Stockholms universitet

Prediktion av studieframgång inom kursen Matematik I på Stockholms universitet

Jonathan Pollack

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2014:11
Matematisk statistik
September 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2014:11**
http://www.math.su.se/matstat

Prediktion av studieframgång inom kursen Matematik I på Stockholms universitet

Jonathan Pollack*

September 2014

Sammanfattning

I denna uppsats använder vi oss av data insamlad under höstterminen 2013 på Stockholms universitet i syfte att prediktera studieframgång och genomströmning i kursen Matematik I. Vi utvecklar prediktiva modeller med hjälp av linjär och logistisk regression, naiv Bayesiansk klassifikation samt klassifikationsträd. Modellerna jämförs sedan och analyseras, särskilt utifrån deras AUC, Brier score och precision. Vi anpassar även modeller på äldre och yngre studenter separat. Av metoderna som används finner vi att logistisk regression och klassifikationsträd är särskilt bra lämpade, och att den prediktiva styrkan allmänt är högre för yngre studenter i jämförelse med äldre. Vi jämför även våra modeller med den allmänt tillämpade urvalsmodellen där man går enbart efter gymnasiebetyg och finner att även denna metod har en godtagbar prediktionsförmåga, särskilt för kursmomentet *analys*.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: jhpollack@yahoo.com. Handledare: Martin Sköld.

Abstract

In this paper we use data collected from students studying the first-year math course 'Matematik I' at Stockholm University in the fall of 2013, with the aim of predicting first-year retention rates. Predictive models are built and developed using linear and logistic regression, naive Bayes classification and decision trees. The models are compared and analyzed, especially in terms of their AUC, Brier score and precision. We also fit models on older and younger students separately. Logistic regression and classification trees are found to perform especially well on the dataset, and the models developed on younger students are found to have higher predictive strength compared to those on the older students. We also compare our models to the common selection procedure in higher education, in which only grades from secondary education are used to predict student success rate. We find that this model too makes predictions at an acceptable level, especially for the *calculus* component of the course.

Förord

Denna uppsats utgör ett självständigt arbete omfattande 15 hp som leder till en kandidatexamen i matematisk statistik.

Jag vill tacka Samuel Lundqvist, studierektor för grundutbildningen i matematik, och min handledare Martin Sköld för det nöjsamma tillfället att ha fått utföra detta arbete och för deras hjälp under dess gång. Tom Everitt och Lisa Niklasson förtjänar också ett tack för deras mycket uppskattade insats i insamlandet av data under HT13. Jag vill även tacka min hustru Honey Rakel Pollack för hennes stöd och tålamod.

Innehåll

1 Introduktion

1.1 Matematik I

Matematik I omfattar 30 högskolepoäng och är Stockholms universitets huvudsakliga grundkurs inom matematik på högskolenivå. Den ges varje termin och är en självklar del inom ett flertal kandidatprogram där matematiska färdigheter är en nödvändighet, men kan även läsas fristående eller på distans. För många studenter utgör kursen den första erfarenheten av matematiska studier på universitetsnivå, eller universitetsstudier överhuvudtaget.

Kursen består av de två delmomenten *algebra* och *analys* som examineras separat. Det är även möjligt att läsa kursen på halvfart och alltså fokusera på endast ett av momenten under en termin. Nytt för höstterminen 2014 är möjligheten att välja ett kursupplägg mer lämpligt för studenter inriktade inom fysik. Universitetet erbjuder även den nätbaserade kursen *Förberedande kurs i matematik* under vår sommar och höst, som är avsedd att fungera som en mjukstart inför matematik på universitetsnivå.

Matematik I har som syfte att utveckla och stimulera elevernas matematiska utveckling och innehåller både frivilliga undervisningsmoment såsom föreläsningar och räkneövningar, såväl som obligatoriska examensmoment. För godkänt slutresultat krävs deltagande i e-tentor, seminarier, laborationer samt godkänt resultat vid tentamen (betyg A-E) inom de respektive momenten. För mer utförlig information om kursupplägget hänvisar vi till kursens utbildningsplan (senast reviderad 2014-05-19) som finns tillgänglig på matematiska institutionens hemsida.

1.2 Syfte och metodik

Syftet med denna uppsats är att undersöka om, och i så fall hur väl det går att prediktera studieframgång inom Matematik I, givet information och bakgrundsvariabler som är typiskt tillgängliga vid urval. För detta syfte kommer vi att tillämpa olika typer av regressioner och klassifikationsmetoder av huvudsakligen binär art, med responsvariabel ej godkänd= 0 och godkänd = 1. Vi kommer att undersöka kursens två huvudmoment algebra och analys separat eftersom de också examineras separat, och vi kommer även att jämföra äldre och yngre studenter. Vi undersöker sedan och jämför de olika klassifikationsmodellerna med avseende på prediktiv styrka, och illustrerar slutligen hur modellerna i sig kan fungera som urvalsmetoder till kursen under olika söktryck. Som underlag använder vi anonymiserad data insamlad under kursens gång höstterminen 2013.

