Kandidatuppsats i Statistik

Undersökning hur medarbetarundersökningar kan prediktera anställdas avsikt att lämna arbetsplatsen

Kopplingar mellan att inte se sig jobba kvar och variabler så som trivsel och kompetensutveckling

Oskar Storberg
William Wiik

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

Vårterminen, 2023 | ISRN-nummer

Handledare: Annika Tillander, universitetslektor

Examinator: Linda Wänström, universitetslektor, Docent

Sammanfattning

Syftet med denna studie är att utveckla en prediktiv modell för att identifiera anställda som inte kan se sig fortsätta arbeta på sin nuvarande arbetsplats. Vidare strävar studien efter att undersöka effekten mellan att inte arbeta kvar på sin nuvarande arbetsplats och frågor från medarbetarundersökning som handlar om trivsel, lika behandling, kompetensutveckling, ledarskap och våld. Data är från medarbetsundersökningar som har tillhandahållits av Webropol och efter omfattande hantering resulterade data i två separata datasets. Det första datasetet omfattade 4465 respondenter och var från sjukvårdssektorn, medan det andra datasetet omfattade 7665 respondenter och var från byggsektorn. Med denna informationen ville uppsatsen identifiera vilka frågor som tycks ha en inverkan på om en anställd inte kan se sig vara kvar på sin nuvarande arbetsplats. Därför har en grundmodell med logistisk regression använts för att tolka sambanden och mer avancerade metoder som Random Forest och XGBoosting används i klassificeringssyfte. De modeller som kunde särskilja klasserna mest distinkt var Random Forest och XGBoosting med träffsäkerheter på 79% respektive 75%. Dessa modeller hade högst AUC-värden och en balanserad specificitet och sensitivitet. Den fråga som hade störst betydelse för båda branscherna om en anställd inte kan se sig jobba kvar var frågan “Jag känner motivation i mitt arbete”, där en mindre motivation ökar risken för att en anställd inte kan se sig jobba kvar på sin arbetsplats.

Abstract

The purpose of the study is to create a model that can predict whether an employee cannot see themselves working at their current workplace. Furthermore, the study aims to investigate the relationship between leaving theirs current workplace and employee survey questions relating to wellbeing, equal treatment, skill development, leadership, and workplace violence. The data is sourced from employee surveys provided by Webropol, and after extensive data processing, it resulted in two separate datasets. The first dataset contained 4465 respondents from the healthcare sector, while the second dataset included 7665 respondents from the construction sector. With this information, the thesis aimed to identify which questions appear to have an impact on whether an employee cannot see themselves staying at their current workplace. Hence, a logistic regression model was estimated to interpret the associations, and more advanced methods such as Random Forest and XGBoosting were utilized for classification purposes. The models that exhibited the most distinct ability to differentiate the classes were Random Forest and XGBoosting, achieving accuracies of 79% and 75%, respectively. These models yielded the highest AUC values and a balanced specificity and sensitivity. The question that held the greatest significance for both industries regarding if an employee cannot see themselves working was "I feel motivated in my work," where lesser motivation increased the risk of an employee not being able to see themselves remaining at their workplace.

Förord

Först och främst vill vi rikta vårt tack till Webropol, det företag som har varit vår samarbetspartner och tillhandahållit den värdefulla data för vår analys av medarbetarundersökningar.

Vidare skulle vi vilja uttrycka vårt djupa tack till vår handledare, Annika Tillander. Hennes engagemang, expertis och ständiga stöd har varit värdefullt för oss genom hela uppsatsarbetet.

Slutligen vill vi uttrycka vår uppskattning till alla de anonyma respondenter som deltog i medarbetarundersökningarna och generöst bidrog med sin tid och sina svar.

Innehåll

Figurer

Tabeller

1 Introduktion

1.1 Bakgrund

Personalomsättning är en utmaning för företag på grund av dess negativa effekter, som innefattar höga kostnader för rekrytering och personalutveckling samt störningar i produktionsprocessen. För att minska personalomsättningen behöver företag förstå de faktorer som påverkar den, särskilt de som ökar den. Föregående forskning har funnit att det finns ett negativ samband mellan arbetstillfredsställelse och avsikt att säga upp sig. Det finns också potential i att använda medarbetarundersökningsdata för att prediktera om en anställd ligger i riskzonen för att säga upp sig (Gomomo, 2015).

En studie av Foley (2019) undersökte användningen av en Random Forest modell för att förutsäga uppsägning av anställda inom försvarsmakten. Resultaten visade att modellen hade en träffsäkerhet på 89% och en precision på 72%. Foley noterade dock att det begränsade datamaterialet (1500 observationer av totalt 20 000) var en bidragande faktor till modellens låga precision. Andra maskininlärningsmodeller kan vara användbara för att vidare undersöka denna problematik.

En annan relevant medarbetarundersökning av Xin (2022) undersökte sambandet mellan arbetsförhållandenas påfrestningar och de psykologiska aspekterna av de anställda. Studien genomfördes med japanska medborgare födda senast 1985 som var anställda inom olika akademiska sektorer. Resultaten visade ett signifikant negativt samband mellan psykologiskt kapital, som inkluderar variabler som hopp, optimism, uthållighet och självmedvetenhet, och benägenheten att lämna arbetsplatsen.

Genom att kombinera resultat från tidigare forskning och tillämpa statistiska modeller kan företag analysera och förstå de faktorer som påverkar personalomsättningen. Detta kan hjälpa företag att vidta åtgärder för att förbättra arbetsmiljön, öka arbetstillfredsställelsen och behålla sina anställda.

1.2 Syfte

Syftet med denna studie är att utveckla en prediktiv modell för att identifiera anställda som inte kan se sig fortsätta arbeta på sin nuvarande arbetsplats. Vidare strävar studien efter att undersöka effekten mellan att inte arbeta kvar på sin nuvarande arbetsplats och frågor från medarbetarundersökning som handlar om trivsel, lika behandling, kompetensutveckling, ledarskap och våld.

1.2.1 Frågeställningar

- Går det att göra en bra modell som predikterar om en anställd inte kan se sig jobba kvar på sin nuvarande arbetsplats?
- Vilka specifika frågor inom medarbetarundersökningar har störst effekt på att en anställd inte kan se sig arbeta kvar på sin nuvarande arbetsplats?

1.3 Etiska och samhälleliga aspekter

Undersökningar kan generera känslig information från respondenter, vilket ökar behovet av att utforma en etiskt ansvarsfull undersökningsprocess. När det gäller att undersöka samband mellan om en anställd kan se sig jobba kvar är det viktigt att ta hänsyn till de etiska samhälleliga frågor som kan uppstå.

För att säkerställa respondenternas integritet och skydda deras anonymitet är det viktigt att hantera data på ett säkert sätt där röjande variabler kan krypteras eller exkluderas (Karlsson et al., 2011).

När det gäller undersökningar som syftar till att undersöka samband mellan "om en individ kan se sig jobba kvar" och diverse andra frågor är det viktigt att vara medveten om de potentiella konsekvenserna av att identifiera respondenterna. Uppsägning kan vara en känslig fråga som kan leda till stigmatisering och diskriminering, särskilt om undersökningen syftar till att undersöka orsakerna bakom varför individen inte kan se sig jobba kvar.

2 Data

Datamaterialet består av 29 624 respondenter uppdelat på 75 företag i 17 olika branscher som kommer ifrån Webropols medarbetsundersökningar mellan åren 2016–2022. Inom varje företag har olika frågor ställts till medarbetarna, vilket har gjort att frågorna har behövts gått igenom manuellt för varje företag. Frågorna skilde sig mycket mellan bransch till bransch, men inom varje bransch har frågorna varit liknande, vilket har gjort att analyserna kommer göras branschvis.

I tabell 1 nedan visas ett exempel hur data ursprungligen kan se ut på ett företag med fiktiva svar.

Tabell 1: Fiktiv data

Ålder	Kön	Jag trivs på mitt företag/arbetsplats	Jag kan påverka min arbetssituation	...	...	Jag ser mig själv jobba kvar på X om 5 år
2	1	5	4	...	...	1
3	2	4	4	...	...	1
2	2	4	4	...	...	1
1	1	4	4	...	...	1
1	1	3	2	...	...	2

2.1 Avgränsningar

Rapporten har avgränsat sig till de två branscher som har flest svar på responsvariabeln, "Jag tror att jag kommer att arbeta kvar i företaget om X år", vilket är bygg och sjukvård. Det var alltså 15 branscher som ej undersöktes, organisationer, fastighet, utbildning, data & IT, företagstjänster, kultur & nöje, tillverkning, bank & finans, partihandel, hotell & restaurang, juridik, bemanning, motorfordon, transport och teknik.

I tabell 2 nedan visas antal företag på de två branscherna, bygg och sjukvård.

Tabell 2: Antal företag för branscherna

Bransch	Antal företag
Sjukvard	11
Bygg	7

Sjukvårdsbranschen har 11 företag och byggbranschen har 7 företag.

För att skapa ett datamaterial med få antal saknade värden, gjordes även vissa avgränsningar för företagen inom branscherna.

Varje företag hade mellan 30-120 frågor i sina medarbetarundersökningar. Dessa frågor granskades manuellt för att identifiera gemensamma frågor inom branschen. Ibland fanns det liknande frågor där det var svårt att avgöra om de skulle betraktas som samma och företagen använde också olika skalor i sina undersökningar. Därför togs företag med få antal observationer (<100) bort från analysen, eftersom det skulle kräva mycket arbete och tid utan att förväntas ha en betydande inverkan i analysen.

Det har också funnits företag som ej har haft med rapportens responsvariabel, "Jag tror att jag kommer att arbeta kvar i företaget om X år", vilket har gjort att de företagen har tagits bort.

Vissa företag hade också mycket få frågor jämfört med andra företag inom samma bransch. Dessa företag exkluderades för att prioritera ett större antal gemensamma frågor istället för fler observationer.

Nedan visas en lista över alla företag inom båda branscherna, inklusive antalet observationer per företag och vilka företag som har behållits (de som har grön bakgrund i tabell 3).

Tabell 3: Företag i bygg- och sjukvårdsbranschen tillsammans med deras observationer, företag med grön bakgrund har behållts

Företag	Antal observationer
Bygg, Företag 1	1737
Bygg, Företag 2[b]	122
Bygg, Företag 3	3437
Bygg, Företag 4	2491
Bygg, Företag 5[a]	23
Bygg, Företag 6[a]	22
Bygg, Företag 7[a]	22
Sjukvård, Företag 1[a]	74
Sjukvård, Företag 2[b]	2465
Sjukvård, Företag 3	501
Sjukvård, Företag 4	3836
Sjukvård, Företag 5	776
Sjukvård, Företag 6	749
Sjukvård, Företag 7[c]	323
Sjukvård, Företag 8[c]	317
Sjukvård, Företag 9[c]	296
Sjukvård, Företag 10[c]	315
Sjukvård, Företag 11	985

[a] Företaget har för få observationer [b] Företaget har ej med responsvariabeln [c] Företaget har för få frågor

I tabell 3 ovan syns det att fyra stycken företag har tagits bort i byggbranschen och att sex stycken företag har tagits bort i sjukvårdsbranschen. Anledningen till att företagen togs bort framkommer i fotnoterna under tabellen.

2.2 Gemensamma frågor inom branscherna

Nedan visas tabell 4 där varje gemensam fråga för företagen inom byggbranschen är utskriven tillsammans med en variabelbeteckning. Den andra kolumnen, "Kategori" är en benämning som skribenterna själv har gjort genom att gruppera de frågor som behandlar samma ämne. Den sista kolumnen "Företag" beskriver vilka företag som har haft med frågan i sin medarbetsundersökning.

Tabell 4: Varje gemensam fråga för företagen inom byggbranschen

Variabelnamn	Kategori	Frågans_Formulering	Företag
arb_mil_1	Arbetsmiljö	Jag har en bra fysisk arbetsmiljö	1,3,4
arb_mil_2	Arbetsmiljö	Min arbetsmiljö är trygg och säker	1,3,4
ålder	Ålder	Ålder	1,3,4
utveck_1	Kompetensutveckling	Jag får den kompetensutveckling jag behöver	1,3,4
utveck_2	Kompetensutveckling	Jag tar initiativ till att få den kompetensutveckling jag behöver	1,3,4
utveck_3	Kompetensutveckling	Jag tycker att det finns tillräckliga utvecklingsmöjligheter för mig inom X	1,3,4
utveck_4	Kompetensutveckling	Jag anser att min kompetens och erfarenhet tas tillvara på ett bra sätt	1,3,4
utveck_5	Kompetensutveckling	Jag känner att det satsas på mig rent utvecklingsmässigt inom X	1,3,4
utveck_6	Kompetensutveckling	Jag är nöjd med min utveckling i arbetet det senaste året	1,3,4
kräk_sär_1	Kränkande_särbehandling	På mitt arbete är det tydligt att kränkande särbehandling inte accepteras	1,3,4
kräk_sär_2	Kränkande_särbehandling	Jag tycker att det finns bra rutiner på X för att hantera kränkande särbehandling	1,3,4
kräk_sär_3	Kränkande_särbehandling	Jag vet vem jag ska vända mig till om kränkande särbehandling förekommer	1,3,4
kräk_sär_4	Kränkande_särbehandling	På min arbetsplats får den som är utsatt för kränkande särbehandling snabbt hjälp	1,3,4
ledar_1	Ledarskap	Jag får tydlig feedback på min arbetsinsats från min närmaste chef	1,3,4
kön	Kön	Kön	1,3,4
trivs_1	Trivsel	Jag trivs på mitt företag/min arbetsplats	1,3,4
trivs_2	Trivsel	Jag kan påverka min arbetssituation	1,3,4
trivs_3	Trivsel	Jag trivs med min yrkesroll/arbetsuppgifter	1,3,4
trivs_8	Trivsel	Jag känner motivation i mitt arbete	1,3,4
trivs_13	Trivsel	Samarbetet med mina kollegor på min arbetsplats fungerar bra	1,3,4
våld_1	Våld	Har du, under det senaste året, blivit utsatt för våld eller hot i ditt arbete?	1,3,4
våld_2	Våld	Tycker du att det förekommer hot eller våld på X?	1,3,4
jobb_kva_5	Jobba_kvar	Jag tror att jag kommer att arbeta på X om fem år	1,3,4

Tabell 5 och 6 nedan visar en sammanställning av gemensamma frågor för företag inom sjukvårdsbranschen, tillsammans med respektive variabelbeteckning. Observera att vissa frågor har kombinerats till en enda variabel, eftersom frågorna är så likartade att de kan betraktas som samma. Denna bedömning har gjorts av författarna till rapporten.

