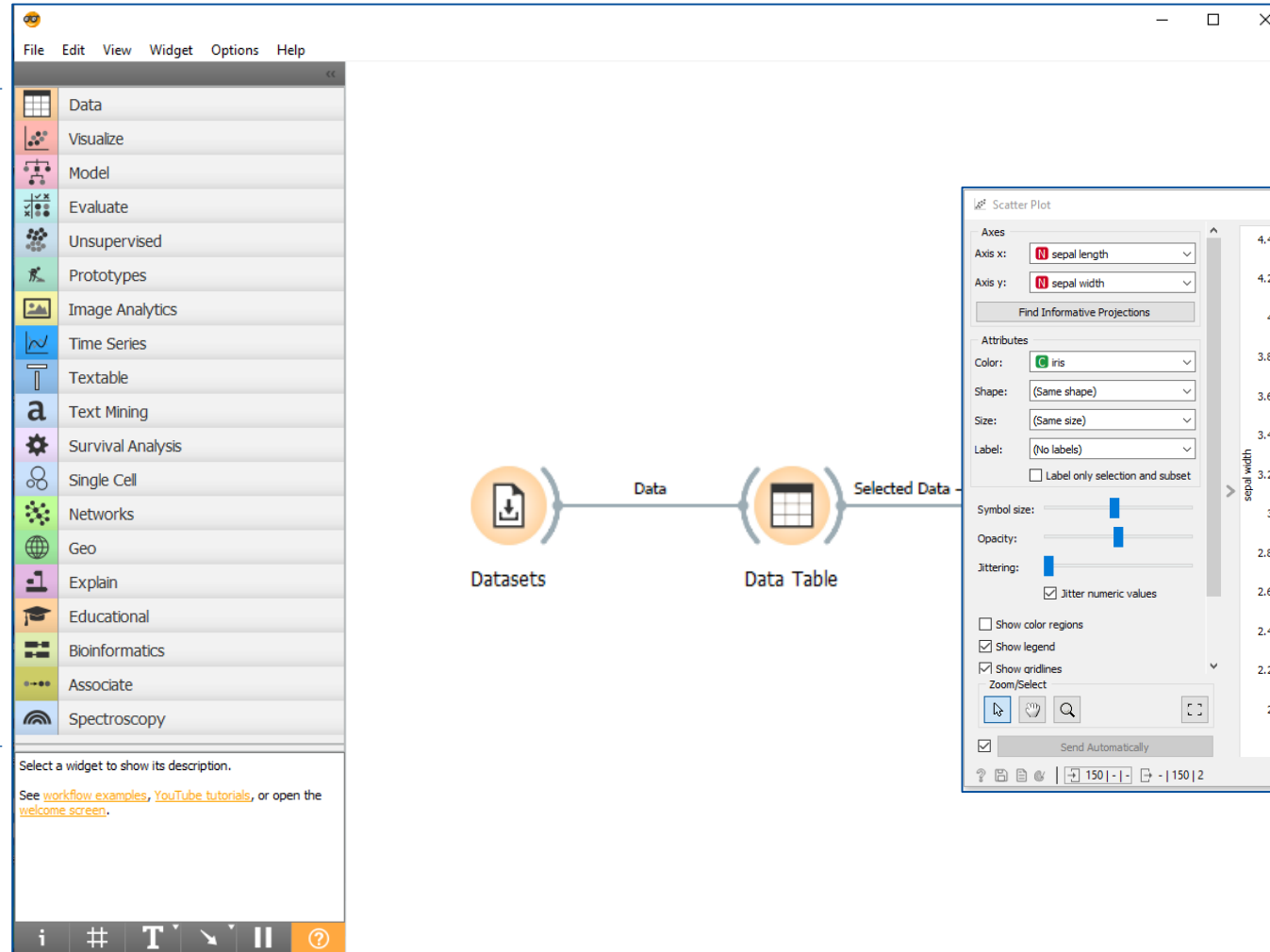
	[[52 0]
	[2 2]]