Samtliga beräkningar och illustrationer har gjorts med hjälp av programvaran `R`.

1.3 Urval och genomströmning: tidigare studier

Det är av intresse för både studenter och lärosäten att urvalet till högre utbildning fungerar så effektivt och rättvist som möjligt och leder till en god genomströmning. Om ett urvalsinstrument prioriterar felaktigt och bidrar till att en hög andel antagna saknar förkunskaper eller är bristfälliga i sin förmåga att fullfölja studierna, får detta konsekvenser både på individnivå och för utbildningens kvalitet i stort, och kan på lång sikt leda till sämre produktivitet och dålig matchning på arbetsmarknaden.

I Sverige används huvudsakligen snittbetyg från gymnasiet samt resultat från högskoleprovet vid urval, som två skilda inkvoteringsgrupper. Provet har ett uttalat syfte att fungera som en andra chans och bredda rekryteringen till högre studier. Enligt Wikström och Wikström (2012) visar de flesta studier att prediktionsförmågan av genomströmning i det svenska urvalsförfarandet är relativt låg, men något bättre för betygsmedelvärdet än för högskoleprovet. Det bakomliggande skälet tros vara att snittbetyget är en bättre indikation på en individs kapacitet att på lång sikt studera och ta till sig relevanta kunskaper. De menar också att det saknas kunskap om hur urvalsinstrumenten fungerar för olika grupper av sökanden, särskilt för sökande av olika åldrar som erhållit sina betyg vid olika tillfällen; en relevant fråga i takt med att sökande till högre utbildning blivit en alltmer heterogen grupp med avseende på ålder och social bakgrund. I sin egen studie undersöker de om gymnasiebetygen förmåga att prediktera slutförda universitetspoäng på ett ekonomprogram påverkas av tiden mellan utbildningarna, med slutsatsen att inget signifikant samband finns kvar om det gått tre år eller mer mellan gymnasium och antagning. En liknande analys som är mer relevant för matematik bedrevs av Pugh och Lowther (2004), som sammanfattat hur akademisk prestation inom matematik på en nationell nivå i USA kan delvis förklaras av under vilket år man tagit sina senaste matematikpoäng på high school, och på vilken nivå.

Studier kring prediktionsförmågan av genomströmningen inom högre utbildningar är ett område inom vilket det blir vanligare att metodik från maskininlärning (dvs. klassificeringsmetoder från artificiell intelligens) används i större utsträckning. Herzog (2006) gjorde en studie i vilken ett flertal sådana metoder användes och jämfördes med avseende på sin förmåga att förutsäga genomströmning. Chong et al. (2010) har tillämpat och jämfört prediktionsförmågan av klassifikationsträd, MARS (multivariate adaptive regression splines) och neuronnät på genomströmningen av studenter under första studieåret på Arizona State University. Även Alkhasawneh (2011) har utvecklat en modell med hjälp av neuronnät på genomströmningen inom STEM-ämnen på Virginia Commonwealth University, med särskilt intresse för hur den ter sig bland minoritetsgrupper.

Anledning till att maskininlärningsmetoder ökat i popularitet inom området är troligen att det är välkänt att genomströmning påverkas av många komplexa variabler bortom dem som är typiskt tillgängliga vid urval, särskilt på lång sikt, och att alltmer sofistikierade algoritmer som kan upptäcka komplexa strukturer

samtidigt utvecklats inom områden som datautvinning och mönsterigenkänning. Det kan finnas många skäl till att man väljer att avbryta sina studier, inte minst sociala faktorer som kan vara bakomliggande men även kretsa kring utbildningen och arbetsmarknaden i sig. Det är särskilt dessa man är intresserad av att kartlägga i många av de tidigare nämnda studierna. En sådan djupanalys ligger dock bortom omfånget för denna uppsats, men för ett par relevanta studier på läget här i Sverige kan vi hänvisa till Eriksson (2010) som kartlagt orsaker till studieavbrott för dåvarande Högskoleverket, samt Dryler (2013) som har analyserat genomströmning i högskolan i relation till social bakgrund för UK-ämbetet.