Tabell 5: Varje gemensam fråga för företagen inom sjukvårdsbranschen

Variabelnamn	Kategori	Frågans_Formulering	Företag
info_1	Information	Jag tar själv ansvar för att söka den information jag behöver i mitt arbete	3,4,5,6,11
info_2	Information	Jag har lätt att hitta den information jag behöver i mitt arbete	3,4,5,6,11
utveck_1	Kompetensutveckling	Jag ges tillräckliga möjligheter att utveckla min kompetens inom mitt arbetsområde	4,6,11
	Kompetensutveckling	Jag får den kompetensutveckling/utbildning som jag behöver i mitt arbete	3,5
utveck_2	Kompetensutveckling	Jag tar själv initiativ för att få den kompetensutveckling jag behöver	3,4,5,6,11
utveck_3	Kompetensutveckling	Det finns bra utvecklingsmöjligheter inom X för mig	3,4,5,6,11
utveck_4	Kompetensutveckling	Jag anser att min kunskap tas tillvara på ett bra sätt	4,5
	Kompetensutveckling	Jag anser att min kompetens och erfarenhet tas tillvara på ett bra sätt	3,6
	Kompetensutveckling	Jag känner att min kompetens tas tillvara på min arbetsplats	11
ledar_2	Ledarskap	Jag är som helhet nöjd med min närmaste chef	3,4,5,6,11
ledar_3	Ledarskap	Min närmaste chef tar snabbt tag i problem som rör konflikter och relationer på arbetsplatsen	3,4,5,6,11
ledar_4	Ledarskap	Min närmaste chef arbetar för att skapa ett öppet, inkluderande och tillåtande klimat	3,4,5,6,11
lika_beh_1	Likabehandling	På min arbetsplats behandlas alla lika oavsett kön	3,4,5,6,11
lika_beh_2	Likabehandling	På min arbetsplats behandlas alla lika oavsett ålder	3,4,5,6,11
lika_beh_3	Likabehandling	På min arbetsplats behandlas alla lika oavsett etnisk tillhörighet	3,4,5,6,11
lika_beh_4	Likabehandling	På min arbetsplats behandlas alla lika oavsett könsöverskridande identitet eller uttryck	3,4,5,6,11
lika_beh_5	Likabehandling	På min arbetsplats behandlas alla lika oavsett religion eller annan trosuppfattning	3,4,5,6,11
lika_beh_6	Likabehandling	På min arbetsplats behandlas alla lika oavsett funktionsnedsättning	3,4,5,6,11

Tabell 6: Varje gemensam fråga för företagen inom sjukvårdsbranschen (fortsättning)

Variabelnamn	Kategori	Frågans_Formulering	Företag
lika_beh_7	Likabehandling	På min arbetsplats behandlas alla lika oavsett sexuell läggning	3,4,5,6,11
trivs_2	Trivsel	Jag kan påverka min arbetssituation	3,4,5,6,11
trivs_4	Trivsel	Jag har en rimlig arbetsbelastning (varken för stor/inte för liten)	4,11
trivs_6	Trivsel	Jag har en rimlig arbetsbelastning	3,5,6
trivs_7	Trivsel	Det finns tid för återhämtning efter perioder av stress på jobbet	3,4,5,6,11
trivs_8	Trivsel	Jag är som helhet nöjd med min arbetssituation	3,4,5,6,11
trivs_9	Trivsel	Jag känner motivation i mitt arbete	3,4,5,6,11
trivs_10	Trivsel	Jag är som helhet nöjd med X som arbetsgivare	3,4,5,6,11
trivs_11	Trivsel	Min arbetssituation uppfyller de förväntningar som jag har på mitt arbete	3,4,5,6,11
trivs_12	Trivsel	Jag är stolt över att arbeta i X	3,4,5,6,11
våld_1	Våld	Jag rekommenderar gärna min arbetsgivare för andra	3,4,5,6,11
jobb_kva_2	Jobba_kvar	Har du, under det senaste året, blivit utsatt för våld eller hot i ditt arbete?	5,6
jobb_kva_3	Jobba_kvar	Jag tror att jag kommer att arbeta kvar i X om 2 år	3,4,11
		Tror du att du kommer att arbeta kvar inom X om 3 år?	

2.3 Skalor

Bland företagen har olika skalor använts på samma frågor, vilket kräver att detta hanteras korrekt. Det är viktigt att göra detta för att få tillförlitliga resultat och dra korrekta slutsatser från medarbetsundersökningarna.

2.3.1 Bygg

Nedan i tabell 7 visas vilka skalor som har använts på frågorna i medarbetsundersökningarna för byggbranschen. Kolumnen “Företag” beskriver vilka företag som har haft den skalan.

Tabell 7: Skalor på variabler för företag inom byggbranschen

Variabler	Skala	Företag
arb_mil_1-arb_mil_2, utveck_1-utveck_6, trivs_1-trivs_3, trivs_8, trivs_13, ledar_1	1-5 skala: 1=Instämmer inte alls, 5=Instämmer helt, 6=Vet ej	1,3,4
våld_1-våld_2, kräk_sär_1-kräk_sär_4	1-2 skala: 1=Ja, 2=Nej, 3=Vill ej uppge	1,3,4
ålder	1-6 skala: 1=25 år eller yngre, 2=26-35 år, 3=36-45 år, 4=46-55 år, 5=56-65 år, 6=Äldre än 65 år, 7=Vill ej uppge	1,4
	1-3 skala: 1=25 år eller yngre, 2=26-35 år, 3=36 år eller äldre, 4=Vill ej uppge	3
kön	1-2 skala: 1=Kvinna, 2=Man, 3=Annat, 4=Vill ej uppge	1,4
	1-2 skala: 1=Kvinna, 2=Man, 3=Vill ej uppge	3
jobb_kva_5	1-2 skala: 1=Ja, 2=Nej, 3=Vet ej	1,4
	1-5 skala: 1-5 skala: 1=Instämmer inte alls, 5=Instämmer helt, -1=Vet ej	3

Det skiljer sig bland företagen när det kommer till skalor, vilket gör att detta måste åtgärdas så det blir samma för varje företag. Först och främst kommer alla svar "vet ej" och "vill ej uppge" hanteras som saknade värden förutom variabeln **jobb_kva_5**. Detta beror på att "vet ej" och "vill ej uppge" hade en liten frekvens på alla variabler förutom variabeln **jobb_kva_5**.

Variabeln **ålder** kommer att ändras till gemensam 1-2 skala för de tre företagen. Detta gör att 1-6 skalan som företag 1 och 4 måste ändras till en 1-2 skala. 1 till 2 slås ihop till kategorin "<=35år" och 3 till 6 slås ihop till kategorin ">35år", vilket visas nedan.

$$\text{1: <=35år} \begin{cases} \text{1=25 år eller yngre} \\ \text{2=26-35 år} \end{cases} \qquad \text{2: >35år} \begin{cases} \text{3=36-45 år} \\ \text{4=46-55 år} \\ \text{5=56-65 år} \\ \text{6=Äldre än 65 år} \end{cases} \tag{1}$$

Företag 3 använder sig av en 1-3 skala, vilket gör att den kommer ändras till en 1-2 skala genom att 1 till 2 slås ihop till kategorin "<=35år". Detta visas nedan.

$$1\text{: <=35år} \begin{cases} 1\text{=25 år eller yngre} \\ 2\text{=26-35 år} \end{cases} \qquad 2\text{: >35år} \begin{cases} 3\text{=36 år eller äldre} \end{cases} \tag{2}$$

Inom variabeln **kön** fanns respondenter som valde alternativet "annat", men eftersom antalet personer som valde detta alternativ var en klar minoritet jämfört med de som valde "man" och "kvinna", fanns en risk för att deras identitet skulle kunna röjas om kategorin "annat" var kvar. För att undvika denna potentiella risk beslutades det att inte inkludera "annat" som ett separat alternativ.

Det sista variabeln som skiljer skalorna åt är responsvariabeln **jobb_kva_5**. En binär skala har använts, vilket har gjort att alla företags skalor har behövts ändras. Responsvariabeln har två klasser "nej" och "inte nej". Olika omkodningar har testats för att få två omkodningar som kan antas ha samma fördelning och de slutgiltiga omkodningarna syns nedan i tabell 8 och 9 tillsammans med ett χ^2-test i tabell 10. χ^2-testet är en statistisk metod som används för att analysera om det finns en signifikant skillnad mellan två fördelningar (Newbold et al., 2013).

Tabell 8: Omkodning på variabel **jobb_kva_5** för företag 3

Företag 3 gamla skala	Företag 3 ny skala	Betydelse
1	2	Nej
2 3 -1 4 5	1	Inte nej

Tabell 9: Omkodning på variabel **jobb_kva_5** för företag 1 och 4

Företag 1 och 4 gamla skala	Företag 1 och 4 ny skala	Betydelse
2	2	Nej
1 3	1	Inte nej

$$H_0 : \text{Det finns ingen skillnad i fördelningen av svaren mellan omkodningen i företag 3 och omkodningen i företag 1 och 4} \tag{3}$$

$$H_1 : \text{Det finns skillnad i fördelningen av svaren mellan omkodningen i företag 3 och omkodningen i företag 1 och 4} \tag{4}$$

Tabell 10: χ^2-test för skillnad i fördelningen omkodning på variabel **jobb_kva_5**

	Inte nej	Nej
Omkodningen i företag 3	3066	370
Omkodningen i företag 1 och 4	3719	509

[a] $\chi^2_{test} = 2.89$, p-value = 0.089

I χ^2-testet överstiger p-värdet signifikansnivån på 5%, vilket gör att det inte finns stöd för att fördelningarna har olika fördelningar, se tabell 10.

I tabell 11 nedan visas en sammanställning efter all omkodning av variablernas skalor inom byggbranschen.

Tabell 11: Skalor på variabler för företag inom byggbranschen efter omkodning

Variabler	Skala	Företag
arb_mil_1-arb_mil_2, utveck_1-utveck_6, trivs_1-trivs_3, trivs_8, trivs_13, ledar_1	1-5 skala: 1=Instämmer inte alls, 5=Instämmer helt	1,3,4
våld_1-våld_2, kräk_sär_1-kräk_sär_4	1-2 skala: 1=Ja, 2=Nej	1,3,4
ålder	1-2 skala: 1= <=35år, 2= <35 år	1,3,4
kön	1-2 skala: 1=Kvinna, 2=Man	1,3,4
jobb_kva_5	1-2 skala: 1=Inte nej, 2=Nej	1,3,4

2.3.2 Sjukvård

Nedan i tabell 12 visas vilka skalor som har använts på frågorna i medarbetsundersökningarna för sjukvårdsbranschen.

Tabell 12: Skalor på variabler för företag inom sjukvårdsbranschen

Variabler	Skala	Företag
info_1-info_2, utveck_1-utveck_4, ledar_2-ledar_4, lika_beh_1-lika_beh_7, trivs_2, trivs_4, trivs_6-trivs_12	1-5 skala: 1=Instämmer inte alls, 5=Instämmer helt, 6/-1/0=Vet ej	3,4,5,6,11
våld_1	1-2 skala: 1=Ja, 2=Nej, 3=Vill ej uppge	3,4,5,6,11
jobb_kva_2	1-5 skala: 1-5 skala: 1=Instämmer inte alls, 5=Instämmer helt, 6=Vet ej	5,6
jobb_kva_3	1-5 skala: 1-5 skala: 1=Instämmer inte alls, 5=Instämmer helt, 6/-1=Vet ej	3
jobb_kva_3	1-2 skala: 1=Ja, 2=Nej, 3=Vill ej uppge, 4=Vet ej	11,4

Liksom byggbranschen kommer "vet ej"- och "vill ej uppge"-svaren hanteras som saknade värden förutom inom variabeln **jobb_kva_2** och **jobb_kva_3**.

I sjukhusbranschen har två väldigt liknande frågor ställts, "Jag tror att jag kommer att arbeta kvar i X om 2 år" (**jobb_kva_2**) och "Tror du att du kommer att arbeta kvar inom X om 3 år?" (**jobb_kva_3**). Det ska undersökas om dessa frågor kan antas vara lika och ha samma fördelning för att i så fall slå ihop frågan. Detta görs genom ett χ^2-test. Dock behöver skalorna vara samma för alla företag och därför kommer en binär skala skapas för alla företag på samma vis i byggbranschen. Olika omkodningar har testats för att få två omkodningar som kan antas ha samma fördelning och de slutgiltiga omkodningarna syns nedan i tabell 13 och 14 tillsammans med ett χ^2-test i tabell 15 som testar om fördelningarna är lika eller inte.

Tabell 13: Omkodning på variabel **jobb_kva_2** / **jobb_kva_3** för företag 3,5 och 6

Företag 3,5 och 6 gamla skala	Företag 3,5 och 6 ny skala	Betydelse
1 2	2	Nej
3 -1 6 4 5	1	Inte nej

Tabell 14: Omkodning på variabel **jobb_kva_2** / **jobb_kva_3** för företag 4 och 11

Företag 4 och 11 gamla skala	Företag 4 och 11 ny skala	Betydelse
2	2	Nej
1 3 4	1	Inte nej

$$H_0 : \text{Det finns ingen skillnad i fördelningen av svaren mellan omkodningen i företag 3,5,6 och omkodningen i företag 4 och 11} \tag{5}$$

$$H_1 : \text{Det finns skillnad i fördelningen av svaren mellan omkodningen i företag 3,5,6 och omkodningen i företag 4 och 11} \tag{6}$$

Tabell 15: χ^2-test för skillnad i fördelningen på variabel **jobb_kva_2** / **jobb_kva_3**

	Inte nej	Nej
Omkodningen i företag 3,5 och 6	455	46
Omkodningen i företag 4 och 11	3550	414

[a] $\chi^2_{test} = 0.64$, p-value = 0.425

I χ^2-testet är p-värdet över 5%, vilket menas med att det inte finns något stöd för att fördelningarna har olika fördelningar, se tabell 15.

Nedan visas hypoteserna och χ^2-test om **jobb_kva_2** och **jobb_kva_3** har samma fördelning.

$$H_0 : \text{Det finns ingen skillnad i fördelningen av svaren mellan variablerna } \textbf{jobb\_kva\_2} \text{ och } \textbf{jobb\_kva\_2} \tag{7}$$

$$H_1 : \text{Det finns skillnad i fördelningen av svaren mellan variablerna } \textbf{jobb\_kva\_2} \text{ och } \textbf{jobb\_kva\_2} \tag{8}$$

Tabell 16: χ^2-test mellan **jobb_kva_2** och **jobb_kva_3**

	Inte nej	Nej
jobb_kva_3	2631	309
jobb_kva_2	1374	151

[a] $\chi^2_{test} = 0.34$, p-value = 0.560

I tabell 16 antas **jobb_kva_2** och **jobb_kva_3** ha samma fördelning, då det inte finns något stöd för att det är olika fördelningar (p-värde > 5%) och därför kommer dessa variabler slås ihop och bilda den nya variabeln **jobb_kva_23**.

I tabell 17 nedan visas en sammanställning efter all omkodning av variablernas skalor inom sjukvårdsbranschen.

Tabell 17: Skalor på variabler för företag inom sjukvårdsbranschen efter omkodning

Variabler	Skala	Företag
info_1-info_2, utveck_1-utveck_4, ledar_2-ledar_4, lika_beh_1-lika_beh_7, trivs_2, trivs_4, trivs_6-trivs_12	1-5 skala: 1=Instämmer inte alls, 5=Instämmer helt	3,4,5,6,11
våld_1	1-2 skala: 1=Ja, 2=Nej	3,4,5,6,11
jobb_kva_23	1-2 skala: 1=Inte nej, 2=Nej	3,4,5,6,11

2.4 Saknade värden

För att fullständiga datamaterialet är det nödvändigt att utforska och hantera eventuella saknade värden som återstår. Inom detta kapitel kommer variabler tas bort där det förekommer stor andel saknade värden. Det bör noteras att dessa variabler inte tas bort när XGBoost-modeller tränas, eftersom XGBoost kan hantera saknade värden på förklaringsvariablerna (Chen et al., 2016).

2.4.1 Bygg

Det första steget är att ta bort alla respondenter med saknat värde på responsvariabeln **jobb_kva_5**, då dessa respondenter inte är till någon nytta för analysdelen. Av 7 665 respondenter är det ynka 1 person som har saknat värde på **jobb_kva_5**, vilket ger totalt antal respondenter till 7 664.

Nedan visas tabell 18 med alla variabler på byggbranschen tillsammans med antal och andel saknade värden på variabeln.

Tabell 18: Antal saknade värden på varje variabel inom byggbranschen

Variabel	Antal saknade värden	Andel saknade värden
ålder	43	0.01
kön	76	0.01
trivs_1	7	0
trivs_2	50	0.01
trivs_8	8	0
trivs_13	19	0
utveck_1	130	0.02
utveck_2	96	0.01
utveck_3	246	0.03
utveck_4	89	0.01
utveck_5	207	0.03
utveck_6	126	0.02
ledar_1	159	0.02
arb_mil_1	2854	0.37
arb_mil_2	2849	0.37
våld_1	1670	0.22
våld_2	1685	0.22
kräk_sär_1	2263	0.3
kräk_sär_2	3429	0.45
kräk_sär_3	1998	0.26
kräk_sär_4	4567	0.6
jobb_kva_5	0	0

Det syns i tabell 18 att det är flera variabler som har stor andel saknade värden, vilket gör att variabler enligt kriteriet nedan tas bort helt.

- Mer än 20% saknade värden

arb_mil_1,**arb_mil_2**, **våld_1**, **våld_2** och **kräk_sär_1** - **kräk_sär_4** tas bort.

2.4.2 Sjukvård

Det första steget är att ta bort alla respondenter med saknat värde på **jobb_kva_23**, då dessa respondenter inte är till någon nytta för analysdelen. Av 5 827 respondenter är det 1 362 som har saknat värde på **jobb_kva_23**, vilket ger totalt antal respondenter till 4 465.

Nedan visas tabell 19 med alla variabler på sjukvårdsbranschen tillsammans med antal och andel saknade värden på variabeln.