Python

Användbar tredje-parts-programvara: Orange



Verktyslådor

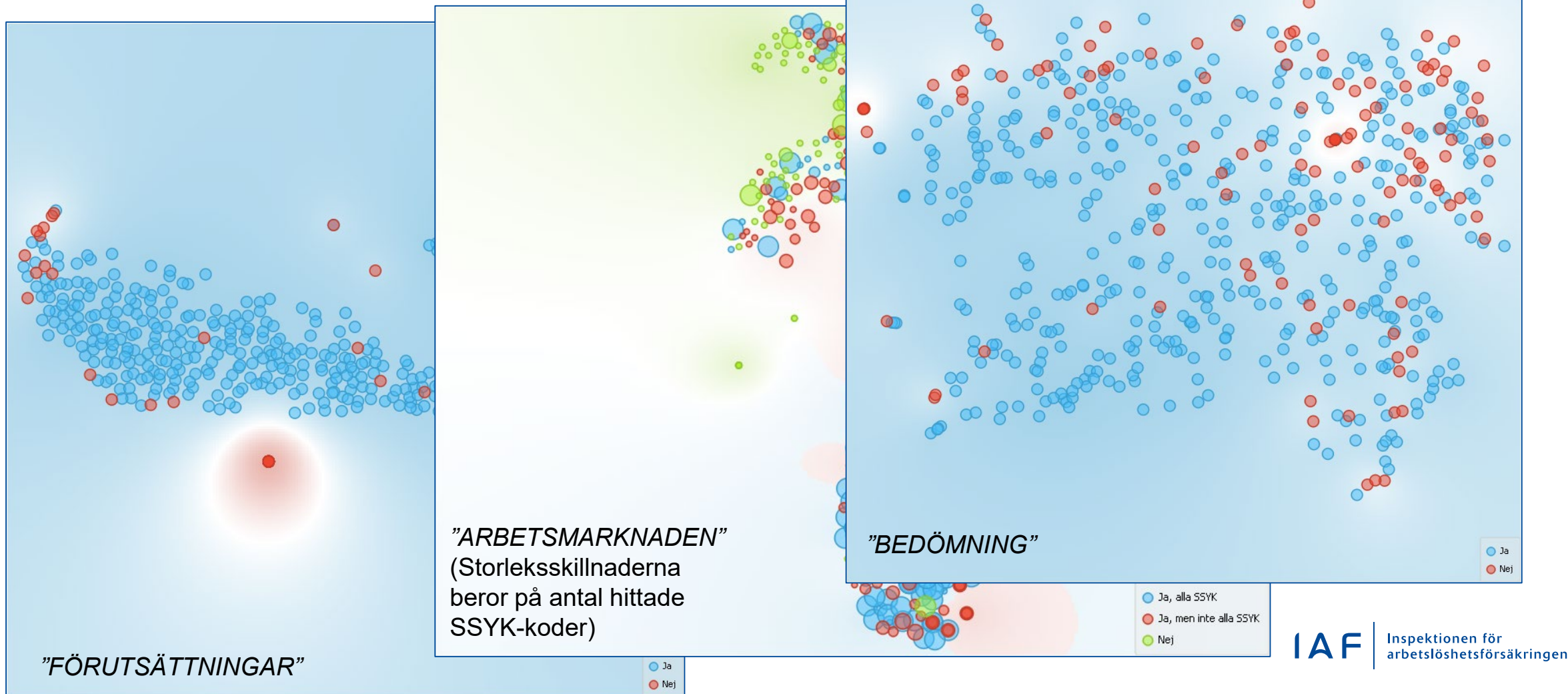


Exempel 1 – Automatiserad bedömning av fritextfält i AF:s planeringsverktyg

- Utgår från frågeställningarna i *Återkommande granskningar av handlingsplaner och aktivitetsrapporter, 2020:15*.
- Specifikt: från den arbetsmarknadspolitiska bedömningen i planeringsverktyget:
 1. Finns det relevant beskrivande text under rubriken "FÖRUTSÄTTNINGAR"? ("Ja" / "Nej")
 2. Finns det relevant beskrivande text under rubriken "ARBETSMARKNADEN"? ("Ja, alla SSYK" / "Ja, men inte alla SSYK" / "Nej")
 3. Finns det relevant beskrivande text under rubriken "BEDÖMNING"? ("Ja" / "Nej")
- Resultatet presenteras som förstudie i intern rapport 2021/55: *Automatiserad bedömning av fritextfält i Arbetsförmedlingens planeringsverktyg*

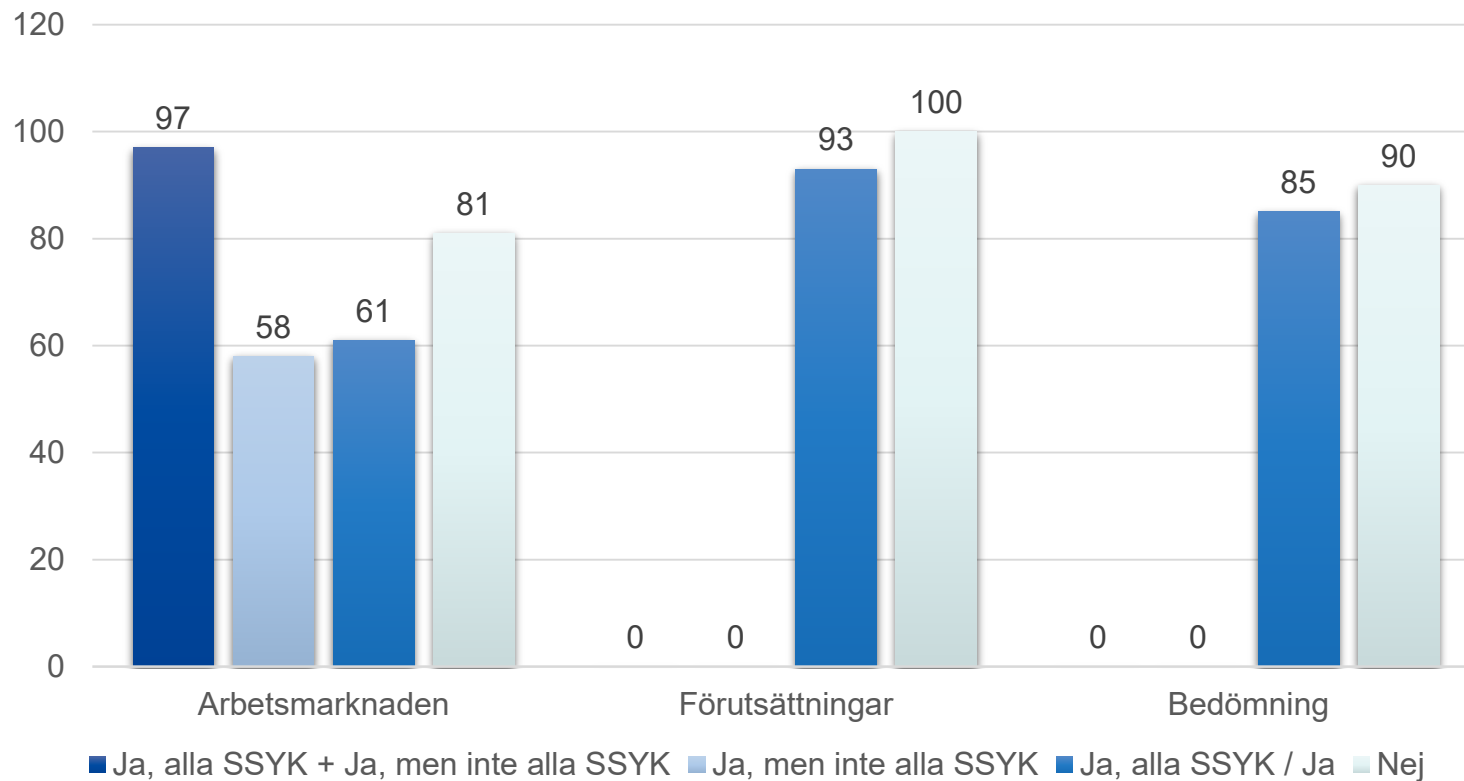
Grafisk representation av texterna (t-SNE)