2 Beskrivning av data

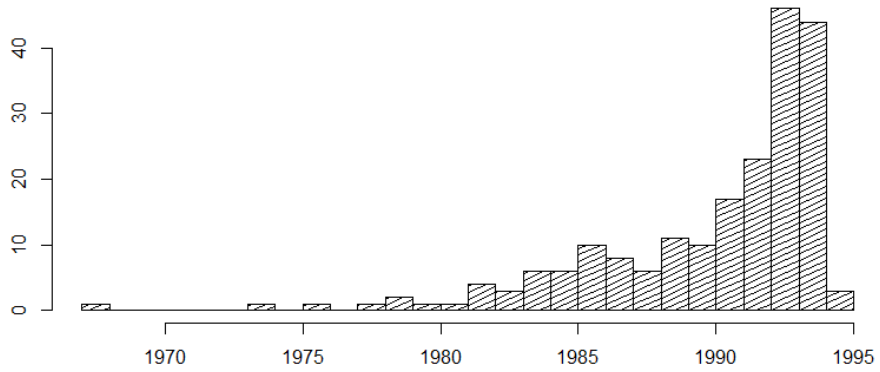
Datamängden samlades in under höstterminen 2013 och innefattar uppgifter på kön, födelseår, program, tidigare snitt- och matematikbetyg, eventuellt resultat på högskoleprov samt prestationer under själva kursens gång på samtliga elever som var registerade på kursen Matematik I under perioden, totalt 498 individer. Av dessa var 188 tidigare registrerade elever, 73 nyregistrerade distanselever och 237 nyregistrerade elever som deltog i undervisningen på campus vid Stockholms universitet.

På grund av höga proportioner av bortfall i de två förstnämnda grupperna kommer vi att fokusera vår analys på de nyregistrerade eleverna vid campus. 41% av tidigare registrerade elever och 63% av nyregistrerade distanselever avbröt sina studier i förtid under terminens gång, medan samma siffra låg på 13% för de nyregistrerade campuseleverna. Vissa individer faller ändå bort ur studien eftersom det av olika skäl saknas motsvarande uppgifter på snittbetyg eller betyg från tidigare matematikkurser; detta beror i enskilda fall på att man t.ex. har utländskt betyg, läst på folkhögskola, eller att uppgifter på betyg helt enkelt saknas.

Efter indelning i de respektive kursmomenten har vi slutligen data på 205 studenter som läst momentet *algebra* samt 192 som läst momentet *analys*. Andelen godkända på respektive kursmoment ligger kring 25%, med något högre genomströmning i algebra-delen. Yngre elever (födda efter 1991) tenderar också att klara kursen i högre utsträckning än äldre. Se tabell A1 i appendix för mer detaljer.

Könsfördelningen bland campus-eleverna ligger på ca 77% män och 23% kvinnor oavsett kursmoment och åldersgrupp, och medianen av deras födelseår är 1992. Ungefär hälften av observationerna kommer alltså från elever som är relativt nyanlända från gymnasiet medan det för övriga har gått minst några år emellan; se figur 1. (Vi kommer senare att dela upp datan kring denna median och jämföra de två grupperna.) För att ge en överblick över vilken typ av studenter som läser kursen redovisar vi även fördelningen av elever över olika kandidatprogram, se tabell A2 i Appendix. Denna information använder vi dock inte i de prediktiva modeller vi senare kommer att utveckla pga den mycket skeva fördelningen (det finns t.ex. endast en representant från lärarprogrammet).

Två tredjedelar av studenterna har skrivit högskoleprovet men endast ett fåtal av dem har resultat som överstiger 1.5. Det saknas nog därför tillräckligt med belägg för att dra några egentliga slutsatser om högskoleprovets inverkan på genomströmningen i kursen.



Figur 1: Fördelning av födelseår bland studenterna.

Eftersom fokus för denna studie ligger på hur väl prediktion kan göras givet några utvalda variabler kommer vi inte att ägna oss åt något modellval i deskriptiv mening, dvs. att försöka hitta modeller med det urval av variabler som bäst beskriver datan. Vi väljer istället ut en mängd variabler på förhand och använder dessa för att anpassa samtliga prediktiva modeller, i jämförande syfte.

De variabler som vi kommer att behålla i analysen för att försöka prediktera studieframgången på Matematik I är: kön; födelseår; betygsmedelvärde; andel MVG bland matematikbetyg från gymnasiet; huruvida man läst Matte E eller ej; huruvida man läst universitetets förberedande kurs i matematik; närvaro av högskoleprovpoäng ≥ 1.5, samt antal år sedan senast tagna matematikpoäng.

3 Teori

I denna del beskrivs kortfattat de statistiska metoder som använts under arbetet. Vi börjar med en genomgång av de olika regressions- och klassifikationsmetoderna som kommer att tillämpas, och går sedan över till beskrivningar av tabeller, kurvor och mått som är vanliga vid klassificeringmodeller i allmänhet.

De flesta metoder som vi kommer att använda oss av är av binär art, och har alltså en responsvariabel Y som antingen är lika med 0 eller 1 (logistisk regression, naiv Bayesiansk klassifikation och trädmodeller). Givet ny data erhålls med dessa metoder sedan motsvarande sannolikhet på intervallet $(0, 1)$ att $Y = 1$.

Då vi har tillgång till studenternas tentamensresultat kommer vi även att tillämpa linjär regression, ur vilken vi erhåller ett predikterat tentamensresultat på intervallet $(0, 30)$. Eftersom 15 poäng räcker för att vara godkänd på kursen kan vi sedan genom division med 30 istället få värden på intervallet $(0, 1)$ och tolka dessa som sannolikheter att klara kursen, och därmed kunna jämföra dem med de andra modellerna med precis samma metodik.