Tabell 19: Antal saknade värden på varje variabel inom sjukvårdsbranschen

Variabel	Antal saknade värden	Andel saknade värden
trivs_2	67	0.02
trivs_4	37	0.01
trivs_6	63	0.01
trivs_7	31	0.01
trivs_8	16	0
trivs_9	65	0.01
trivs_10	81	0.02
trivs_11	114	0.03
trivs_12	143	0.03
utveck_1	151	0.03
utveck_2	128	0.03
utveck_3	624	0.14
utveck_4	72	0.02
ledar_2	224	0.05
ledar_3	778	0.17
ledar_4	328	0.07
lika_beh_1	542	0.12
lika_beh_2	405	0.09
lika_beh_3	660	0.15
lika_beh_4	971	0.22
lika_beh_5	738	0.17
lika_beh_6	795	0.18
lika_beh_7	880	0.2
info_1	28	0.01
info_2	43	0.01
våld_1	60	0.01
jobb_kva_23	0	0

Samma kriterier som användes i byggbranschen för att ta bort variabler kommer att användas här.

- Mer än 20% saknade värden

En variabel, **lika_beh_4**, har mer än 20% saknade värden, vilket gör att den kommer tas bort.

2.5 Beskrivande statistik

I detta kapitel kommer vi att undersöka fördelningen av utvalda frågor för att ge en övergripande bild av datamaterialet.

2.5.1 Bygg

Tabell 20: Fördelningen på variabeln **trivs_8** (Jag känner motivation i mitt arbete) för byggbranschen

Svar betydelse	Svar	Andel
Instämmer inte alls	1	0.02
...	2	0.05
...	3	0.16
...	4	0.38
Instämmer helt	5	0.4

Det är svarsalternativ 4 och 5 som har störst frekvens på frågan "Jag känner motivation i mitt arbete" för byggbranschen, se tabell 20.

Tabell 21: Könsfördelning inom byggbranschen

Kön	Andel
kvinna	0.25
man	0.75

Det är 75% män och 25% kvinnor i datamaterialet för byggbranschen, se tabell 21.

Tabell 22: Fördelning om man tror att man kommer jobba kvar efter 5 år (**jobb_kva_5**)

inte nej /nej	Andel
inte nej	0.89
nej	0.11

Det är 11% som har svarat "nej" på responsvariabeln "Jag tror att jag kommer att arbeta kvar i X om 5 år" och 89% som har svarat "inte nej", vilket inkluderar de som har svarat "ja", "vet ej", "vill ej uppge", se tabell 22.

2.5.2 Sjukvård

Tabell 23: Fördelningen på variabeln **trivs_8** (Jag känner motivation i mitt arbete) för sjukvårdsbranschen

Svar betydelse	Svar	Andel
Instämmer inte alls	1	0.02
...	2	0.06
...	3	0.18
...	4	0.38
Instämmer helt	5	0.37

Det är svarsalternativ 4 och 5 som har störst frekvens på frågan “Jag känner motivation i mitt arbete” för sjukvårdsbranschen, se tabell 23.

Tabell 24: Fördelningen på variabeln **trivs_9** (Jag är som helhet nöjd med X som arbetsgivare) för sjukvårdsbranschen

Svar betydelse	Svar	Andel
Instämmer inte alls	1	0.06
...	2	0.1
...	3	0.24
...	4	0.31
Instämmer helt	5	0.3

Inom frågan “Jag är som helhet nöjd med X som arbetsgivare” är det störst frekvens på svarsalternativ 4 och 5, följt av svarsalternativ 3, se tabell 24.

Tabell 25: Fördelning om man tror att man kommer jobba kvar efter 2/3 år (**jobb_kva_2/jobb_kva_3**)

inte nej /nej	Andel
inte nej	0.9
nej	0.1

Det är 10% som har svarat “nej” på responsvariabeln “Jag tror att jag kommer att arbeta kvar i X om 2/3 år” och 90% som har svarat “inte nej”, vilket inkluderar de som har svarat “ja”, “vet ej”, “vill ej uppge”, se tabell 25.

3 Metod

I detta metodkapitel beskrivs de olika metoderna som används i studien för att kunna besvara frågeställningen. De modeller som kommer tas upp är logistisk regression, klassificeringsträd, random forest och XGBoosting, tillsammans med utvärderingsmått för att bedöma modellernas prestanda.

Tabell 26: Metodbeteckningar

Beteckning	Betydelse
x_i	Observation i
y_i	Responsvariabel i
β_k	Lutningsparameter k
β_0	Intercept
x_{ik}	Observation i för k förklaringsvariabel
$X.k$	Förklaringsvariabel k
AUC	Area under ROC-kurva
$r(z)$	ROC-kurvans funktion
G	Gini-index
W	Antal klasser
p_{uw}	Sannolikheten att en observation i nod u tillhör klass w
$\hat{y}_b(x_i)$	Prediktion från modell b för observation i
q	Delmängd av de förklarande variablerna
p	Antal förklaringsvariabler
B	Antal stickprov i bagging
M	Antal iterationer i Gradient Boosting
TP	Sanna Positiva
FP	Falska Positiva
FN	Falska Negativa
TN	Sanna Negativa

I tabell 26 syns beteckningar tillsammans med deras betydelse som har använts i metodkapitlet.

3.1 Logistisk regression

Om responsvariabeln är binär kan väntevärdet modelleras med en logistisk regression. Den logistiska regressionen tar värden och transformerar de till värden mellan 0 och 1 (Tan et al., 2020).

$$P(y_i = 1|x_i) = \frac{1}{1 + exp[-\boldsymbol{X}\boldsymbol{\beta}]} \tag{9}$$

där

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & x_{i2} & \dots & x_{ik} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \tag{10}$$

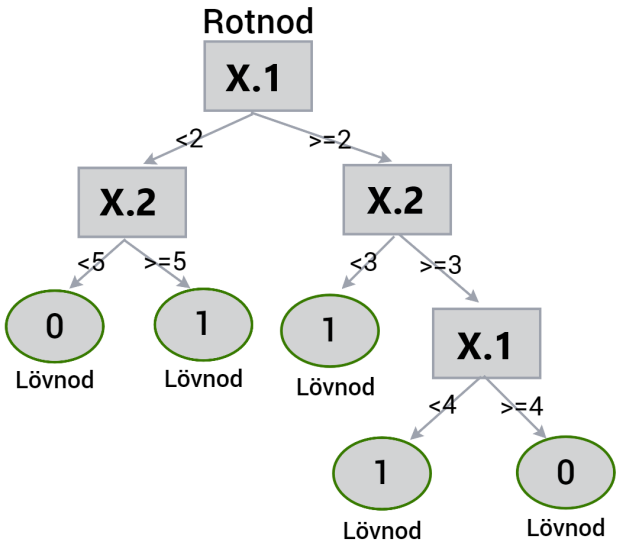
β_k är lutningsparametrar där β_0 är intercept.

x_{ik} där i står för rad och representerar en observation, samt k står för kolumn och representerar en förklaringsvariabel. Parametrarna β_k skattas genom maximum likelihood-metoden. $P(y_i = 1|x_i)$ tolkas som en sannolikhet för $y_i = 1$ och för att ange vilken klass y_i predikteras till används $P(y_i = 1|x_i) > 0.5$ som beslutgräns om klasserna är balanserade. Vid fallet av obalanserade klasser kan beslutgränsen variera för att få bättre resultat (Tan et al., 2020).

3.2 Klassificeringsträd

Klassificeringsträd är en icke-parametrisk metod som används för att prediktera klasser. Ett beslutsträd modelleras via rekursiv bearbetning, där trädet startar med en ensam rotnod som förknippas med all träningsdata. Förknippas en nod med mer än en klass skapas ett kriterium för att dela upp noden. Detta kriterium kontrollerar om det finns information att utvinna till modellen genom att dela upp en nod och när en uppdelning av en nod inte uppfyller kriteriet refereras denna slutliga nod som en lövnod (Tan et al., 2020).

Nedan visas en bild som visuellt beskriver hur klassificeringsträd fungerar.



Figur 1: En illustration av klassificeringsträd

I figur 1 ovan visas rotnoden och den representerar all tillgänglig data. Därefter skapas en förgrening genom att välja en specifik variabel, $X.1$ och skapa ett splittringkriterium baserat på denna variabel. Vilket resulterar i två nya noder som är förgrenade från rotnoden (Tan et al., 2020).

Genom att följa förgreningen i trädet kan man observera att till vänster om den nya noden skapas två lövnoder baserat på variabeln $X.2$. Om den högra noden istället följs, skapas en lövnod till vänster som möjliggör prediktion av klassen, följt av en ny nod till höger. Denna nya nod kan sedan delas upp ytterligare, vilket resulterar i skapandet av två nya lövnoder där klasserna kan predikteras (Tan et al., 2020).

Överanpassning är ett fenomen som förekommer ofta vid anpassningen av trädmodeller, överanpassning är när modellen klassificerar tränings data perfekt vilket innebär att modellen har fångat upp brus och inte skattar det "generella" i data-materialet. Detta resulterar i att valideringsdata inte klassificeras så väl. Det finns två generella metoder för att hantera detta dessa är förbeskärning och efterbeskärning. Förbeskäring är när trädet byggs efter angivna begränsningar som exempelvis max djup och minst antal observationer i varje löv. Efterbeskärning är när trädet beskärs efter det fulla trädet har skapats (Tan et al., 2020).

3.2.1 Gini

Gini-indexet är ett vanligt mått för att mäta total variation inom W antal klasser. Det används ofta som ett kriterium för att göra uppdelningar i trädmodeller. Ett högt Gini-värde indikerar att en nod innehåller information från flera klasser, vilket tyder på att noden behöver delas upp (Tan et al., 2020).

I formeln nedan representerar p_{uw} sannolikheten för att en observation i nod u tillhör klass w. Gini-indexet ger information om hur mycket orenhet det finns i en nod, där 0 betyder perfekt klassificering och 1 betyder maximal orenhet (Tan et al., 2020).

Gini-indexet för en nod kan beräknas enligt följande formel:

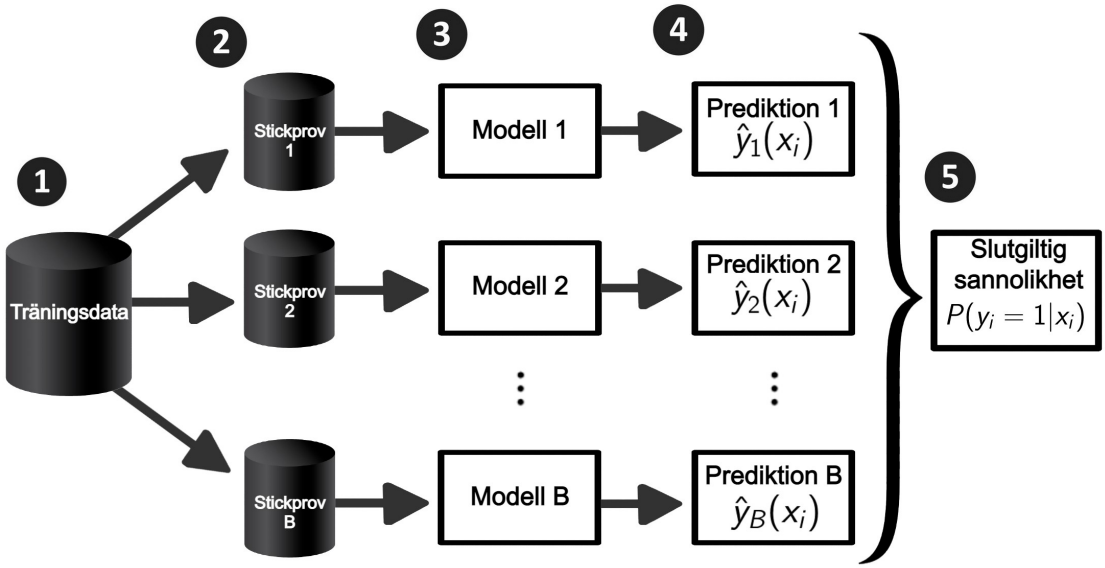
$$G = \sum_{w=1}^{W} p_{uw}(1 - p_{uw}) \tag{11}$$

3.3 Bagging

Bagging är en ensemblemetod som används för att minska variansen hos en modell genom att skapa flera modeller som tränas på olika delmängder av datamängden. Metoden bygger på bootstrapping, där B olika stickprov med återläggning skapas från testdata. Efter det tränas en modell på varje stickprov och slutligen kombineras alla modeller genom att ta medelvärdet av prediktionerna $\hat{y}_b(x_i)$ från varje modell. Den slutgiltiga sannolikheten $P(y_i = 1|x_i)$ räknas ut på följande sätt (Tan et al., 2020).

$$P(y_i = 1|x_i) = \frac{1}{B} * \sum_{b=1}^{B} \mathbb{I}(\hat{y}_b(x_i) = 1) \tag{12}$$

Där $P(y_i = 1|x_i)$ är den slutgiltiga sannolikheten från bagging-modellen för observationen i, $\hat{y}_b(x_i)$ är prediktionen från modell b för observationen i. $\mathbb{I}$ är en indikatorfunktion som är lika med 1 om uttrycket inom parentes är sant och 0 annars. B är antal stickprov/modeller. Nedan visas en bild som visuellt beskriver hur bagging fungerar, se figur 2.



Figur 2: En illustration av bagging

Det finns dock en nackdel med att bara använda bagging, vilket är att de B stickproven kan vara korrelerade. Detta kan påverka modellens noggrannhet negativt.

3.4 Random Forest

För att undvika korrelationen mellan de olika stickproven i bagging används Random Forest metoden som bygger på att slumpmässigt ändra modellerna. I Random Forest används träd som modeller och för varje träd används endast en delmängd q av de förklarande variablerna, där q oftast väljs till $\sqrt{p}$, ($p =$ antal förklaringsvariabler) för klassificering (Tan et al., 2020).

Detta minskar kostnaden vid varje uppdelning och förhindrar att träden blir starkt korrelerade med varandra. Detta ökar i sin tur modellens prestanda och gör det möjligt att hantera en större mängd förklaringsvariabler utan att drabbas av överanpassning. Sedan skattas prediktionen genom bagging-metoden som togs upp innan (Tan et al., 2020).

3.5 Boosting

Boosting är en ensemble-metod där flera svaga prediktorer kombineras för att bilda en stark prediktor. Boosting-algoritmen skiljer sig från bagging genom att fokusera på de missklassificerade observationerna och justerar

därmed vikterna på dessa observationer i varje iteration. Istället för att bygga flera oberoende modeller som i bagging, bygger man en serie av modeller som bygger på varandra (Zhang et al., 2022).

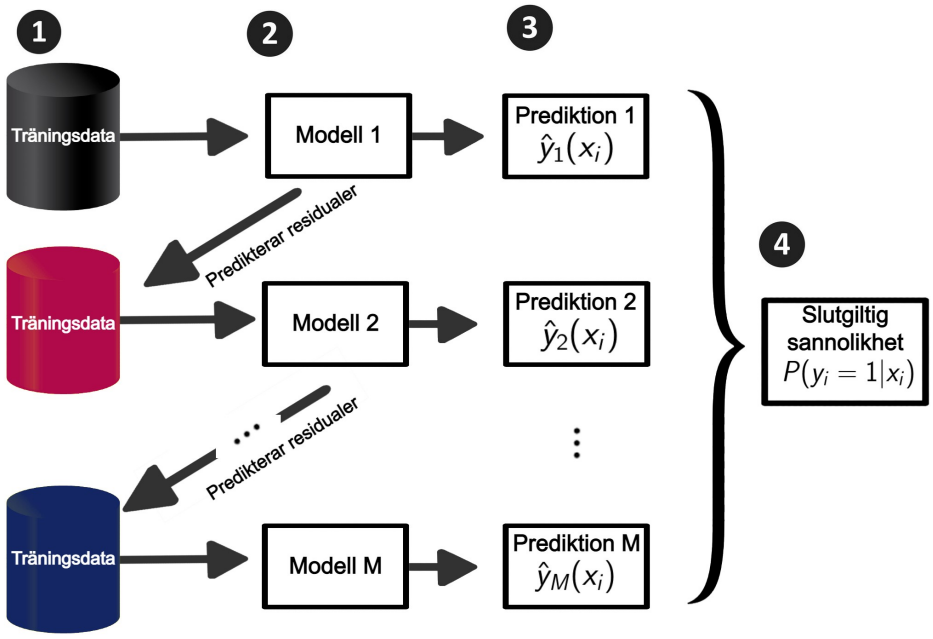
I en boosting-algoritm tränas en sekvens av modeller i serie, där varje modell fokuserar på att korrigera de misstag som gjorts av föregående modell. Varje modell tränas på hela datamängden och kombinerar flera svaga modeller till en stark modell genom att fokusera på de exempel som tidigare var svåra att klassificera (Zhang et al., 2022).