- Sensmoral: klustrade datapunkter är svårare att separera i tydliga klasser

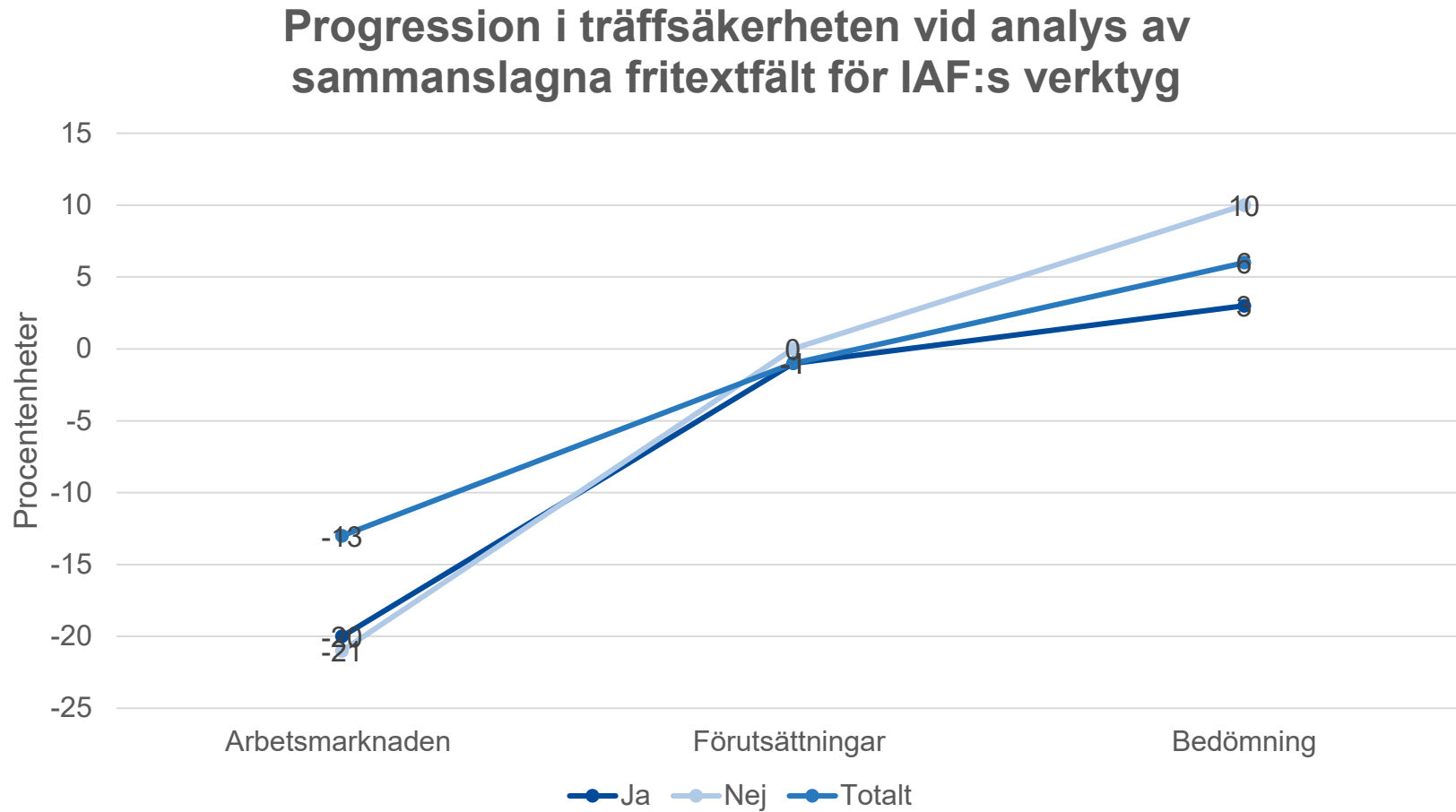


Resultat – klassificering av texter fr 2019

Genomsnittlig individuell träffsäkerhet för IAF:s verktyg (%)