På grund av den begränsade datamängden kommer vi tyvärr inte att kunna partitionera datamaterialet i olika delar för träning, testande och validerande av modellerna, som annars är en vanlig procedur då man utvecklar predikterande modeller (särskilt inom maskininlärning), och vi kommer alltså istället att använda hela datamängden både för modellbyggande och för prediktering.

Sammanfattningsvis använder vi oss av följande variabler i analysen:

X_1	Kön	`gender`	Män = 0, kvinnor = 1
X_2	Födelseår	`year`	Heltal mellan 67 och 94
X_3	Betygssnitt	`snitt`	Medelbetyg gymnasiet
X_4	Andel MVG	`mvg`	En siffra mellan 0 och 1
X_5	Matte E	`matte.E`	Nej= 0, ja = 1
X_6	Förberedande kurs	`prep`	Nej= 0, ja = 1
X_7	Högskoleprov ≥ 1.5	`hskp1.5`	Nej= 0, ja = 1
X_8	Senaste matte	`latest`	Antal år sedan matematikpoäng
Y	Godkänd	`alg/an`	Nej= 0, ja = 1 (binär)
Y	Tentamenspoäng	`alg.p/an.p`	Tentamenspoäng 0-30 (linjär)

3.1 Metoder

3.1.1 Linjär regression

Inom regression är man intresserad av att härleda värdet på en reellvärd responsvariabel Y givet en förklarande variabel X, dvs $\mathsf{E}(Y|X)$. Om X är en vektor av observationer $X^T = (X_1, X_2, ..., X_p)$ kallas detta ofta för *multipel regression*. Med en linjär regressionmodell antar man att sambandet mellan Y och X kan approximeras genom ett linjärt hyperplan av formen

$$\mathsf{E}(Y|X) = f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j.$$

Vanligtvis anpassar man modellen och skattar koefficienterna β genom att minimera det kvadratiska avståndet mellan Y och $f(X)$. Under förutsättning att vi har N par av observationer (x_i, y_i) där varje x_i är en vektor $(x_{i1}, x_{i2}, ..., x_{ip})^T$ vill vi alltså minimera

$$Q(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \quad (1)$$

Låt nu $\mathbf{X}$ beteckna $N\times(p+1)$-matrisen där varje rad är av formen $(1, x_{i1}, x_{i2}, ..., x_{ip})$, β vektorn av samtliga koefficienter och $\mathbf{y}$ vara vektorn av samtliga responsvärden. Vi kan då skriva (1) som skalärprodukten

$$Q(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2,$$

som är en kvadratisk funktion med $p+1$ parametrar och som kan minimeras genom att sätta derivatorna med avseende på $\beta_0, \beta_1, ...\beta_p$ till 0 under förutsättning att $\mathbf{X}$ är av full rang. Därigenom erhålls att skattningarna $\hat{\beta}$ av koefficienterna måste lösa det linjära ekvationssystemet $(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\mathbf{y}$, vilket ger den unika lösningen $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Detta kan även tolkas som att vi gör en ortogonal projektion $\hat{\mathbf{y}}$ på ett underrum U till $\mathbb{R}^N$, där

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Prediktioner av nya observationer x^* kan sedan ges av $\hat{y}^* = x^*\hat{\beta}$.

(Hastie et al., 2008, kap. 3.2.)

3.1.2 Logistisk regression

Då responsvariablen Y istället är binär har vi endast två möjliga responser för varje vektor av observationer X, som vi kan koda som $Y = 0$ eller 1. Beteckna sannolikheten att $Y = 1$ givet att vi har observerat $X = x$ som $\pi(x)$, dvs. $\pi(x) = \mathsf{P}(Y = 1 \mid X = x)$. Med *logistisk regression* antar man att det föreligger något icke-linjärt samband mellan $\pi(x)$ och X, men att $\pi(x)$ istället förändras monotont och kontinuerligt med avseende på X, närmare bestämt genom sambandet

$$\pi(x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} x_j\beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} x_j\beta_j)} = \frac{\exp \mathbf{X}\beta}{1 + \exp \mathbf{X}\beta}.$$

Detta är ofta ett rimligt antagande när responsen är binär eftersom det medför att en förändring i x har en större inverkan på $\pi(x)$ om dess värde är nära 0.5 jämfört med fallet då $\pi(x)$ redan är i närheten av 0 eller 1. Att förändringen är monoton och kontinuerlig innebär att logaritmen av oddskvoten, som även betecknas $\text{logit}[\pi(x)]$, har ett linjärt samband med observationerna:

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \mathbf{X}\beta.$$

Eftersom log-oddskvoten kan anta vilka reella värden som helst har vi därför ett linjärt samband som tidigare, ur vilket vi kan skatta $\hat{\beta}$, men med skillnaden att koefficienterna β_j nu har tolkningen att oddskvoten förändras multiplikativt med faktorn e^{β_j} för varje enhetsökning av x_j.