Det finns många olika varianter av boosting-algoritmer, AdaBoost, Gradienet Boosting, XGBoosting (Extreme Gradient Boosting) med mera. Algoritmerna skiljer sig när det kommer till att korrigera de misstag som gjorts av föregående modell, till exempel använder Gradient Boosting förlustfunktioner, medan AdaBoost använder sig av vikter. Denna uppsats kommer fokusera på XGBoosting, men för att förstå hur XGBoosting fungerar måste Gradient Boosting förklaras först (Zhang et al., 2022).

3.6 Gradient Boosting

Gradient boosting arbetar genom att först träna en grundmodell som förutsäger de initiala klasserna på data. Därefter förutsäger varje efterföljande modell residualerna som har skapats av föregående modeller, istället för att prediktera de faktiska klasserna. Efter att residualerna har predikterats används de för att träna nästa modell i serien. Modellen försöker prediktera residualerna från föregående modell, så att skillnaden mellan de faktiska och förutsagda klasserna minskar. Detta görs genom att hitta den riktning i residualrummet som minskar förlustfunktionen mest. Gradienten av förlustfunktionen används för att justera parametrarna i modellen för att fokusera på de kvarvarande residualerna och minska förlusten så mycket som möjligt. Inom binär klassificering används ofta log loss som förlustfunktion (Zhang et al., 2022).

Nedan visas en bild som visuellt beskriver hur Gradient Boosting fungerar, se figur 3.



Figur 3: En illustration av boosting

3.7 XGBoosting

En variant av gradient boosting är XGBoosting. XGBoosting är en optimerad version av gradient boosting-algoritmen som använder flera tekniker för att förbättra prestandan och undvika överanpassning. En av dessa tekniker är att använda en regleringsparameter som kallas för "shrinkage" eller "learning rate", vilket kontrollerar hur mycket varje efterföljande modell bidrar till den slutgiltiga modellen. Genom att sätta en lägre shrinkage kan man undvika överanpassning och skapa en mer generaliserad modell (Chen et al., 2016).

En annan teknik som används i XGBoosting är något som kallas för **tree pruning**. Detta innebär att algoritmen trimmar de träd som inte bidrar till att förbättra prestandan, vilket kan minska överanpassning (Chen et al., 2016).

XGBoosting använder också en speciell funktion som kallas för "weighted quantile sketch", som kan snabba upp beräkningarna av förlustfunktionen och gradienten. Detta gör XGBoost extremt effektivt för att hantera stora datamängder (Chen et al. 2016).

En annan viktig funktion inom XGBoosting är möjligheten att hantera saknade värden. Detta görs genom att man lägger saknade värden i ett eget trädgren, vilket undviker att saknade värden påverkar prediktionen på fel sätt (Chen et al., 2016).

I XGBoost finns det väldigt många hyperparameter, men vissa är mer centrala och inflytelserika. Nedan beskrivs fem centrala hyperparametrar (XGBoost, 2022)

- max_depth: Maximalt antal nivåer i trädstrukturen. Högre värden tillåter mer komplexitet, men kan öka risken för överanpassning.

- nrounds: Antalet iterationer av gradient boosting, vilket motsvarar antalet träd som byggs under träningen av XGBoost-modellen. Högre värden kan göra modellen mer kraftfull, men ökar också beräkningstiden och risken för överanpassning.
- η: Kontrollerar inlärningshastigheten för modellen. Låga värden ger en mer konservativ modell som motverkar överanpassning, medan höga värden kan ge snabbare konvergens men ökad risk för överanpassning.
- γ: Den minsta vinsten som krävs för att fortsätta dela en nod. Högre värden på γ kan bidra till att undvika överanpassning.
- scale_pos_weight: Förhållandet mellan antalet negativa och positiva exempel i träningsdatan. Används för att hantera obalans mellan klasserna.

Mer om hur XGBoosting fungerar finns att läsa på i skriften *XGBoost: A Scalable Tree Boosting System* av Chen och Guestrin (2016).

3.8 Utvärderingsmått

3.8.1 Förväxlingsmatris

Förväxlingsmatrisen kan användas för att mäta hur bra en modell har klassificerat sina klasser. Matrisen visas nedan i tabell 27 och de gula cellerna är antal observationer som modellen har klassificerat rätt (Tan et al., 2020).

Tabell 27: Förväxlingsmatris

		Predikterad klass	
		Klass = 1	Klass = 0
Sann klass	Klass = 1	TP = Sanna Positiva	FN = Falska Negativa
	Klass = 0	FP = Falska Positiva	TN = Sanna Negativa

Sedan kan träffsäkerhet räknas ut som är ett mått på hur stor andel av alla observationer som har klassificerats rätt. Detta mått kan dock var missvisande i fall där klasserna är obalanserade.

$$\text{Träffsäkerhet} = \frac{TP + TN}{TP + FN + FP + TN} \tag{13}$$

Sensitivitet och specificitet är två mått som är klasspecifika. Formlerna och beskrivningarna kommer utgå från klass = 1, men principen är densamma för klass = 0.

Sensitivitet beräknas som antal korrekta klass = 1 prediktioner dividerat på totala antalet som tillhör klass = 1 (Tan et al., 2020).

$$\text{Sensitivitet} = \frac{TP}{TP + FN} \tag{14}$$

Specificitet beräknas som antal korrekta klass = 0 prediktioner dividerat på totala antalet som tillhör klass = 0 (Tan et al., 2020).

$$\text{Specificitet} = \frac{TN}{TN + FP} \tag{15}$$

3.8.2 AUC

AUC (Area Under the Curve) är ett viktigt mått som används för att bedöma prestandan hos en binär klassificeringsmodell. Istället för att fokusera på ROC-kurvan, som visar förhållandet mellan True Positive Rate (TPR) och False Positive Rate (FPR), mäter AUC den totala prestandan hos modellen genom att beräkna arean under ROC-kurvan (Carrington at el., 2023).

En ROC-kurva illustrerar hur väl modellen kan särskilja mellan positiva och negativa exempel genom att variera den beslutsgräns som används för att göra klassificeringen. Genom att avläsa ROC-kurvan kan man identifiera de beslutsgränser som ger bäst prestanda för modellen (Carrington at el., 2023).

AUC ger ett mått på den balanserade sensitiviteten och specificiteten hos modellen. Genom att mäta arean under ROC-kurvan kan man bedöma hur väl modellen kan särskilja mellan positiva och negativa exempel. För att hitta den optimala beslutsgränsen kan man introducera nya beslutsgränser och söka efter det maximala värdet av AUC. En intressant användning av AUC är när man arbetar med datamaterial som har obalanserade klasser. Genom att maximera AUC kan man förbättra träffsäkerheten för den underrepresenterade klassen (Carrington at el., 2023).

Formeln för att beräkna AUC är:

$$AUC = \int_0^1 r(z)\, dz \tag{16}$$

Där AUC står för arean och $r(z)$ är ROC-kurvans funktion.

3.9 Korsvalidering

Korsvalidering för hyperparametrar i trädmodeller är en viktig metod för att optimera modellens prestanda. En vanlig teknik för att utforska olika kombinationer av hyperparametrar är grid search.

Vid grid search definieras ett förutbestämt antal värden för varje hyperparameter. Därefter testas systematiskt alla möjliga kombinationer av dessa värden. För varje kombination av hyperparametrar tränas trädmodellen och dess prestanda utvärderas med hjälp av utvärderingsmått som AUC, träffsäkerhet, sensitivitet och specificitet. Genom att systematiskt testa alla kombinationer kan den optimala uppsättningen av hyperparametrar identifieras för att maximera modellens prestanda (Géron, 2019).

3.10 R-paket

Rpart (Therneau & Atkinson, 2022) är ett paket som skattar rekursiva partitionerings- och regressionsträd, där används funktionen som heter rpart för att skatta klassificeringsträd.

Paketet *randomForest* (Breiman & Cutler, 2022) skapar klassificerings- och regressionträd baserat på slumpade skogar(forest) av träd. Här användes funktionen randomForest för att skatta en genomsnittlig representation av de slumpande träden.

xgboost (Chen et al., 2023) Exreme Gradient Boosting är en effektivare implementering av gradient boosting, eftersom xgboosting skattar flera modeller samtidigt. Funktionen XGBoost används för att skatta klassificering- samt regressionsträd.

4 Resultat

I resultatdelen redovisas resultaten från fyra modeller för både bygg- och sjukhusbranschen. Varje modell har delat upp data i tränings- och valideringsdata, 70% träning och 30% validering.

4.1 Logistisk regression

4.1.1 Bygg

Nedan i tabell 28 visas de förklarande variablerna tillsammans med interceptet för en logistisk regression för byggbranschen. För att unvidka multikollinearitet har maximalt en variabel ur varje frågekategori (trvisel, kompetensutveckling) använts i modellen. Det har sedan testats olika kombinationer av modeller för att identifiera de frågor som hade störst påverkan genom att jämföra modellers enskilda p-värden och AUC.

Tabell 28: Förklarande variabler för byggbranschen, logistisk regression

Fråga	Koefficient	Oddskvot	P-värde
(Intercept)	2.66		
ålder>35 år	-0.53	0.59	0.00
Jag känner motivation i mitt arbete	-0.73	0.48	0.00
Det finns bra utvecklingsmöjligheter inom X för mig	-0.49	0.61	0.00

De tre variablerna, "ålder > 35" som är dummykodad, "Jag känner mig motiverad i mitt arbete" och "Det finns bra utvecklingsmöjligheter inom X för mig" är tydligt signifikanta, då p-värdet understiger 5% signifikansnivå, se tabell 28. Oddskvoten visar att högre ålder, större motivation på arbetsplatsen och fler utvecklingsmöjligheter minskar risken för att en individ ska se sig sluta, förutsatt att de andra variablerna hålls konstanta.

Tabell 29: Förväxlingsmatris för byggbranschen, logistisk regression

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	1581	378	3637	941
	nej	70	176	197	398

Tabell 29 visar förväxlingsmatrisen för validerings- respektive träningsdata, modellen har svårt att korrekt identifiera de anställda som inte kan se sig jobba kvar.

Tabell 30: Träffsäkerhet för byggbranschen, logistisk regression

Valideringsdata	Träningsdata
0.797	0.78

Träffsäkerheten för validerings- och träningsdata uppgår till 80% respektive 78%, se tabell 30.

Tabell 31: Sensitivitet, specificitet och AUC för byggbranschen, logistisk regression

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.807	0.715	0.761	0.794	0.669	0.732

Tabell 31 visar sensitivitet, specificitet och AUC för validerings- respektive träningsdata. Det utläses från valideringsdata att sensitiveten är 80% och specificiteten 72%, vilket innebär att modellen har svårare att klassificera de anställda som inte ser sig jobba kvar efter 5 år. AUC-värdet för valideringsdata är 76%.

4.1.2 Sjukvård

Nedan i tabell 32 visas de förklarande variablerna tillsammans med interceptet för en logistisk regression för sjukvårdsbranschen. För att unvidka multikollinearitet har maximalt en variabel ur varje frågekategori (trvisel, kompetensutveckling, ledarskap, lika behandling, information) använts i modellen. Det har sedan testats olika kombinationer av modeller för att identifiera de frågor som hade störst påverkan genom att jämföra modellers enskilda p-värden och AUC.

Tabell 32: Förklarande variabler för sjukvårdbranshen, logistisk regression

Fråga	Koefficient	Oddskvot	P-värde
(Intercept)	0.66		
Jag känner motivation i mitt arbete	-0.60	0.55	0.000
Jag är som helhet nöjd med X som arbetsgivare	-0.59	0.55	0.000
Jag tar själv ansvar för att söka den information jag behöver i mitt arbete	0.30	1.35	0.001

De tre variablerna som ingår i modellen är: “Jag känner motivation i mitt arbete”, “Jag är som helhet nöjd med X som arbetsgivare” och “Jag tar själv ansvar för att söka den information jag behöver i mitt arbete”, se tabell 32. Samtliga koefficienter är signifikanta vid en signifikansnivå på 5%, vilket innebär att oddskvoterna kan tolkas. De två första variablerna har en oddskvot under 1, vilket innebär att ökning i skalan på frågorna minskar risken för en anställd att se sig sluta. Frågan “Jag tar själv ansvar för att söka den information jag behöver i mitt arbete” har istället en ökande risk för att en anställd ser sig sluta, vilket indikerar att mer självinitiativ i att skaffa information ökar risken för att en anställd ska se sig sluta.

Tabell 33: Förväxlingsmatris för sjukvårdsbranschen, logistisk regression

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	857	323	2012	732
	nej	31	99	89	229

Tabell 33 visar förväxlingsmatrisen för validerings- respektive träningsdata.

Tabell 34: Träffsäkerhet för sjukvårdsbranschen, logistisk regression

Valideringsdata	Träningsdata
0.73	0.732

I Tabell 34 presenteras träffsäkerheten för både valideringsdata och träningsdata. Både för valideringsdata och träningsdata är träffsäkerheten ungefär 73%.

Tabell 35: Sensitivitet, specificitet och AUC för sjukvårdsbranschen, logistisk regression

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.726	0.762	0.744	0.733	0.72	0.727

Inom valideringsdata är sensitiviteten 73% och specificiteten 76%, vilket innebär att modellen har enklare att klassificera de anställda som inte ser sig jobba kvar efter 2/3 år, se tabell 35. Detta skiljer sig från byggbranschens logistiska regression, där modellen hade svårare att klassificera de anställda som inte ser sig jobba kvar efter 5 år.

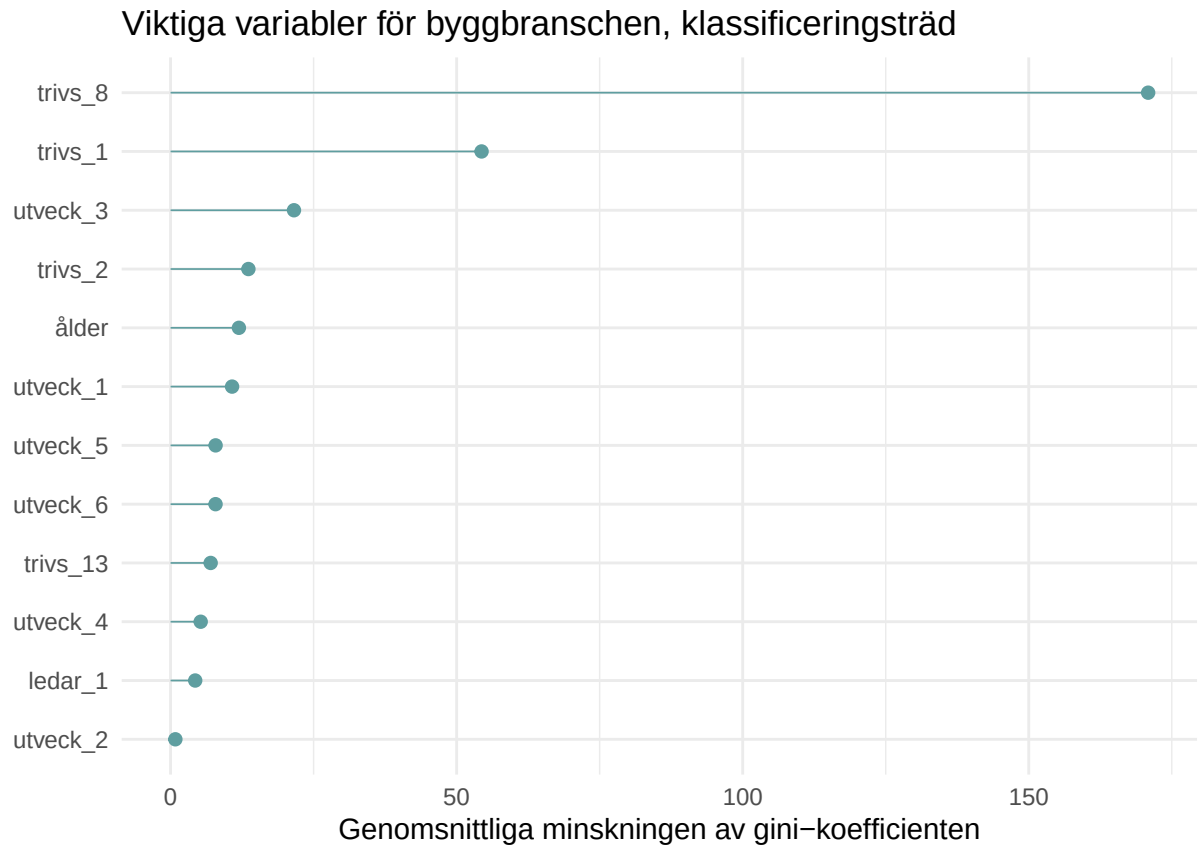
4.2 Klassificeringsträd

4.2.1 Bygg

Ett klassificeringsträd för byggbranschen har skattats genom att använda förberskärning där två begräsningar har använts, maxdjup och minsta antal observationer inom ett löv. Korsvalidering har utförts för att hitta de optimala begränsningarna som ger hög AUC och undviker överanpassning. Efter korsvalideringen har följande begränsningar valts:

- Maxdjup: 6
- Minst antal observationer inom ett löv: 50

Trädet blir för stort att visualisera och därför kommer ett diagram visas nedan på vilka variabler som är viktigast för att klassificera. Detta har räknats ut genom att använda den genomsnittliga minskningen av gini-koefficienten. Ju högre värde på genomsnittlig minskning av gini, desto högre betydelse har variabeln i modellen.