Resultat – klassificering av texter fr 2019

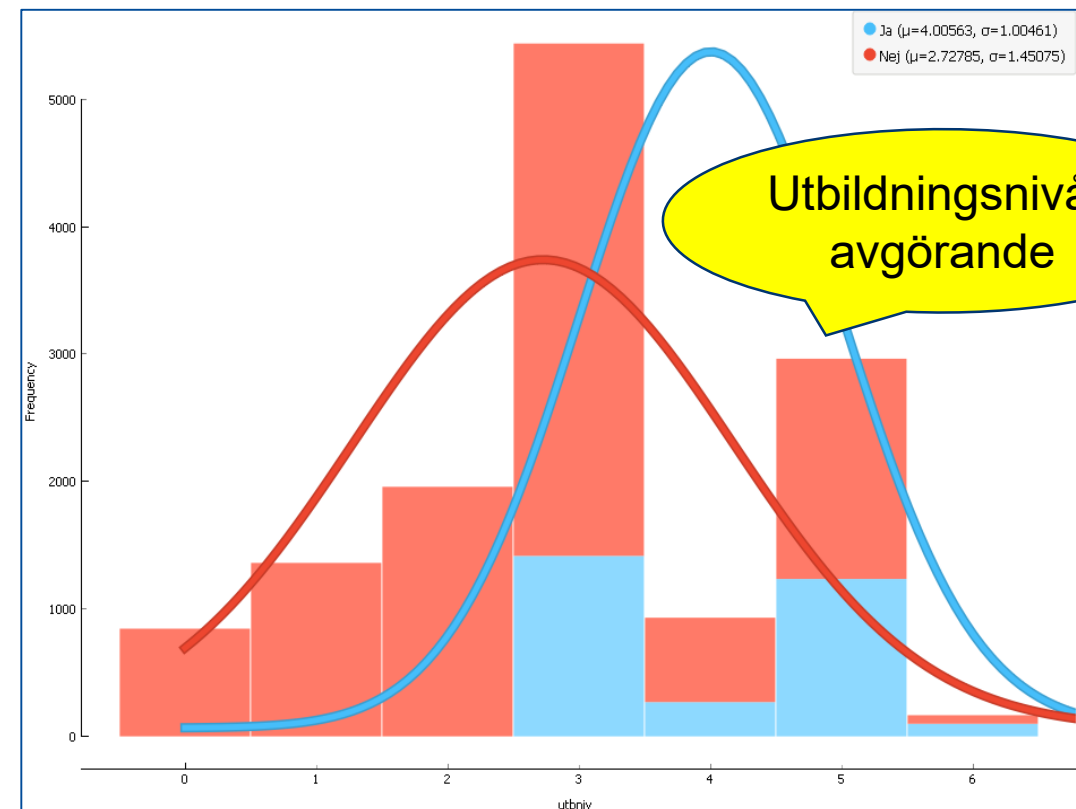
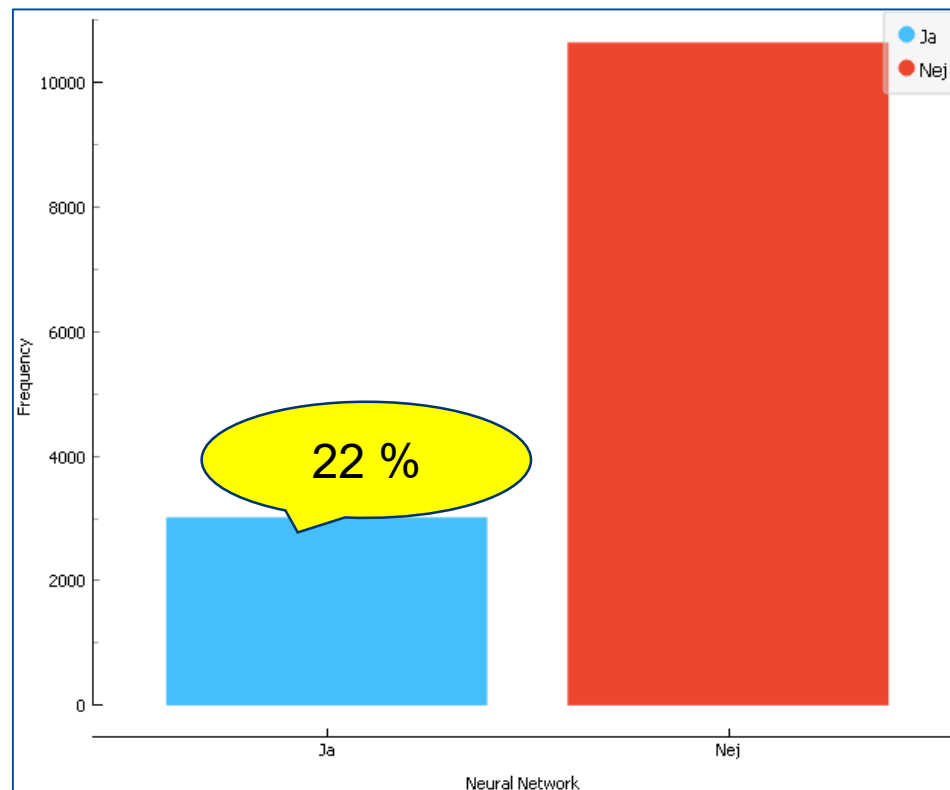


Exempel 2 – långtidsarbetslöshetens utveckling

- Träningsdata: Personer inskrivna på Arbetsförmedlingen under 2019, var arbetslösa i minst 12 månader och sedan fick arbete eller inte inom ytterligare 6 månader.
- Data med okänt utfall: personer som blev långtidsarbetslösa september 2021.
- Populationen uppgick till totalt 112 500 personer
- Typiskt klassificeringsproblem med två klasser: "Ja" och "Nej"

Exempel 2 – långtidsarbetslöshetens utveckling

- Hur många som blir långtidsarbetslösa sept 2021 får arbete inom 6 månader?