För varje ny observation x^* kan vi nu skatta sannolikheten att $Y = 1$ genom sambandet

$$\hat{\pi}(x^*) = \frac{\exp x^*\beta}{1 + \exp x^*\beta}.$$

(Agresti, 2013, s. 119-120, 163.)

3.1.3 Naiv Bayesiansk klassifikation

Detta är en klassifikationsmetod som bygger på Bayesiansk statistik. Som namnet antyder är det en metod med något starka antaganden, nämligen att samtliga element i en given vektor av observationer X_i är sinsemellan oberoende givet Y. Under detta antagande är det en beräkningsmässigt enkel sak att med hjälp av Bayes sats beräkna aposteriori-sannolikheter ur apriori-sannolikheter som skattats utifrån träningsdata. Enligt Bayes sats har vi nämligen att

$$\mathsf{P}(Y = 1 \mid X_1, ..., X_p) = \frac{\mathsf{P}(Y = 1)\mathsf{P}(X_1, ..., X_p \mid Y = 1)}{\sum_{k=0}^{1} \mathsf{P}(Y = k)\mathsf{P}(X_1, ..., X_p \mid Y = k)},$$

som under antagandet om oberoende blir:

$$\mathsf{P}(Y = 1 \mid X_1, ..., X_p) = \frac{\mathsf{P}(Y = 1) \prod_{j=1}^{p} \mathsf{P}(X_j \mid Y = 1)}{\sum_{k=0}^{1} \mathsf{P}(Y = k) \prod_{j=1}^{p} \mathsf{P}(X_j \mid Y = k)}.$$

En enkel klassifikationsregel ges då av

$$\underset{k}{\operatorname{argmax}}\, \mathsf{P}(Y = k) \prod_{j=1}^{p} \mathsf{P}(X_j \mid Y = 1),$$

som i det binära fallet ($k = 1$ eller 0) kan reduceras till att undersöka om kvoten

$$\frac{\mathsf{P}(Y = 1)}{\mathsf{P}(Y = 0)} \prod_{j=1}^{p} \frac{\mathsf{P}(X_j \mid Y = 1)}{\mathsf{P}(X_j \mid Y = 0)}$$

är större eller mindre än 1.

Trots sina starka antaganden har denna metod visat sig vara förvånandsvärt effektiv på att klassificera inom många områden, inte minst textanalys (där den t.ex. tillämpas flitigt som spamfiltrering) och andra områden där antalet parametrar kan vara mycket stort.

(Zhang, 2004; Hastie et al., 2008, s. 210-211)

3.1.4 Klassifikationsträd

Klassifikationsträd är en samling icke-parametriska metoder som enligt olika algoritmer kan partitionera data i regioner med avseende på de förklarande variablerna, i syfte att hitta de strukturer som bäst kan klassificera responsen. Den bästa binära partitioneringen av en given datamängd antas vara den som kan bäst separera data i grupper så att de resulterande undergrupperna har tydliga majoriteter av en viss klass. Algoritmen kan sedan fortsätta rekursivt med att dela upp datan i ännu mindre partitioner, med syftet att en ökning i homogenitet sker i varje splittring i relation till den mängd som splittrats. Den resulterande binära splittringen av data kan visualiseras mycket tydligt som

förgreningar och noder i ett träd med hierarkisk struktur. Det är vanligt att man först låter en trädmodell växa sig stort och nästan överbestämt på datan, för att sedan beskära den till en lämplig nivå. I denna analys använder vi oss av programpaketet `rpart`, som tillämpar algoritmen CART inom vilken måttet Gini-index används för att mäta graden av förorening (impurity) i en eventuell splittring. Om vi i en nod m som representerar en region R_m innehållande N_m observationer låter

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

beteckna proportionen av observationer i m av klass k, kan vi klassifiera observationerna i nod m som tillhörande klassen $\text{argmax}_k \, \hat{p}_{mk}$. Motsvarande Gini-index för noden är $1 - \sum_k \hat{p}_{mk}^2$, som i det binära fallet blir lika med $2p(1-p)$, om p är proportionen av observationer i den ena klassen. När p är 0 eller 1 har vi fullständig homogenitet och indexet blir lika med 0; när p istället är 0.5 finns ingen tydlig majoritet i noden och indexet når sitt maximum.