Figur 4: Viktiga variabler för byggbranschen, klassificeringsträd

Variabeln **trivs_8** har överlägset högst genomsnittlig minskning av gini, följt av **trivs_1**, se figur 4. Motivation i arbetet har således störst betydelse om en anställd inte kan se sig jobba kvar efter 5 år inom byggbranschen, följt av trivsel på arbetsplatsen, se figur 4.

Tabell 36: Variabelnamn med tillhörande fråga för byggbranschen, klassificeringsträd

Variabelnamn	Fråga
trivs_8	Jag känner motivation i mitt arbete
trivs_1	Jag trivs på mitt företag
utveck_3	Det finns bra utvecklingsmöjligheter inom X för mig
trivs_2	Jag kan påverka min arbetssituation
ålder	ålder
utveck_1	Jag får den kompetensutveckling jag behöver
utveck_5	Jag känner att det satsas på mig rent utvecklingsmässigt inom X
utveck_6	Jag är nöjd med min utveckling i arbetet det senaste året
trivs_13	Samarbetet med mina kollegor på min arbetsplats fungerar bra
utveck_4	Jag känner att min kompetens tas tillvara på min arbetsplats
ledar_1	Jag får tydlig feedback på min arbetsinsats från min närmaste chef

Tabell 36: Variabelnamn med tillhörande fråga för byggbranschen, klassificeringsträd *(continued)*

Variabelnamn	Fråga
utveck_2	Jag tar själv initiativ för att få den utveckling jag behöver

Tabell 36 innehar alla variabelbeteckningar med dess respektive fråga från enkätundersökningarna som används i trädet.

Tabell 37: Förväxlingsmatris för byggbranschen, klassificeringsträd

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	1694	350	3911	830
	nej	89	167	240	383

Tabell 37 visar förväxlingsmatrisen för både validerings- och träningsdata.

Tabell 38: Träffsäkerhet för byggbranschen, klassificeringsträd

Valideringsdata	Träningsdata
0.809	0.801

Träffsäkerheten för validerings- samt träningsdata ligger på 81% respektive 80%, se tabell 38.

Tabell 39: Sensitivitet, specificitet och AUC för byggbranschen, klassificeringsträd

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.829	0.652	0.741	0.825	0.615	0.72

För valideringsdata är sensitiveten 83%, specificiteten 65% och AUC-värdet 74%, vilket innebär att modellen har svårare att klassificera de anställda som inte ser sig jobba kvar efter 5 år, se tabell 39.

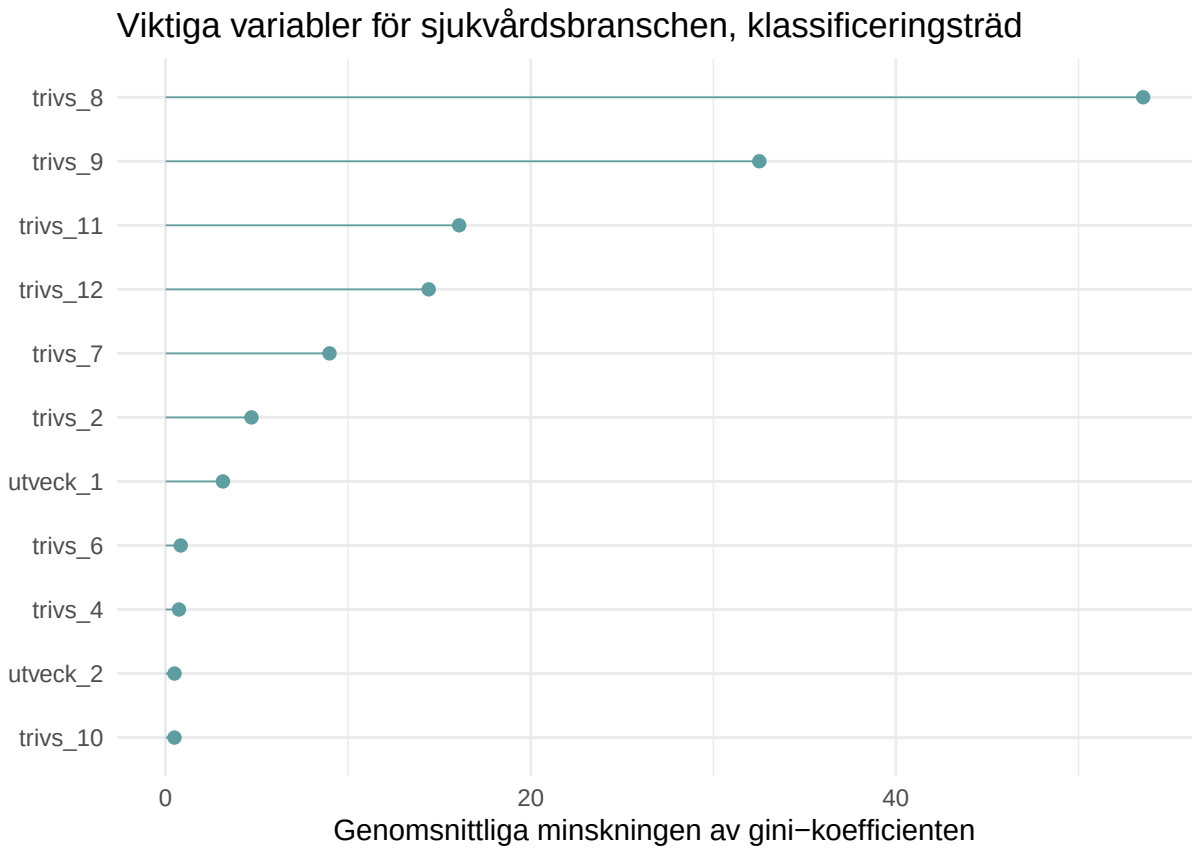
4.2.2 Sjukvård

Ett klassificeringsträd för sjukvårdsbranschen har skattats genom att använda förberskärning där två begränsningar har använts, maxdjup och minsta antal observationer inom ett löv. Korsvalidering har utförts för att hitta de optimala begränsningarna som ger hög AUC och undviker överanpassning. Efter korsvalideringen har följande begränsningar valts:

- Maxdjup: 4

- Minst antal observationer inom ett löv: 25

Nedan visas figur 5 som är ett diagram på vilka variabler som är viktigast för att klassificera för trädet. Detta har räknats ut genom att använda den genomsnittliga minskningen av gini-koefficienten. Ju högre värde på genomsnittlig minskning av gini, desto högre betydelse har variabeln i modellen.



Figur 5: Viktiga variabler för sjukvårdsbranschen, klassificeringsträd

Variabeln **trivs_8** har högst genomsnittlig minskning av gini, följt av **trivs_9**, se figur 5. Motivation i arbetet har därmed störst betydelse om en anställd inte kan se sig jobba kvar efter 2/3 år inom sjukvårdsbranschen, följt av nöjdhet av sin arbetsgivare.

Tabell 40: Variabelnamn med tillhörande fråga för sjukvårdsbranschen, klassificeringsträd

Variabelnamn	Fråga
trivs_8	Jag känner motivation i mitt arbete
trivs_9	Jag är som helhet nöjd med X som arbetsgivare
trivs_11	Jag är stolt över att arbeta i X
trivs_12	Jag rekommenderar gärna min arbetsgivare för andra
trivs_7	Jag är som helhet nöjd med min arbetssituation

Tabell 40: Variabelnamn med tillhörande fråga för sjukvårdsbranschen, klassificeringsträd *(continued)*

Variabelnamn	Fråga
trivs_2	Jag kan påverka min arbetssituation
utveck_1	Jag får den kompetensutveckling jag behöver
trivs_6	Det finns tid för återhämtning efter perioder av stress på jobbet
trivs_4	Jag har en rimlig arbetsbelastning
utveck_2	Jag tar själv initiativ för att få den utveckling jag behöver
trivs_10	Min arbetssituation uppfyller de förväntningar som jag har på mitt arbete

Tabell 40 innehar alla variabelbeteckningar med dess respektive fråga från enkätundersökningarna som används i trädet.

Tabell 41: Förväxlingsmatris för sjukvårdsbranschen, klassificeringsträd

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	1033	173	2403	396
	nej	60	74	137	189

Tabell 41 visar förväxlingsmatrisen för valideringsdata respektive träningsdata.

Tabell 42: Träffsäkerhet för sjukvårdsbranschen, klassificeringsträd

Valideringsdata	Träningsdata
0.826	0.829

Träffsäkerheten för valideringsdata samt träningsdata är 83%, vilket innebär att modellen inte anses överanpassad, se tabell 42.

Tabell 43: Sensitivitet, specificitet och AUC för sjukvårdsbranschen, klassificeringsträd

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.857	0.552	0.704	0.859	0.58	0.719

För valideringsdata är sensitiviteten 86%, specificiteten 55% och AUC-värdet 70%, vilket innebär att modellen har mycket svårare att klassificera de anställda som inte ser sig jobba kvar efter 2/3 år, se tabell 43.

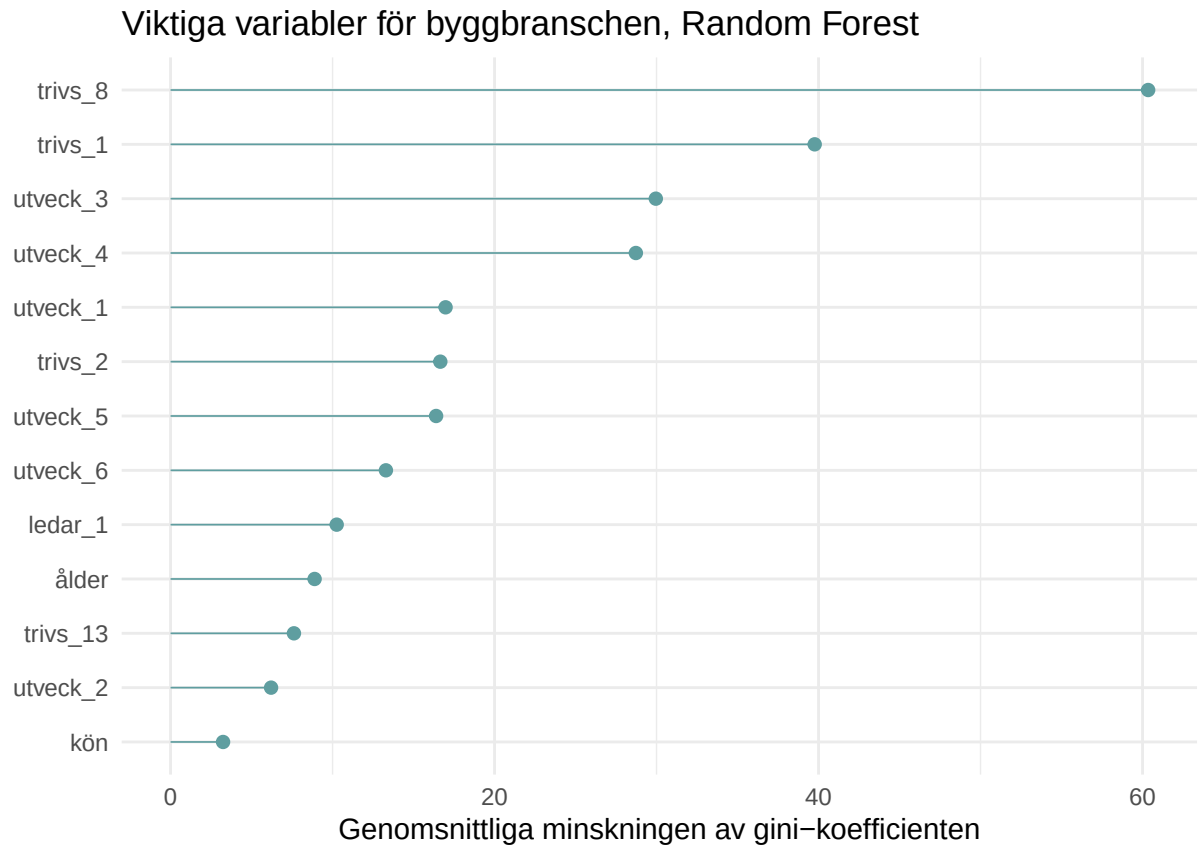
4.3 Random Forest

4.3.1 Bygg

En Random Forest modell för byggbranschen har skattats genom fyra hyperparametrar, antal trä att gro, antalet variabler som slumpmässigt väljs vid varje split i trädet, antalet observationer som krävs för att ett löv i trädet ska kunna skapas och maximala antalet noder som tillåts i trädet. Antalet variabler som slumpmässigt väljs vid varje split i trädet har valts till $\lfloor\sqrt{p}\rfloor = \lfloor\sqrt{13}\rfloor = 3$ där p är antal förklarande variabler. För de tre andra hyperparametrarna har korsvalidering utförts för att hitta de optimala begränsningarna som ger hög AUC och undviker överanpassning. Efter korsvalideringen har följande begränsningar valts:

- Antal träd att gro: 450
- Antalet variabler som slumpmässigt väljs vid varje split i trädet: 3
- Antalet observationer som krävs för att ett löv i trädet ska kunna skapas: 30
- Maximala antalet noder som tillåts i trädet: 50

Nedanför återfinns ett diagram som visar de variabler som har störst betydelse för klassificeringen av träden. Ju högre värde på den genomsnittliga minskningen av Gini-koefficienten, desto större betydelse har variabeln i modellen.



Figur 6: Viktiga variabler för byggbranschen, Random Forest

Variabeln **trivs_8** har högst genomsnittlig minskning av gini, följt av **trivs_1**, se figur 6. Motivation i arbetet har störst betydelse om en anställd inte kan se sig jobba kvar efter 5 år inom byggbranschen, följt av trivsel på jobbet.

Tabell 44: Variabelnamn med tillhörande fråga för byggbranschen, Random Forest

Variabelnamn	Fråga
trivs_8	Jag känner motivation i mitt arbete
trivs_1	Jag trivs på mitt företag
utveck_3	Det finns bra utvecklingsmöjligheter inom X för mig
utveck_4	Jag känner att min kompetens tas tillvara på min arbetsplats
utveck_1	Jag får den kompetensutveckling jag behöver
trivs_2	Jag kan påverka min arbetssituation
utveck_5	Jag känner att det satsas på mig rent utvecklingsmässigt inom X
utveck_6	Jag är nöjd med min utveckling i arbetet det senaste året
ledar_1	Jag får tydlig feedback på min arbetsinsats från min närmaste chef
ålder	ålder
trivs_13	Samarbetet med mina kollegor på min arbetsplats fungerar bra

Tabell 44: Variabelnamn med tillhörande fråga för byggbranschen, Random Forest *(continued)*

Variabelnamn	Fråga
utveck_2	Jag tar själv initiativ för att få den utveckling jag behöver
kön	kön

Tabell 44 visar variablerna som ingår i Random Forest modellen och dess motsvarande fråga från enkätundersökningen.

Tabell 45: Förväxlingsmatris för byggbranschen, Random Forest

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	1483	370	3522	798
	nej	60	164	152	399

Tabell 45 visar förväxlingsmatrisen för validerings- respektive träningsdata.