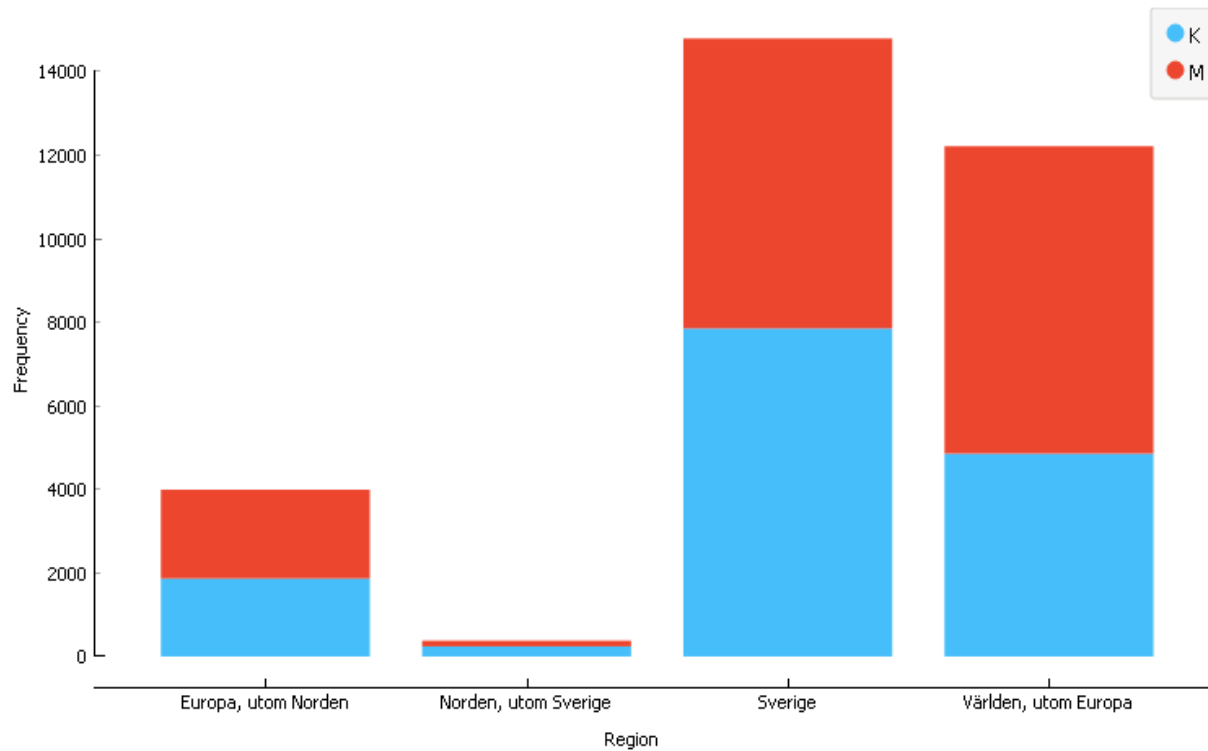


Exempel 3 – potentiella felaktiga utbetalningar (FUT) från a-kassorna

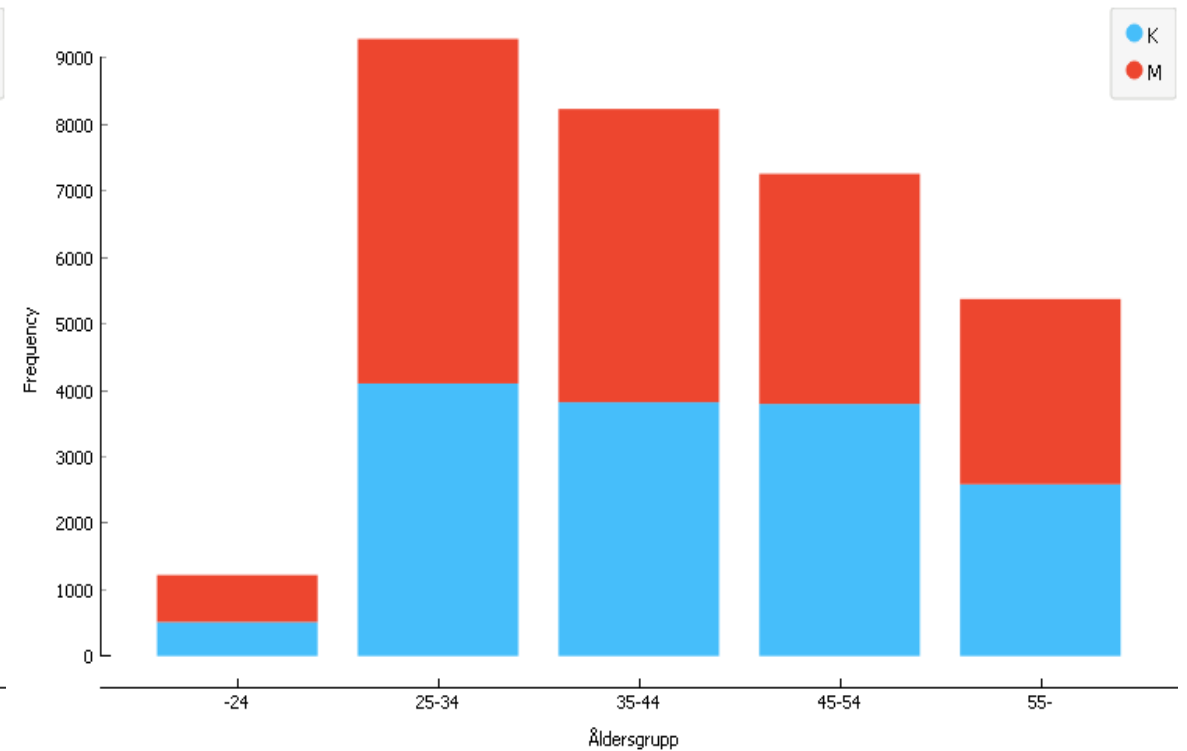
- Idé: identifiera utbetalningar som inte genererat återkrav men som *kanske* borde ha gjort det
- Träningsdata: 265 000 utbetalningar fr 2020 som inte genererat återkrav ("Nej"), och sådana som genererat återkrav ("Ja")
- Till utbetalningarna knyts ärendedata och ärendehistorik, etc.
- Bland "felklassificerade" utbetalningar utan återkrav kan vi identifiera potentiella FUT som med viss sannolikhet borde tillhöra gruppen "Ja"