(Maindonald and Braun, 2010, kap. 11; Hastie et al., 2008, kap. 9.2)

3.2 Tabeller, kurvor och mått

3.2.1 Klassifikationstabeller

När man har utvecklat en binär klassifikationsmodell är man intressad av att mäta hur väl modellen gör prediktioner om huruvida $y = 0$ eller $y = 1$ utifrån nya observationer. Om de sanna värdena är kända kan vi jämföra dessa med de motsvarande predikterade sannolikheterna $\hat{\pi}_i$ i en sk. *klassifikationstabell*. För att erhålla en tydlig binär respons för varje $\hat{\pi}_i$ låter vi prediktionen för observation i vara $\hat{y} = 1$ då $\hat{\pi}_i > \pi_0$ och $\hat{y} = 0$ då $\hat{\pi}_i \leq \pi_0$, för något tröskelvärde π_0. Vi kan då summera modellens prediktiva förmåga genom sannolikheterna

$$\begin{aligned} &\mathsf{P}(\hat{y} = 1 \mid y = 1) \text{ (sensitivitet)} \quad \text{och} \\ &\mathsf{P}(\hat{y} = 0 \mid y = 0) \text{ (specificitet).} \end{aligned}$$

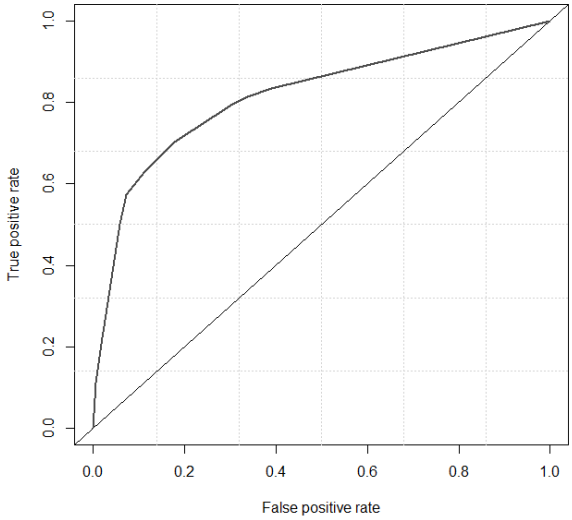
Detta förfaringssätt har dock sina begränsningar. Tröskelvärdet är godtyckligt och resultaten är känsliga för den relativa mängden gånger som y faktiskt är 1 eller 0.

(Agresti, 2013, s. 223; Hosmer and Lemeshow, 2013, s. 169-173)

3.2.2 ROC-kurvor

Med en *ROC-kurva* (Receiver Operating Characteristic curve) kan man istället summera modellens prediktiva förmåga för alla möjliga tröskelvärden mellan 0 och 1 och få en överblick över hur sensitivitet och specificitet varierar.

Sensitiviteten kallas även True Positive Rate och är alltså proportionen av gångerna som instanser av $y = 1$ har blivit korrekt klassificerade, medan $1-$ specificitet kallas False Positive Rate, och är proportionen av gångerna som



Figur 2: Exempel på en ROC-kurva.

en instans av $y = 0$ har blivit felaktigt klassificerade. Dessa värden plottar man mot varandra med varierande tröskel för att erhålla ROC-kurvan. Med ett tröskelvärde $\pi_0 = 0$ hade vi till exempel klassificerat samtliga responser som $y = 1$, och vi hade hamnat i punkten $(1, 1)$ på kurvan. Motsatsen, att klassificera samtliga responser som $y = 0$, ger punkten $(0, 0)$. Bäst prediktiv förmåga indikeras av att kurvan är så långt upp till vänster som möjligt, eftersom det är där både specificitet och sensitivitet är som högst.

Valet av tröskelvärde beror i hög grad på tillämpning. Om vi till exempel utför ett diagnostisk test på någon sjukdom vill vi förmodligen ha en relativt låg tröskel och hellre misstänka sjukdom, för att vara på den säkra sidan. Vid urval till högre utbildning kan man, beroende på prioriteringar, vara intresserad av att maximera specificitet till förmån för sensitivitet, eller försöka hitta någon bra balans mellan dem. En sådan kostnadssensitiv avvägning är någonting som det finns gott om teori kring, se särskilt Elkan (2001).

(Agresti, 2013, s. 223; Hosmer and Lemeshow, 2013, kap. 5.2.4)

3.2.3 AUC

Arean under ROC-kurvan (AUC) ligger någonstans mellan 0.5 och 1.0 och fungerar som ett sammanfattande mått över modellens förmåga att skilja mellan instanser av $y = 1$ och $y = 0$. En AUC på 0.5 indikerar att modellen inte kan diskriminera överhuvudtaget mellan fallen och gör rena gissningar, medan en area på 1.0 motsvarar perfekt diskriminering. Värdet kan tolkas som sannolikheten att en slumpmässigt vald instans av $y = 1$ rankas högre av modellen än en slumpmässigt vald instans av $y = 0$. Enligt Hosmer och Lemeshow (2013,

kap. 5.2.4) kan man som tumregel betrakta ett AUC mellan 0.7 och 0.8 som att modellen har en godtagbar förmåga att diskriminera, medan ett värde på 0.8 eller över kan betraktas som utmärkt.

3.2.4 Precision

Ett annat begrepp från ROC-analys som vi kommer att intressera oss av är *precisionen* (PPV, positive predictive value), vilket är andelen korrekt klassificerade instanser bland klassificeringarna $\hat{y} = 1$, dvs

$$PPV = \frac{TP}{TP + FP},$$

där TP är antalet korrekt klassificerade instanser (True Positives) och FP är antalet felaktigt klassificerade instanser (False Positives).