Tabell 46: Träffsäkerhet för byggbranschen, Random Forest

Valideringsdata	Träningsdata
0.793	0.805

Träffsäkerheten för validerings- samt träningsdata är 80% och 81%, se tabell 46.

Tabell 47: Sensitivitet, specificitet och AUC för byggbranschen, Random Forest

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.8	0.73	0.766	0.82	0.72	0.77

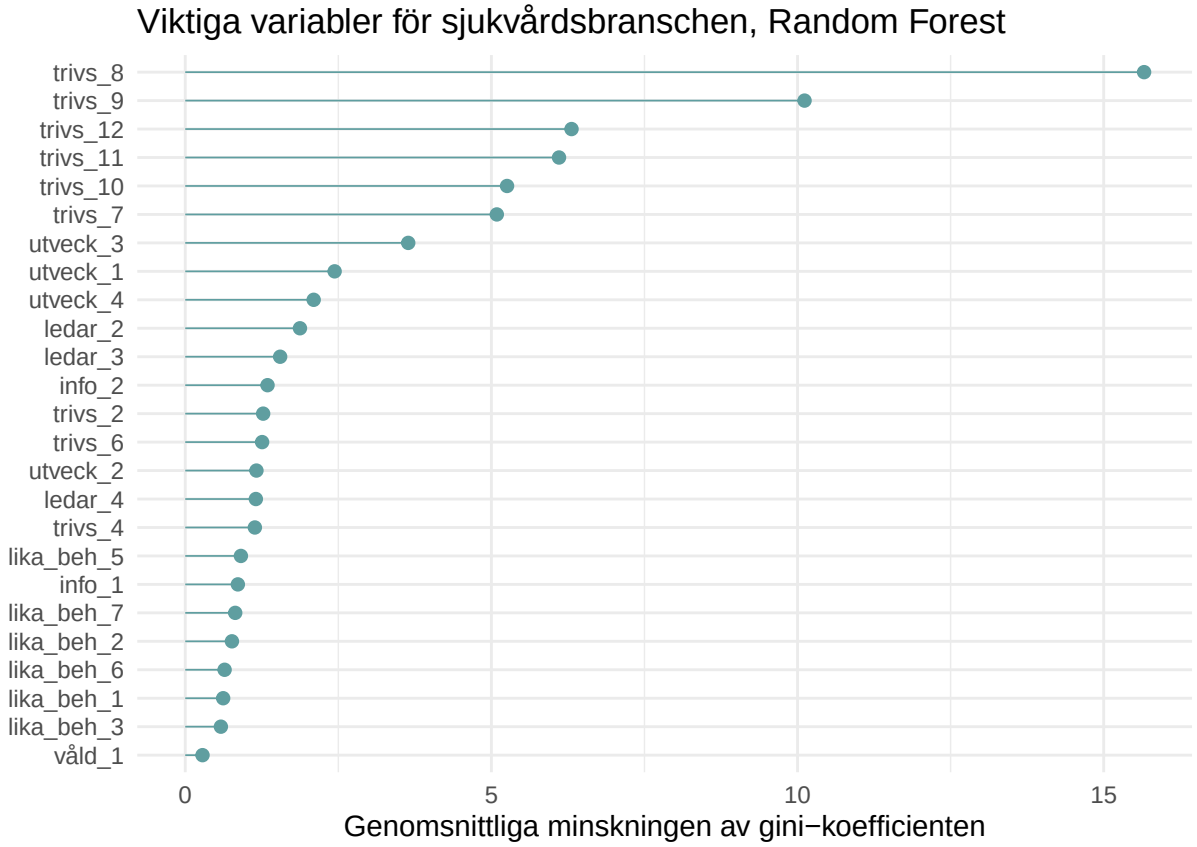
För valideringsdata är sensitiveten 80%, specificiteten 73% och AUC-värdet 77%, se tabell 47. För träningsdata är det liknande procentsatser som valideringsdata, vilket betyder att modellen inte har överanpassats.

4.3.2 Sjukvård

En Random Forest modell för sjukvårdsbranschen har skattats genom fyra hyperparametrar, antal trä att gro, antalet variabler som slumpmässigt väljs vid varje split i trädet, antalet observationer som krävs för att ett löv i trädet ska kunna skapas och maximala antalet noder som tillåts i trädet. Antalet variabler som slumpmässigt väljs vid varje split i trädet har valts till $\lfloor \sqrt{p} \rfloor = \lfloor \sqrt{26} \rfloor = 5$ där p är antal förklarande variabler. För de tre andra hyperparametrarna har korsvalidering utförts för att hitta de optimala begränsningarna som ger hög AUC och undviker överanpassning. Efter korsvalideringen har följande begränsningar valts:

- Antal träd att gro: 500
- Antalet variabler som slumpmässigt väljs vid varje split i trädet: 5
- Antalet observationer som krävs för att ett löv i trädet ska kunna skapas: 15
- Maximala antalet noder som tillåts i trädet: 13

Nedanför återfinns ett diagram som visar de variabler som har störst betydelse för klassificeringen av träden. Ju högre värde på den genomsnittliga minskningen av Gini-koefficienten, desto större betydelse har variabeln i modellen.



Figur 7: Viktiga variabler för sjukvårdsbranshcen, Random Forest

Variabeln **trivs_8** har högst genomsnittlig minskning av gini, följt av **trivs_9**, se figur 7. Motivation i arbetet har således störst betydelse om en anställd inte kan se sig jobba kvar efter 2/3 år inom sjukvårdsbranschen, följt av nöjdhet av sin arbetsgivare.

Tabell 48: Variabelnamn med tillhörande fråga för sjukvårdsbranschen, Random Forest

Variabelnamn	Fråga
trivs_8	Jag känner motivation i mitt arbete
trivs_9	Jag är som helhet nöjd med X som arbetsgivare
trivs_12	Jag rekommenderar gärna min arbetsgivare för andra
trivs_11	Jag är stolt över att arbeta i X
trivs_10	Min arbetssituation uppfyller de förväntningar som jag har på mitt arbete
trivs_7	Jag är som helhet nöjd med min arbetssituation
utveck_3	Det finns bra utvecklingsmöjligheter inom X för mig
utveck_1	Jag får den kompetensutveckling jag behöver
utveck_4	Jag känner att min kompetens tas tillvara på min arbetsplats
ledar_2	Jag är som helhet nöjd med min närmaste chef
ledar_3	Min chef tar snabbt tag i problem som rör konflikter och relationer
info_2	Jag har lätt att hitta den information jag behöver i mitt arbete
trivs_2	Jag kan påverka min arbetssituation
trivs_6	Det finns tid för återhämtning efter perioder av stress på jobbet
utveck_2	Jag tar själv initiativ för att få den utveckling jag behöver
ledar_4	Min chef arbetar för att skapa ett öppet, inkluderande och tillåtande klimat
trivs_4	Jag har en rimlig arbetsbelastning
lika_beh_5	På min arbetsplats behandlas alla lika oavsett religion eller annan trosuppfattning
info_1	Jag tar själv ansvar för att söka den information jag behöver i mitt arbete
lika_beh_7	På min arbetsplats behandlas alla lika oavsett sexuell läggning
lika_beh_2	Jag upplever att alla på X ges lika rättigheter och möjligheter oavsett ålder
lika_beh_6	På min arbetsplats behandlas alla lika oavsett funktionsnedsättning
lika_beh_1	På min arbetsplats behandlas alla lika oavsett kön
lika_beh_3	På min arbetsplats behandlas alla lika oavsett etnisk tillhörighet
våld_1	Har du blivit utsatt för hot eller våld i tjänsten under det senaste året

Tabell 48 innehåller variabelnamnen och tillhörande frågor från enkätundersökningarna som användes i träden.

Tabell 49: Förväxlingsmatris för sjukvårdsbranschen, Random Forest

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	506	138	1292	261
	nej	22	57	44	130

Resultaten i tabell 49 visar förväxlingsmatrisen för både validerings- och träningsdata.

Tabell 50: Träffsäkerhet för sjukvårdsbranschen, Random Forest

Valideringsdata	Träningsdata
0.779	0.823

Träffsäkerheten är lite högre för träningsdata (82%) än för valideringsdata (78%), vilket skulle kunna indikera på en överanpassad modell, se tabell 50.

Tabell 51: Sensitivitet, specificitet och AUC för sjukvårdsbranschen, Random Forest

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.79	0.72	0.754	0.83	0.75	0.79

För valideringsdata är sensitiveten 79%, specificiteten 72% och AUC-värdet 75%, vilket är väldigt liknande resultat som Random Forest modellen för byggbranschen, se tabell 51.

4.4 XGBoosting

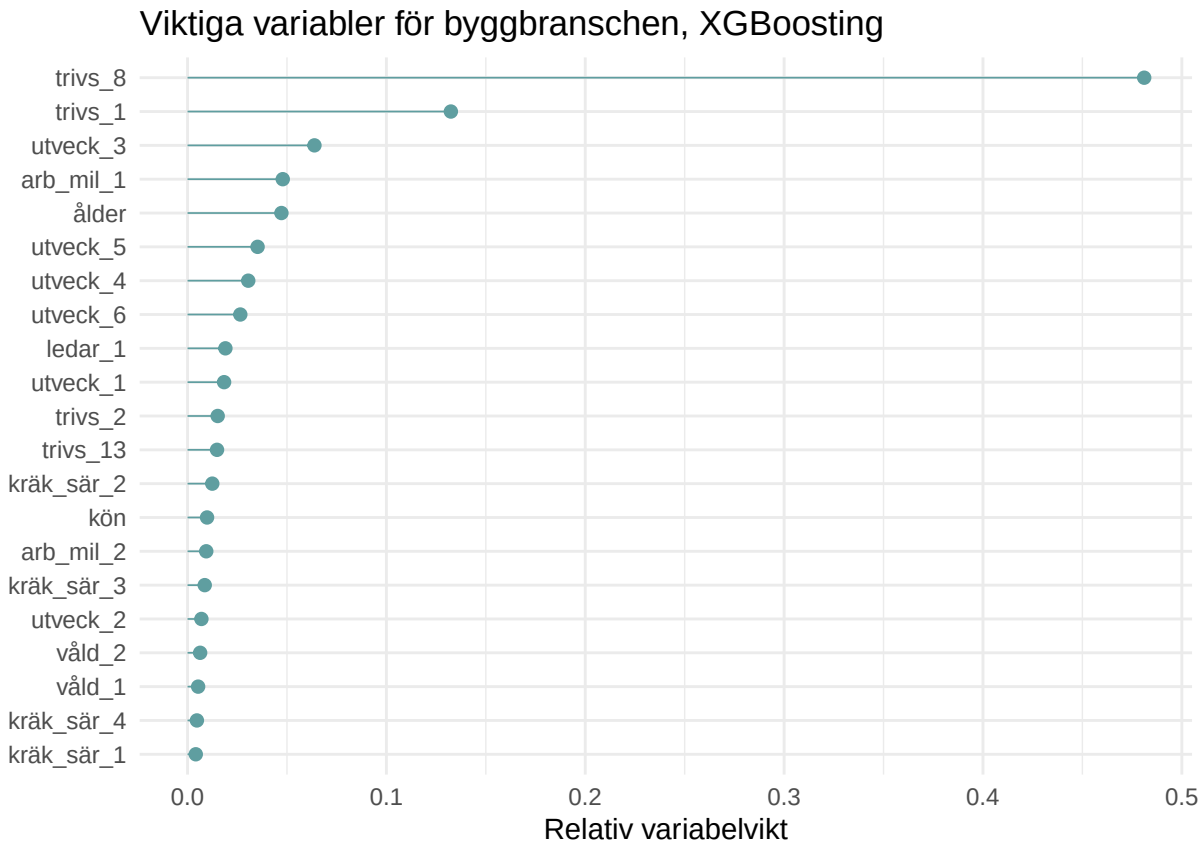
4.4.1 Bygg

En XGBoost modell för byggbranschen har skattats genom fem hyperparametrar, maxdjup, antalet iterationer av gradient boosting, vikt på positiva och negativa klasser, η och γ. Vikt på positiva och negativa klasser har valts till förhållandet mellan "inte nej" och "nej" på responsavariabelns träningsdata.

För de fyra andra hyperparametrarna har korsvalidering utförts för att hitta de optimala begränsningarna som ger hög AUC och undviker överanpassning. Efter korsvalideringen har följande begränsningar valts:

- Maximalt träd-djup: 5
- Antalet iterationer av gradient boosting: 5
- Vikt på positiva och negativa klasser: 0.88
- η: 0.35
- γ: 0.2

Nedanför återfinns ett diagram som visar de variabler som har störst betydelse för klassificeringen av modellen. Ju högre värde på den relativa variabelvikten, desto större betydelse har variabeln i modellen.



Figur 8: Viktiga variabler för byggbranschen, XGBoosting

Variabeln **trivs_8** har överlägset högst relativ variabelvikt, följt av **trivs_1**, se figur 8. Motivation i arbetet har följaktligen störst betydelse om en anställd inte kan se sig jobba kvar efter 5 år inom byggbranschen, följt av trivsel på jobbet.

Tabell 52: Variabelnamn med tillhörande fråga för byggbranschen, XGBoosting

Variabelnamn	Fråga
trivs_8	Jag känner motivation i mitt arbete
trivs_1	Jag trivs på mitt företag
utveck_3	Det finns bra utvecklingsmöjligheter inom X för mig
arb_mil_1	Jag har en bra fysisk arbetsmiljö
ålder	ålder
utveck_5	Jag känner att det satsas på mig rent utvecklingsmässigt inom X
utveck_4	Jag känner att min kompetens tas tillvara på min arbetsplats
utveck_6	Jag är nöjd med min utveckling i arbetet det senaste året
ledar_1	Jag får tydlig feedback på min arbetsinsats från min närmaste chef
utveck_1	Jag får den kompetensutveckling jag behöver
trivs_2	Jag kan påverka min arbetssituation

Tabell 52: Variabelnamn med tillhörande fråga för byggbranschen, XGBoosting *(continued)*

Variabelnamn	Fråga
trivs_13	Samarbetet med mina kollegor på min arbetsplats fungerar bra
kräk_sär_2	Jag tycker att det finns bra rutiner på X för att hantera kränkande särbehandling
kön	kön
arb_mil_2	Min arbetsmiljö är trygg och säker
kräk_sär_3	Jag vet vem jag ska vända mig till om kränkande särbehandling förekommer
utveck_2	Jag tar själv initiativ för att få den utveckling jag behöver
våld_2	Tycker du att det förekommer hot eller våld på X
våld_1	Har du blivit utsatt för hot eller våld i tjänsten under det senaste året
kräk_sär_4	På min arbetsplats får den som är utsatt för kränkande särbehandling snabbt hjälp
kräk_sär_1	På mitt arbete är det tydligt att kränkande särbehandling inte accepteras

Tabell 52 innehar alla variabelbeteckningar med dess respektive fråga från enkätundersökningarna som används i modellen.

Tabell 53: Förväxlingsmatris för byggbranschen, XGBoosting

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	1727	317	4029	712
	nej	79	177	187	436

Resultaten i tabell 53 visar förväxlingsmatrisen för både validerings- och träningsdata.

Tabell 54: Träffsäkerhet för byggbranschen, XGBoosting

Valideringsdata	Träningsdata
0.828	0.832

Träffsäkerheten är 83% för både tränings- och valideringsdata, vilket innebär att modellen inte har överanpassat sig, se tabell 54.

Tabell 55: Sensitivitet, specificitet och AUC för byggbranschen, XGBoosting

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.84	0.69	0.768	0.85	0.7	0.775

För valideringsdata är sensitiveten 84%, specificiteten 69% och AUC-värdet 77%, vilket innebär att modellen har svårare att klassificera de anställda som inte ser sig jobba kvar efter 5 år, se tabell 55.