Exempel 3 – typiska återkrav med 95 % sannolikhet

■ Födelseland

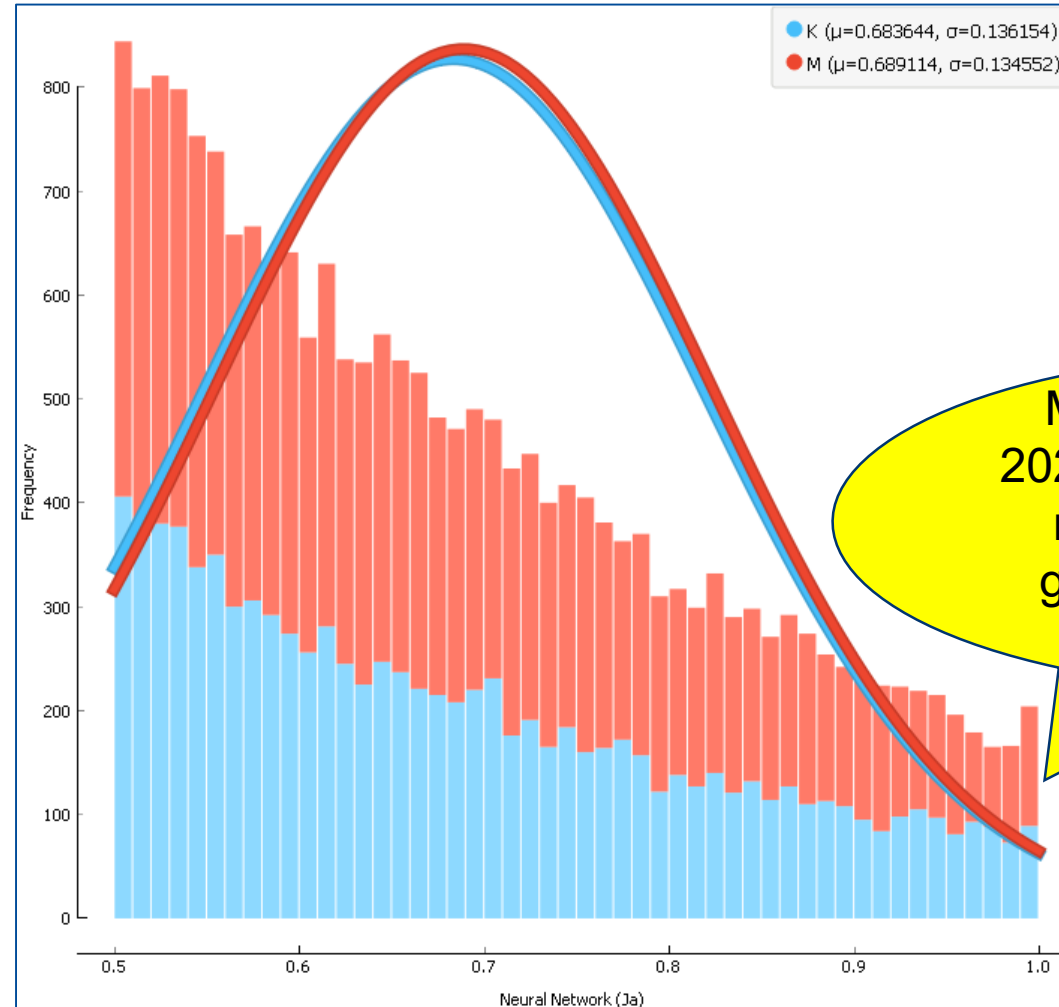


■ Åldersgrupp



Exempel 3 – potentiella FUT

- Finns utbetalningar från a-kassorna som borde genererat återkrav men inte gjort det?



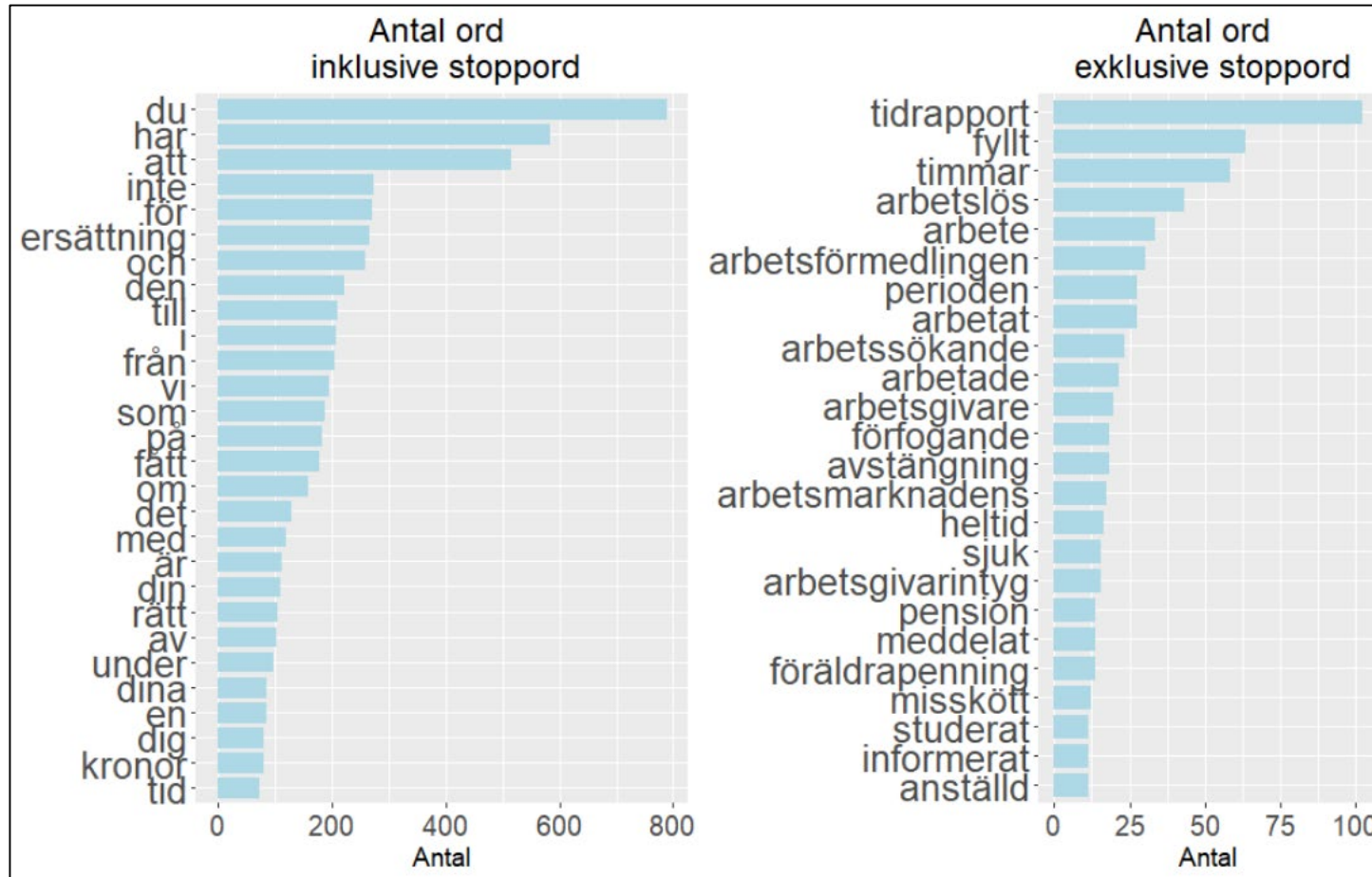
Minst 900 ej återkrav fr
2020 *borde* enligt modellen
med 95 % sannolikhet
genererat ett återkrav –
potentiella FUT!