3.2.5 Brier score

Brier score är ett mått för det kvadratiska avståndet mellan prediktioner och observerade utfall för binära modeller och definieras som

$$BS = \frac{1}{N} \sum_{i=1}^{N} (\hat{\pi}_i - \hat{y}_i)^2. \tag{2}$$

Medan AUC mäter hur bra förmåga modellen har att kunna ursklija effektivt fångar Brier score även upp aspekter av kalibrering. Att en modell är välkalibrerad innebär att de predikterade sannolikheterna matchar fördelningen av de faktiska sannolikheterna väl. Ett Brier score på 0.25 pekar på dålig kalibrering (vilket kan inses genom att sätta $\hat{\pi}_i = 0.5$ i (2)), medan små värden så nära 0 som möjligt indikerar god kalibrering (Steyerberg et al., 2009).

4 Analys

Vi tillämpar nu de nämnda modellerna på datamaterialet och ställer upp ROC-kurvor för att jämföra deras prediktiva styrka med avseende på studieframgång. Därefter tittar vi närmare på strukturer i datan som utrönats av trädmodellerna. Till sist kommer vi att illustrera hur modellerna kan användas för att simulera en urvalsprocess på datamaterialet och mäta motsvarande förväntade genomströmning.

4.1 ROC-kurvor, AUC och Brier score

ROC-kurvor för de predikterande modellerna av studieframgången i de respektive kursmomententen och för de olika åldersgrupperna finns i Appendix som figurer A1-A6. Nedan redovisas även de nämnda modellernas mått på prediktionsstyrka.

Tabell 1: AUC-värden (*algebra*).

AUC	Linjär	Logistisk	Bayes	Träd
Alla	0.7697	0.7844	0.7655	**0.7905**
Yngre	0.8295	**0.8321**	0.8313	0.7963
Äldre	0.7402	**0.7416**	0.7248	0.7230

Tabell 2: AUC-värden (*analys*).

AUC	Linjär	Logistisk	Bayes	Träd
Alla	0.8108	**0.8113**	0.7939	0.7839
Yngre	0.8325	**0.8500**	0.8452	0.7987
Äldre	0.7848	0.7786	**0.7911**	0.7205

Tabell 3: Brier score (*algebra*).

Brier	Linjär	Logistisk	Bayes	Träd
Alla	0.1625	0.1587	0.1927	**0.1470**
Yngre	0.1540	**0.1476**	0.1778	0.1526
Äldre	0.2306	0.1563	0.1807	**0.1535**

Tabell 4: Brier score (*analys*).

Brier	Linjär	Logistisk	Bayes	Träd
Alla	0.1413	0.1360	0.1766	**0.1193**
Yngre	0.1468	0.1350	0.1603	**0.1290**
Äldre	0.1822	0.1272	0.1706	**0.1241**

En jämförelse av kurvorna och värdena i tabellen visar att modellerna liknar varandra ganska mycket med avseende på prediktionsförmåga, men att de logistiska modellerna verkar vara mer effektiva än de övriga. Samtidigt framstår trädmodellerna som de mest välkalibrerade. Modellerna är i allmänhet betydligt bättre på att förutsäga studieframgången för yngre elever jämfört med de äldre.

4.2 Träddiagram

I framtagandet av trädmodellerna hade vi ett visst mått av frihet. Klassifikationsträdsalgoritmernas styrka är att de är relativt simpla, icke-parametriska och kan upptäcka strukturer som är svåra att hitta med konventionella regressionsmetoder, men är i gengäld något okänsliga för nyanser i de specifika situationer som de tillämpas till. I utvecklingen av modellerna styrde vi över parametrarna `minsplit`, som bestämmer hur få observationer som får vara i en nod som splittras, samt `minbucket` som avgjorde minimumantalet för observationer i slutnoderna. De framtagna modellerna är kompromisser mellan överbestämda modeller med endast ett fåtal individer i slutnoderna och modeller med endast enstaka grenar.

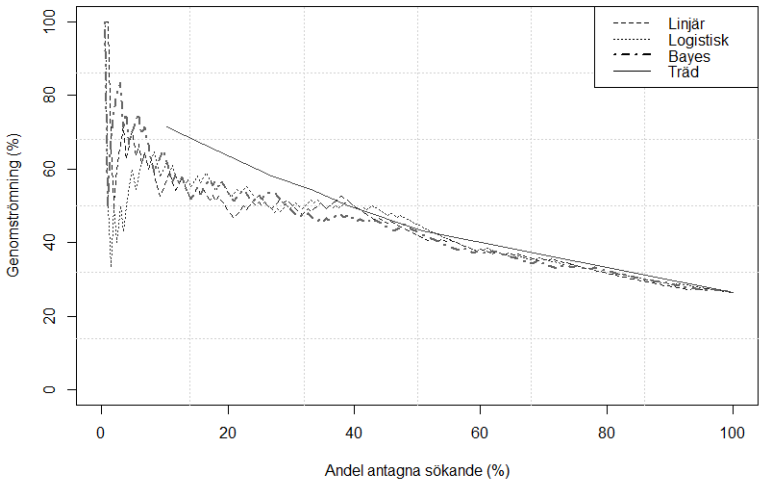
Träddiagramen visar hierarkiska strukturer och karaktäristiska drag i datan på ett mycket enkelt sätt. Dessa redovisas i Appendix som figurer A7-A12, och kan jämföras med resultaten av de logistiska modellerna (tabeller A3-A8). I varje nod visas en nolla eller en etta som indikerar huruvida majoriteten av individer i noden blivit godkända respektive icke godkända på kursen, och under noderna visas siffror på exakt hur många dessa är. Över varje förgrening står även ett påstående som t.ex. "mvg < 0.1", vilket är algoritmens urvalda kriterie för nästa splittring i datan, och de individer för vilka påståendet är sant flyttar ned till nästa vänstra nod, övriga till höger, och så fortsätter trädet vidare.