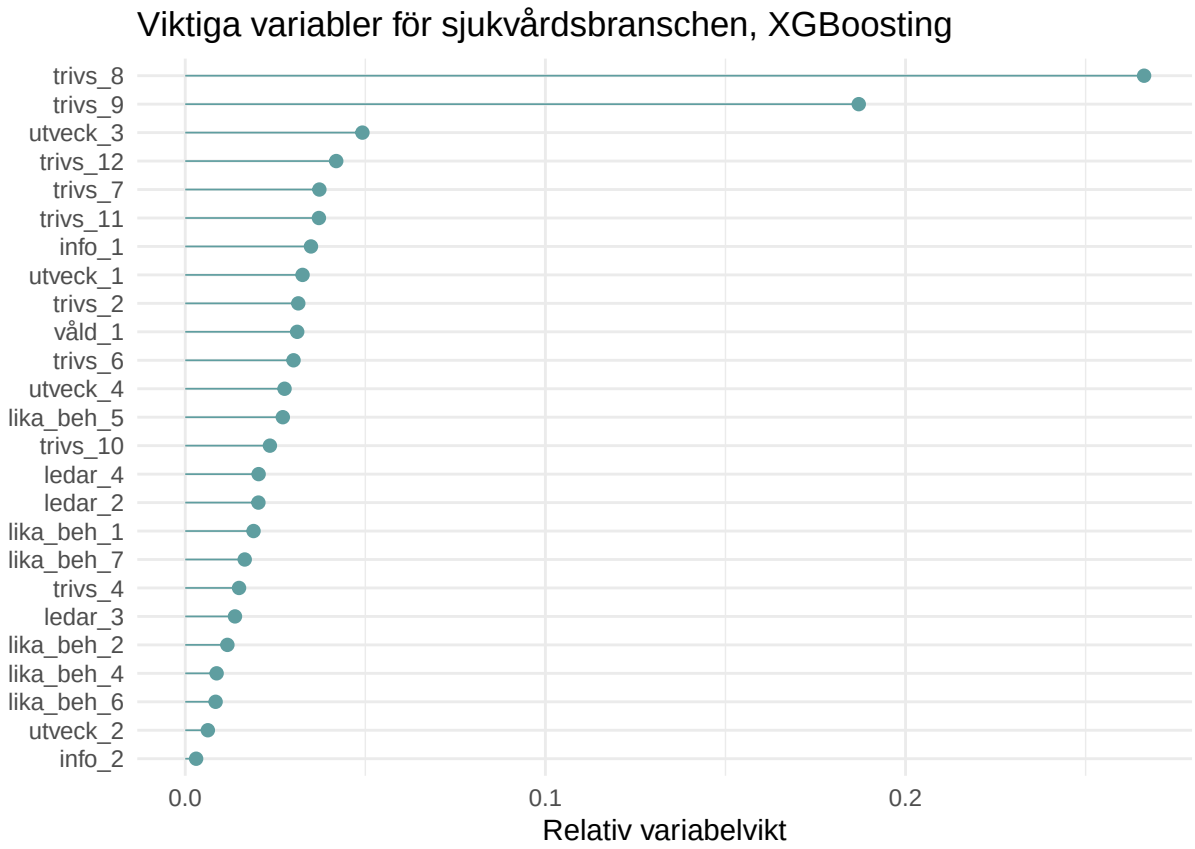
4.4.2 Sjukvård

En XGBoost modell för sjukvårdsbranschen har skattats genom fem hyperparametrar, maxdjup, antalet iterationer av gradient boosting, vikt på positiva och negativa klasser, η och γ. Vikt på positiva och negativa klasser har valts till förhållandet mellan "inte nej" och "nej" på responsavariabelns träningsdata.

För de fyra andra hyperparametrarna har korsvalidering utförts för att hitta de optimala begränsningarna som ger hög AUC och undviker överanpassning. Efter korsvalideringen har följande begränsningar valts:

- Maximalt träd-djup: 5
- Antalet iterationer av gradient boosting: 5
- Vikt på positiva och negativa klasser: 0.90
- η: 0.40
- γ: 0.2

Nedanför återfinns ett diagram som visar de variabler som har störst betydelse för klassificeringen av modellen. Ju högre värde på den relativa variabelvikten, desto större betydelse har variabeln i modellen, se figur 8.



Figur 9: Viktiga variabler för sjukvårdsbranschen, XGBoosting

Variabeln **trivs_8** har högst relativ variabelvikt, följt av **trivs_9**, se figur 9. Motivation i arbetet har således störst betydelse om en anställd inte kan se sig jobba kvar efter 2/3 år inom sjukvårdsbranschen, följt av nöjdhet av sin arbetsgivare.

Tabell 56: Variabelnamn med tillhörande fråga för sjukvårdsbranschen, XGBoosting

Variabelnamn	Fråga
trivs_8	Jag känner motivation i mitt arbete
trivs_9	Jag är som helhet nöjd med X som arbetsgivare
utveck_3	Det finns bra utvecklingsmöjligheter inom X för mig
trivs_12	Jag rekommenderar gärna min arbetsgivare för andra
trivs_7	Jag är som helhet nöjd med min arbetssituation
trivs_11	Jag är stolt över att arbeta i X
info_1	Jag tar själv ansvar för att söka den information jag behöver i mitt arbete
utveck_1	Jag får den kompetensutveckling jag behöver
trivs_2	Jag kan påverka min arbetssituation
våld_1	Har du blivit utsatt för hot eller våld i tjänsten under det senaste året
trivs_6	Det finns tid för återhämtning efter perioder av stress på jobbet
utveck_4	Jag känner att min kompetens tas tillvara på min arbetsplats
lika_beh_5	På min arbetsplats behandlas alla lika oavsett religion eller annan trosuppfattning
trivs_10	Min arbetssituation uppfyller de förväntningar som jag har på mitt arbete
ledar_4	Min chef arbetar för att skapa ett öppet, inkluderande och tillåtande klimat
ledar_2	Jag är som helhet nöjd med min närmaste chef
lika_beh_1	På min arbetsplats behandlas alla lika oavsett kön
lika_beh_7	På min arbetsplats behandlas alla lika oavsett sexuell läggning
trivs_4	Jag har en rimlig arbetsbelastning
ledar_3	Min chef tar snabbt tag i problem som rör konflikter och relationer
lika_beh_2	Jag upplever att alla på X ges lika rättigheter och möjligheter oavsett ålder
lika_beh_4	På min arbetsplats behandlas alla lika oavsett könsöverskridande identitet eller uttryck
lika_beh_6	På min arbetsplats behandlas alla lika oavsett funktionsnedsättning
utveck_2	Jag tar själv initiativ för att få den utveckling jag behöver
info_2	Jag har lätt att hitta den information jag behöver i mitt arbete

Tabell 56 innehar alla variabelbeteckningar med dess respektive fråga från enkätundersökningarna som används i modellen.

Tabell 57: Förväxlingsmatris för sjukvårdsbranschen, XGBoosting

		Predikterad klass			
		Valideringsdata		Träningsdata	
		inte nej	nej	inte nej	nej
Sann klass	inte nej	899	307	2125	674
	nej	29	105	68	258

Resultaten i Tabell 57 visar förväxlingsmatrisen för både validerings- och träningsdata.

Tabell 58: Träffsäkerhet för sjukvårdsbranschen, XGBoosting

Valideringsdata	Träningsdata
0.749	0.763

Träffsäkerheten 75% för träningsdata 76% för valideringsdata, vilket innebär att modellen inte har överanpassat sig, se tabell 58.

Tabell 59: Sensitivitet, specificitet och AUC för sjukvårdsbranschen, XGBoosting

	Valideringsdata			Träningsdata		
	Sensitivitet	Specificitet	AUC	Sensitivitet	Specificitet	AUC
inte nej	0.75	0.78	0.765	0.76	0.79	0.775

För valideringsdata är sensitiveten 75%, specificiteten 78% och AUC-värdet 77%, se tabell 59. Modellen har en högre träffsäkerhet för att klassificera de anställda som inte ser sig jobba kvar efter 2/3 år jämfört med att klassificera de som ej ser sig inte jobba kvar.

4.5 Jämförelse av modeller

Tabell 60: Jämförelse av modeller, byggbranschen

	Träffsäkerhet	Sensitivitet	Specificitet	AUC
Logistisk regression	0.80	0.81	0.72	0.76
Klassificeringsträd	0.81	0.83	0.65	0.74
Random Forest	0.79	0.80	0.73	0.77
XGBoosting	0.83	0.84	0.69	0.77

Tabell 60 visar resultaten för träffsäkerhet, sensitivitet, specificitet och AUC baserat på valideringsdata för bygg. Resultaten visar att logistisk regression har en träffsäkerhet på 80%, sensitivitet på 83%, specificitet på 72% och AUC-värde på 76%. Detta är en av de bättre modellerna när balanseringen av sensitiviteten och specificiteten beaktas.

Klassificeringsträdet visar på lägst AUC-värde 74%, en sensitivitet och specificitet på 83% respektive 65% och en träffsäkerhet på 81%.

Random Forest är den modell som har mest balanserad sensitivitet och specificitet på 80% respektive 73%. Det utläses även att AUC-värdet är på 77% vilket är ett av de högsta bland modellerna. Dock har modellen en träffsäkerhet på 79%, vilket är 4 procentenheter lägre än den modellen med högst träffsäkerhet.

XGBoosting visar på den högsta träffsäkerheten (83%) samtidigt som sensitiviteten är högre än specificiteten 84% respektive 69% med ett AUC-värde på 77%.

Tabell 61: Jämförelse av modeller, sjukvårdsbranschen

	Träffsäkerhet	Sensitivitet	Specificitet	AUC
Logistisk regression	0.73	0.73	0.76	0.74
Klassificeringsträd	0.83	0.86	0.55	0.70
Random Forest	0.78	0.79	0.72	0.75
XGBoosting	0.75	0.75	0.78	0.76

Tabell 61 visar träffsäkerhet, sensitiviteten, specificitet och AUC för valideringsdata baserat på sjukvård. Det utläses att logistisk regression har lägst träffsäkerhet på 73%, vilket är 10 procentenheter lägre än den modellen med högst träffsäkerhet. Sensitiviteten och specificiteten är balanserad på 73% respektive 76% samt ett AUC-värde på 74%.

Klassificeringsträdet visar högsta träffsäkerhet på 83%, men denna höga träffsäkerhet beror främst på en obalanserad sensitivitet och specificitet på 86% respektive 55%, vilket återspeglas i det låga AUC-värdet på 70%. Detta betyder att modellen har svårt att klassificera de individer som inte kan se sig jobba kvar.

Random Forest har relativt hög träffsäkerhet 78% och kan särskilja klasserna tydligare än klassificeringsträdet, då sensitiviteten och specificiteten är 79% respektive 72% med ett AUC-värde på 75%.

XGBoosting har en träffsäkerhet på 75%, sensitivitet och specificitet på 75% respektive 78% samt ett AUC-värde på 76%.

5 Diskussion

I frågeställningarna undersöks det om det är möjligt att skapa en "bra modell" som kan prediktera om en anställd inte kan se sig jobba kvar på sin nuvarande arbetsplats. En bra modell anser vi därför vara en modell som har hög träffsäkerhet för att klassificera de anställda som inte kan se sig arbeta kvar (specificitet), samtidigt som modellen har en hög total träffsäkerhet. Därför används AUC som ett mått för att identifiera den bästa modellen, eftersom det tar hänsyn till både specificitet och träffsäkerhet.

Baserat på högst AUC-värde anses Random Forest och XGBoost vara de bästa modellerna för att prediktera om en anställd inte kan se sig arbeta kvar på sin arbetsplats för bygg- och sjukvårdsbranschen. Dessa modeller har balanserad sensitivitet och specificitet, vilket ökar träffsäkerheten på den underrepresenterade klassen som låg i fokus samtidigt som den totala träffsäkerheten är relativ hög.

Foley (2019) stötte på ett liknande problem med obalanserade klasser, men valde att inte fokusera på att balansera sensitivitet och specificitet, vilket ledde till att hennes modell hade högre träffsäkerhet 89% jämfört med våra modeller som hade lägre träffsäkerheter. Hon fokuserade på att maximera träffsäkerheten och därmed försummade specificiteten, vilket gjorde att Foleys modell var mindre tillförlitlig för att särskilja de två klasserna åt.

Av de samtliga klassificeringsträd för bygg- och sjukvårdsdata har frågan "Jag känner mig motiverad i mitt arbete" varit den viktigaste variabeln. I modeller uppbyggda på träd går det ej att tolka effekten av variabler, vilket gör de modellerna är mindre tolkningsbara än den logistiska regressionen.

Den logistiska regressionsmodellen för bygg hade de tre signifikanta variablerna "ålder", "Jag känner mig motiverad i mitt arbete" och "Det finns bra utvecklingsmöjligheter inom X för mig". Oddskvoterna under 1 på de två sistnämnda variablerna var rimliga, då de som svarade "Instämmer helt" hade en minskad risk att se sig sluta. Variabeln "ålder" hade också en oddskvot under 1, vilket innebar att de personer som är äldre än 35 år minskar risken att se sig sluta.

Den logistiska regressionsmodellen för sjukvård hade också tre signifikanta variabler: "Jag känner motivation i mitt arbete", "Jag är som helhet nöjd med X som arbetsgivare" och "Jag tar själv ansvar för att söka den informationen jag behöver i mitt arbete". De två första frågorna hade en minskad risk att en anställd kan se sig sluta om den anställda svara "Instämmer helt". Det som var intressant med den logistiska regression för sjukvård var att den sista frågan "Jag tar själv ansvar för att söka den informationen jag behöver i mitt arbete" hade en omvänd effekt jämfört med de andra två frågorna. Det innebär att individer som svarar "Instämmer inte alls" hade en minskad risk att se sig sluta. Anledningen till detta samband är svårt att förstå, då det inte verkar finnas någon logik i det till skillnad från de andra frågorna där det är mer rimliga samband.

En tidigare studie av Gomomo (2015) undersökte sambanden mellan arbetstillfredsställelse och avsikt att säga upp sig. Här fann författaren ett negativt samband mellan arbetstillfredsställelse och avsikt att säga upp sig, vilket är ett liknande resultat som vi fann i de logistiska regressionsmodellerna på frågan "Jag känner motivation i mitt arbete". Detta stärker resultatet att en anställd har en minskad risk att se sig sluta genom att ha en hög trivsel på sin arbetsplats.

En stor del av arbetet med denna uppsats har varit att hantera datamaterialet och det har uppstått flera situationer där val har gjorts som kan påverka resultatet. Ett exempel är att "vet ej"/"vill ej uppge" svar hanterades som saknade värden på förklaringsvariablerna, vilket kan ha en negativ påverkan på resultatens kvalité och tillförlighet. Detta beror på att det finns en risk att det är en specifik grupp som svarar "vet ej"/"vill ej uppge" på många frågor, vilket gör att denna potentiella grupp inte kommer att ingå i analysen.

Inom responsvariabeln (Jag tror att jag kommer att arbeta kvar i X om X år) har hanteringen av "vet ej"/"vill ej uppge" skiljt sig från förklaringsvariablerna, då de har kodats om till "ja" på responsvariabeln. Detta beror på att det var en stor andel som hade svarat "vet ej"/"vill ej uppge", vilket hade gjort att väldigt många

observationer hade behövts tas bort om de hade kodats som saknat värde. På grund av att frågeställningarna handlar om en anställd inte kan ses jobba kvar, gjordes därför en binär uppdelning på responsvariabeln med de anställda som svara “nej” och de som svara “inte nej”, vilket då inkluderar de som svarade ”vet ej”/“vill ej uppge”. Denna hantering skulle kunna leda till en överdriven andel av ”inte nej” svar, vilket kan påverka modellernas förmåga att korrekt identifiera variabler som är förknippade med att anställda inte kan se sig jobba kvar efter X år. En till nackdel är att det är svårt att tolka resultatet av modellerna, eftersom det är oklart vad “ja” egentligen betyder när det inkluderar ”vet ej”/“vill ej uppge”-svaren.

För att hantera ”vet ej”/“vill ej uppge” svar på ett bättre sätt och minska risken för negativ påverkan på resultatet, finns det olika tekniker som skulle kunna användas. En möjlighet hade varit att använda någon imputationsteknik som innebär att ”vet ej”/“vill ej uppge” ersätts med uppskattade värden. En annan möjlighet hade varit att de hade kategoriserats som en egen kategori, separat från de övriga svarsalternativen.

Denna rapport grundar sig på svar från respondenter från en medarbetsundersökning, vilket gör att vissa av frågorna kommer vara väldigt lika varandra och därmed kommer det finnas multikollinearitet mellan frågorna. Detta blir ett problem när modeller ska tränas och tolkas, då multikollinearitet bidrar till ökad osäkerhet och minskad precision i parameteruppskattningarna, vilket i sin tur kan göra det svårt att dra korrekta slutsatser om vilka frågor som är viktigast för att förklara variationen i responsvariabeln. En möjlig lösning för att hantera multikollinearitet och öka precisionen i parametrarnas uppskattningar skulle vara att använda faktoranalys. Genom att extrahera underliggande faktorer från de korrelerade variablerna kan faktoranalysen reducera antalet variabler och minska problemet med multikollinearitet.