Grunder till återkrav

- Om du får ersättning från a-**kassan** som du inte har rätt till kan du bli återbetalningsskyldig. Om den felaktiga ersättningen beror på att du har lämnat felaktiga uppgifter till a-**kassan** så blir du alltid återbetalningsskyldig.
- Datamaterialet består av texter i form av beslutsmotiveringar
- Datamaterialet är oklassificerat
- Använder icke-vägledda ML metoder
- Textmining (Klusteranalys, temamodellering & PCA)

Deskriptiv analys

Frekvensdiagram över antalet ord



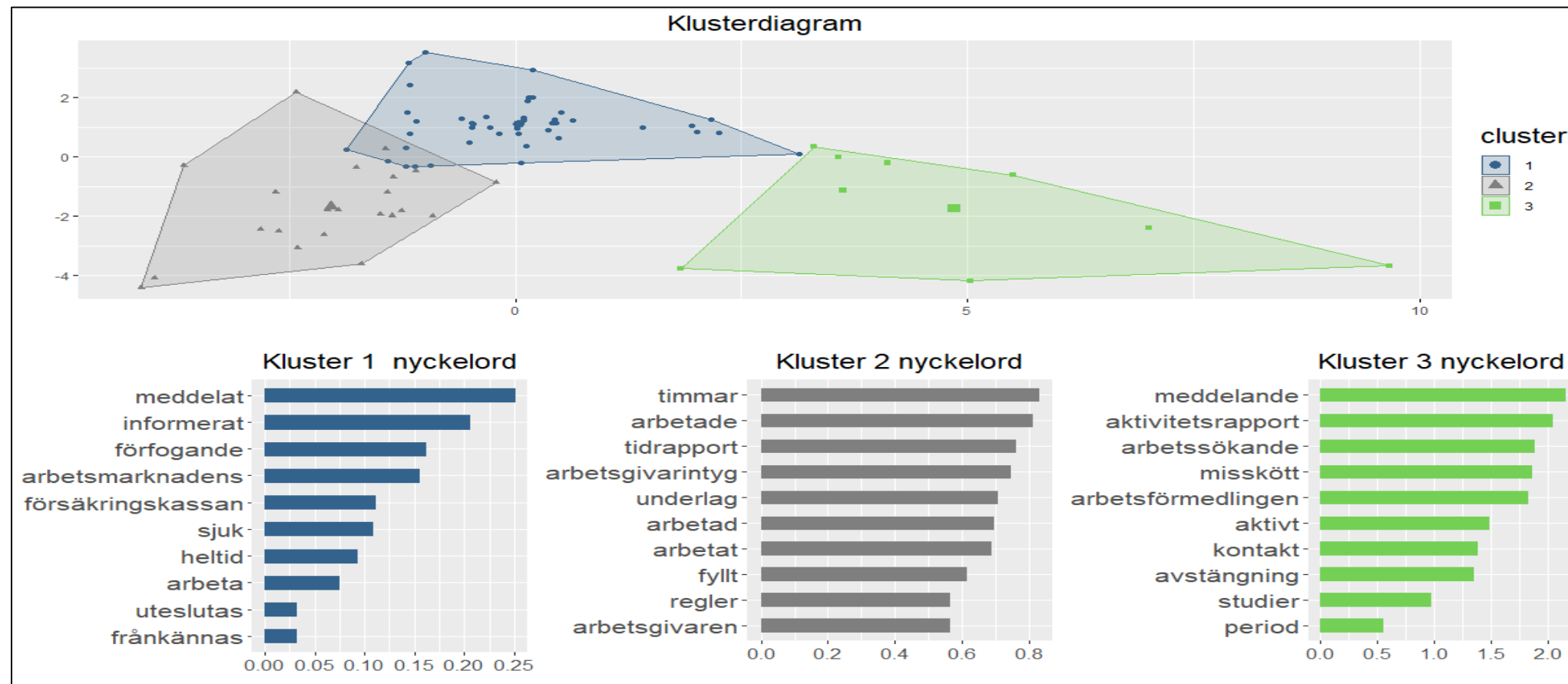
Deskriptiv analys

Ordnätverk av bi-gram



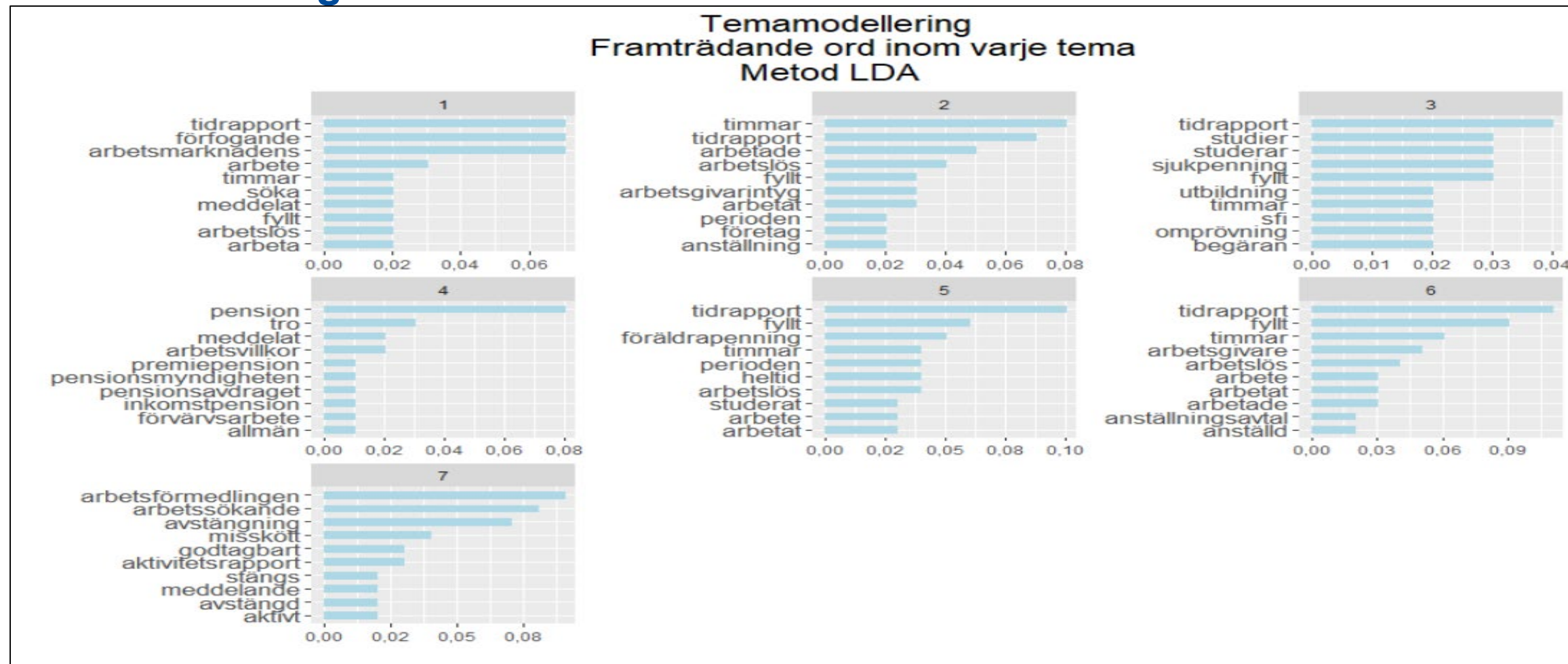
Resultat

Klusteranalys



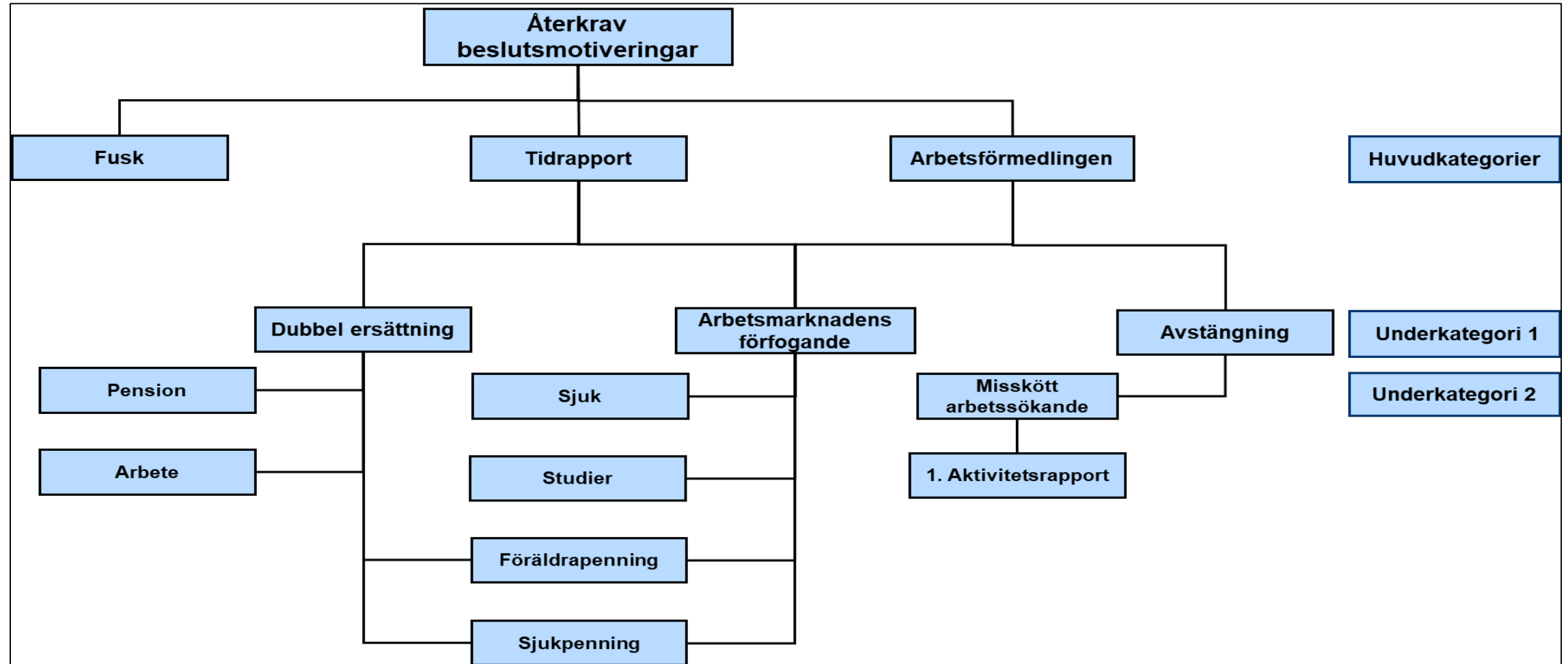
Resultat

Temamodellering



Resultat

Sammanfattning och kategorisering



IAF

Inspektionen för
arbetslöshetsförsäkringen

Frågor eller kommentarer?