6 Slutsats

Syftet med denna studie var att utveckla en prediktiv modell för att identifiera anställda som inte kan se sig fortsätta arbeta på sin nuvarande arbetsplats. Samtidigt som studien strävade efter att undersöka effekter mellan att inte arbeta kvar på sin nuvarande arbetsplats och frågor från medarbetarundersökning som handlar om trivsel, lika behandling, kompetensutveckling, ledarskap och våld. För att besvara syftet ställds två frågeställningar upp.

- Går det att göra en bra modell som predikterar om en anställd inte kan se sig jobba kvar på sin nuvarande arbetsplats?
- Vilka specifika frågor inom medarbetarundersökningar har störst effekt på att en anställd inte kan se sig arbeta kvar på sin nuvarande arbetsplats?

Med en träffsäkerhet 79% och specificitet på 73% går det att prediktera om en anställd inte kan se sig jobba kvar på sin nuvarande arbetsplats efter 5 år inom byggbranschen genom att använda Random Forest, vilket kan ses som en förhållandevis bra modell.

Med en träffsäkerhet på 75% och en specificitet på 78% går det att prediktera om en anställd inte kan se sig jobba kvar på sin nuvarande arbetsplats efter 2/3 år inom sjukvårdsbranschen genom att använda XGBoosting, vilket kan ses som en förhållandevis bra modell.

Frågan “Jag känner motivation i mitt arbete” har störst effekt på att en anställd inte kan se sig arbeta kvar på sin nuvarande arbetsplats.

Referenser

Breiman, L., & Cutler, A. (2022). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. CRAN. https://CRAN.R-project.org/package=randomForest

Carrington, A., Manuel, D., Fieguth, P., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., Mclnnes, M., & Holzinger, A. (2023). *Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation*. IEEE. https://ieeexplore-ieee-org.e.bibl.liu.se/stamp/stamp.jsp?tp=&arnumber=9693294

Chen, T., & Guesrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. arXiv. https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2023). *xgboost: Extreme Gradient Boosting*. CRAN. https://CRAN.R-project.org/package=xgboost

Foley, A. (2019). *Using Machine Learning to Predict Employee Resignation in the Swedish Armed Forces.*[Kandidatuppsats, Kungliga Tekniska högskolan]. Diva. http://kth.diva-portal.org/smash/get/diva2:1376923/FULLTEXT01.pdf

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2 uppl.). O'Reilly Media, Inc.

Gomomo, N. (2014). *An investigation into the relationship of job satisfaction, organisational commitment and the intention to quit among academics and administrative employees at the University of Fort Hare.*[Masteruppsats, University of Fort Hare]. Seals. http://vital.seals.ac.za:8080/vital/access/manager/Repository/vital:27553?site_name=GlobalView

Karlsson, S., Perttersson, M., Huitfeld, B., & Ribe, M. (2011). *Svenska statistikfrämjandets etiska kod för statistiker och statistisk verksamhet*. Statistiskfrämjande. https://statistikframjandet.se/wp-content/uploads/2010/12/etisk_kod_final.pdf

Newbold, P., Carlson, W., & Thorne, B. (2013). *Statistics for business and economics* (8 uppl.). Pearson.

Tan, P., Steinbach, M., & Karpatne, V. (2020). *Introduction to Data Mining*. (2 uppl.). Pearson.

Therneau, T., & Atkinson, B. (2022). *rpart: Recursive Partitioning and Regression Trees*. CRAN. https://CRAN.R-project.org/package=rpart

XGBoost. (2022). *XGBoost Parameters*. XGBoost. https://xgboost.readthedocs.io/en/stable/parameter.html

Zhang, Z., Zhu, X., & Liu, D. (2022). *Model of Gradient Boosting Random Forest Prediction*. IEEE. https://doi.org/10.1109/ICNSC55942.2022.10004112

Xin, Z. (2022). *The Influence of Work Environment on Employee's Psychological Relationship: A Case Study of Japanese Literature*. Hindawi. https://doi.org/10.1155/2022/5387795

7 Bilaga

7.1 Bilaga A

I bilaga A visas den R-kod som har använts för att hantera datamaterialet.

```
# ##############################################################################
# # DATAHANTERING
# ##############################################################################

# FUNKTION SOM ÄNDRAR SKALOR ------------------------------------------------------------------

get_data <- function(sjuk_or_bygg = NULL){

# Alla variabler med skala 1-5
h1_h6 <- 8:41
for(i in h1_h6){

  df[df[,i] == "" ,i]<- "NA"
  df[df[,i] == ("0") | df[,i] == ("-1") | df[,i] == ("6"),i] <- "vet ej"

}

# KÖN
df[df[,7] == "1",7] <- "kvinna"
df[df[,7] == "2",7] <- "man"
df[df[,7] == "Kvinna",7] <- "kvinna"
df[df[,7] == "Man",7] <- "man"
df[df[,7] == "3"|df[,7] == "4"|df[,7] == "5"|
   df[,7] == "Vill ej uppge"| df[,7] == "Annat",7] <- "annat/vill ej uppge"



# våld_1 och våld_2
df[df[,42] == "1",42] <- "ja"
df[df[,42] == "2",42] <- "nej"
df[df[,42] == "3",42] <- "vill ej uppge"


df[df[,43] == "1",43] <- "ja"
df[df[,43] == "2",43] <- "nej"
df[df[,43] == "3",43] <- "vill ej uppge"


# Kränkade särbehandling_1 - Kränkade särbehandling_4
df[df[,44] == "1",44] <- "ja"
df[df[,44] == "2",44] <- "nej"
df[df[,44] == "3",44] <- "vill ej uppge"
```

```
df[df[,45] == "1",45] <- "ja"
df[df[,45] == "2",45] <- "nej"
df[df[,45] == "3",45] <- "vill ej uppge"

df[df[,46] == "1",46] <- "ja"
df[df[,46] == "2",46] <- "nej"
df[df[,46] == "3",46] <- "vill ej uppge"

df[df[,47] == "1",47] <- "ja"
df[df[,47] == "2",47] <- "nej"
df[df[,47] == "3",47] <- "vill ej uppge"


# Ålder
df[df[,6] == "25 år eller yngre",6] <- "<=35 år"
df[df[,6] == "26-35 år",6] <- "<=35 år"
df[df[,6] == "36 år eller äldre",6] <- ">35 år"

df[df[,6] == "1",6] <- "<=35 år"
df[df[,6] == "2",6] <- "<=35 år"
df[df[,6] == "3",6] <- ">35 år"
df[df[,6] == "4",6] <- ">35 år"
df[df[,6] == "5",6] <- ">35 år"
df[df[,6] == "6",6] <- ">35 år"
df[df[,6] == "7",6] <- "Vill ej uppge"




# Responsvariabeln y med bolag som har 1-5 skala
bolag1till5 <- c(7,40,42,43)

for(j in 1:length(bolag1till5)){

  if((df[which(df$number == bolag1till5[j])[1],3] == "NA") &
     (df[which(df$number == bolag1till5[j])[1],4] == "NA")){
    i <- 5
  }else if((df[which(df$number == bolag1till5[j])[1],4] == "NA") &
           (df[which(df$number == bolag1till5[j])[1],5] == "NA")){
    i <- 3
  }else{
    i <- 4
  }

  if(sjuk_or_bygg == "sjuk"){

    # SJUKVÅRD KODNING
    df[df$number == bolag1till5[j] & (df$y5 == "1" | df$y5 == "2" |
                                        df$y3 == "1" | df$y3 == "2" |
```

```
                                        df$y2 == "1" | df$y2 == "2"),i] <- "nej"

    df[df$number == bolag1till5[j] & (df$y5 == "4" | df$y5 == "5" | df$y5 == "3" |
                                        df$y5 == "-1" | df$y5 == "6"|
                                        df$y3 == "4" | df$y3 == "5" | df$y3 == "3" |
                                        df$y3 == "-1" | df$y3 == "6"|
                                        df$y2 == "4" | df$y2 == "5" | df$y2 == "3" |
                                        df$y2 == "-1" | df$y2 == "6"),i] <- "ja"

  }else{

    # BYGG KODNING
    df[df$number == bolag1till5[j] & (df$y5 == "1" |
                                        df$y3 == "1" |
                                        df$y2 == "1" ),i] <- "nej"

    df[df$number == bolag1till5[j] & (df$y5 == "4" | df$y5 == "5" | df$y5 == "3" |
                                       df$y5 == "-1" | df$y5 == "6"|df$y5 == "2" |
                                       df$y3 == "4" | df$y3 == "5" | df$y3 == "3" |
                                       df$y3 == "-1" | df$y3 == "6"|df$y3 == "2" |
                                       df$y2 == "4" | df$y2 == "5" | df$y2 == "3" |
                                       df$y2 == "-1" | df$y2 == "6"| df$y2 == "2"),i] <-"ja"

  }
}



# Responsvariabeln y med bolag som har 1-2 skala
bolag1till2 <- c(5,8,41,48)

for(j in 1:length(bolag1till2)){

  if((df[which(df$number == bolag1till2[j])[1],3] == "NA") &
     (df[which(df$number == bolag1till2[j])[1],4] == "NA")){
    i <- 5
  }else if((df[which(df$number == bolag1till2[j])[1],4] == "NA") &
           (df[which(df$number == bolag1till2[j])[1],5] == "NA")){
    i <- 3
  }else{
    i <- 4
  }

    df[df$number == bolag1till2[j] & (df$y5 == "1" | df$y5 == "3" | df$y5 == "4" |
                                        df$y3 == "1" | df$y3 == "3" | df$y3 == "4" |
                                        df$y2 == "1" | df$y2 == "3" | df$y2 == "4"),i] <- "ja"

    df[df$number == bolag1till2[j] & (df$y5 == "2" |
                                        df$y3 == "2" |
```

```
                                        df$y2 == "2" ),i] <- "nej"


}


df[df == "NA"] <- NA
df[df == "vet ej" | df == "annat/vill ej uppge" |
   df == "vill ej uppge" |df == "Vill ej uppge" ] <- NA

df[,names(df)[c(2:7,42:48)]] <- lapply(df[,names(df)[c(2:7,42:48)]],factor)
df[,names(df)[-c(2:7,42:48)]] <- lapply(df[,names(df)[-c(2:7,42:48)]],as.numeric)


if(sjuk_or_bygg == "bygg"){
  df_bygg <- df[df$type=="Bygg",c(1:3,6:9,15,20,23:27,38:48)]
  return(df_bygg)

}else{
  df_sjukvard <- df[df$type=="Sjukvard",c(1,2,4,5,9,11,13:19,21:24,28:37,40:42,48)]
  return(df_sjukvard)
}

}
```

7.2 Bilaga B

I bilaga B visas den R-kod som har använts för att dela upp data i träning och validering.

```
# ##############################################################################
# # SPLITTAR TRÄNING OCH VALIDERING (TEST )
# ##############################################################################

set.seed(990420)

train_ind <- sample(nrow(df_bygg),nrow(df_bygg)*0.7)
test_b <- df_bygg[-train_ind,]
train_b <- df_bygg[train_ind,]

train_ind <- sample(nrow(df_sjukvard),nrow(df_sjukvard)*0.7)
test_s <- df_sjukvard[-train_ind,]
train_s <- df_sjukvard[train_ind,]
```

7.3 Bilaga C

Bilaga C innehåller R-kod som har använts för att anpassa modellen logistisk regression.

```r
# ##############################################################################
# # LOGISTISK REGRESSION, BYGG OCH SJUKVÅRD
# ##############################################################################


model_b <- glm(jobb_kva_5 ~ ., family = binomial(link = "logit"),
               data = train_b[,c(3,7,11,16)],
               na.action = na.omit)


model_s <- glm(jobb_kva_23 ~ ., family = binomial(link = "logit"),
               data = train_s[,c(7,8,26,29)],
               na.action = na.omit)
```

7.4 Bilaga D

Bilaga D innehåller R-kod som har använts för att anpassa modellen klassificeringsträd.

```r
# ##############################################################################
# # KLASSIFICERINGSTRÄD, BYGG OCH SJUKVÅRD
# ##############################################################################

tree_b <- rpart(
    formula = jobb_kva_5 ~.,
    data =  train_b[,-c(1,2)],
    method = "class",
    control = list(
        minsplit = 50,
        maxdepth = 6,
        cp = 0
        )
)



tree_s <- rpart(
    formula = jobb_kva_23 ~.,
    data = train_s[,-(1:2)],
    method = "class",
    control = list(
        minsplit = 25,
        maxdepth = 4,
        cp = 0
        )
)
```

7.5 Bilaga E

Bilaga E innehåller R-kod som har använts för att anpassa modellen Random Forest.

```
# ##############################################################################
# # RANDOM FOREST, BYGG OCH SJUKVÅRD
# ##############################################################################

tree_random_b <- randomForest(jobb_kva_5~.,
                              data = train_b[,-c(1,2)],
                              ntree = 450,
                              mtry = floor(sqrt(ncol(train_b[,-c(1,2,16)]))),
                              nodesize = 30,
                              maxnodes = 50,
                              na.action = na.omit)

tree_random_s <- randomForest(jobb_kva_23~.,
                              data = train_s[,-(1:2)],
                              ntree = 500,
                              mtry = floor(sqrt(ncol(train_s[,-c(1,2,29)]))),
                              nodesize = 15,
                              maxnodes = 13,
                              na.action = na.omit)
```

7.6 Bilaga F

Bilaga F innehåller R-kod som har använts för att anpassa modellen XGBossting.

```
# ##############################################################################
# # XGBOOSTING, BYGG OCH SJUKVÅRD
# ##############################################################################

# SPLITTAR TRÄNING OCH VALIDERING IGEN PGA ANNAN DATA
train_ind <- sample(nrow(df_bygg_NA),nrow(df_bygg_NA)*0.7)
df_xg_bygg_test <- df_bygg_NA[-train_ind,-c(1:2)]
df_xg_bygg_train <- df_bygg_NA[train_ind,-c(1:2)]

# FORMATERAR OM DATA SÅ SOM XGBOOST VILL HA DEN
df_xg_bygg_train$jobb_kva_5 <-ifelse(df_xg_bygg_train$jobb_kva_5=="ja",0,1)
df_xg_bygg_test$jobb_kva_5 <-ifelse(df_xg_bygg_test$jobb_kva_5=="ja",0,1)

negative_cases <- sum(df_xg_bygg_train$jobb_kva_5==0)
postive_cases <- sum(df_xg_bygg_train$jobb_kva_5==1)


model_bygg <- xgboost(data = data.matrix(df_xg_bygg_train[,-22]),
                      label = df_xg_bygg_train$jobb_kva_5,
                      max.depth = 5,
```

```
                  nround = 5,
                  eta = 0.35,
                  gamma = 0.2,
                  scale_pos_weight = (negative_cases/(negative_cases+postive_cases)),
                  objective = "binary:logistic")



# SPLITTAR TRÄNING OCH VALIDERING IGEN PGA ANNAN DATA
train_ind <- sample(nrow(df_sjukvard_NA),nrow(df_sjukvard_NA)*0.7)
df_xg_sjuk_test <- df_sjukvard_NA[-train_ind,-c(1:2)]
df_xg_sjuk_train <- df_sjukvard_NA[train_ind,-c(1:2)]

# FORMATERAR OM DATA SÅ SOM XGBOOST VILL HA DEN
df_xg_sjuk_train$jobb_kva_23 <-ifelse(df_xg_sjuk_train$jobb_kva_23=="ja",0,1)
df_xg_sjuk_test$jobb_kva_23 <-ifelse(df_xg_sjuk_test$jobb_kva_23=="ja",0,1)

negative_cases <- sum(df_xg_sjuk_train$jobb_kva_23==0)
postive_cases <- sum(df_xg_sjuk_train$jobb_kva_23==1)


model_sjuk <- xgboost(data = data.matrix(df_xg_sjuk_train[,-c(27)]),
                  label = df_xg_sjuk_train$jobb_kva_23,
                  max.depth = 5,
                  nround = 5,
                  eta = 0.40,
                  gamma = 0.2,
                  scale_pos_weight = (negative_cases/(negative_cases+postive_cases)),
                  objective = "binary:logistic") %>% quiet()
```