Det är tydligt utifrån dessa diagram att betygsmedelvärde och tidigare matematikbetyg är särskilt viktiga faktorer för genomströmningen i utbildningen. Dessa variabler är även signifikanta i de logistiska modellerna (dock ej för äldre studenter). Förutom detta har träddiagramen mest hittat partitioneringar med avseende på födelseår. I somliga träd framstår även antalet år sedan senaste matematikpoäng som en viktig faktor. För hela campus (figur A10) gäller allmänt att studenter med kort studieuppehåll klarar sig bättre, medan motsatsen ser ut att gälla om man ser på endast de äldre studenterna (figur A9).

4.3 Genomströmning

Låt oss nu leka med tankeexperimentet att vi tillämpar några av klassificeringsmodellerna som urvalsmetod för nya sökande till Matematik I. Under antagandet att de vid tillfället sökande till kursen råkar likna studenterna HT13 i sin sannolikhetsfördelning kan vi då simulera den förväntade genomströmningen av en antagningsprocedur under olika söktryck:

Låt oss säga att vi har N sökande individer till kursmomentet algebra och att vi ur dessa individer måste göra ett urval av precis n stycken. Då har vi alltså klassificerat totalt n av N individer som antagna, i tron att de är de mest lämpade till kursen. Precisionen, dvs. den predikterade genomströmningen av detta urval borde då under antagandet vara densamma som precisionen som motsvaras av exakt det tröskelvärde för vilket andelen positivt klassificerade individer är n/N. Vi kan plotta modellernas precisionsvärden mot denna andel för samtliga tröskelvärden för att få en bild av hur genomströmningen hade sett ut för olika andelar antagna (se figur 3).



Figur 3: Förväntad genomströmning då olika andelar av populationen blir urvalda till *algebra*.

I fallet $n = N$ måste modellerna klassificera samtliga individer som antagna, och genomströmningen hamnar givetvis på den verkliga nivån, ca 26%.

Kurvorna borde givetvis realistiskt sett vara helt avtagande; att de inte är det torde härröra från slumpfel i prediktionerna.

4.4 Jämförelse med nuvarande urvalsmodell

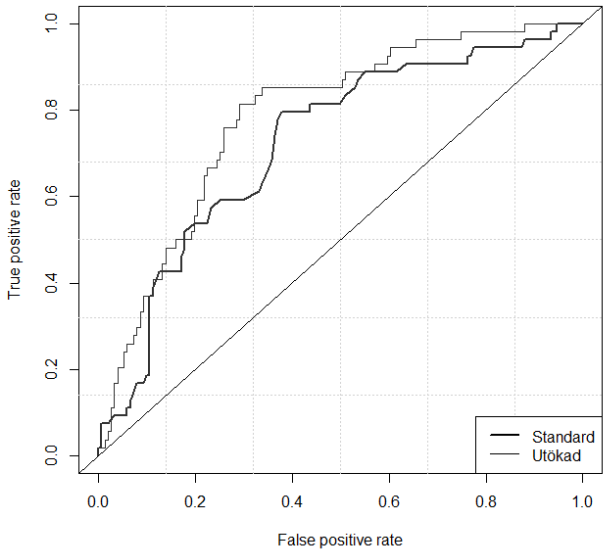
En naturlig fråga är hur pass bra våra framtagna modeller framstår i kontrast till den nuvarande allmänt tillämpade urvalsmodellen. Som vi redan nämnt har vi för lite data för att säga någonting om högskoleprovet, men vi kan jämföra med varianten där enbart snittbetyget används som urvalskriterie. För detta ändamål tar vi fram sådana modeller med hjälp av logistisk regression och jämför med de tidigare framtagna logistiska modellerna. Vi benämner dessa som ”utökade” modeller, medan vi kallar den allmänt tillämpade varianten med enbart snittbetyg för ”standard”.

Nedan redovisas en jämförelse i modellernas AUC, och vi visar även ROC- och genomströmningskurvor för fallet algebra (hela campus).

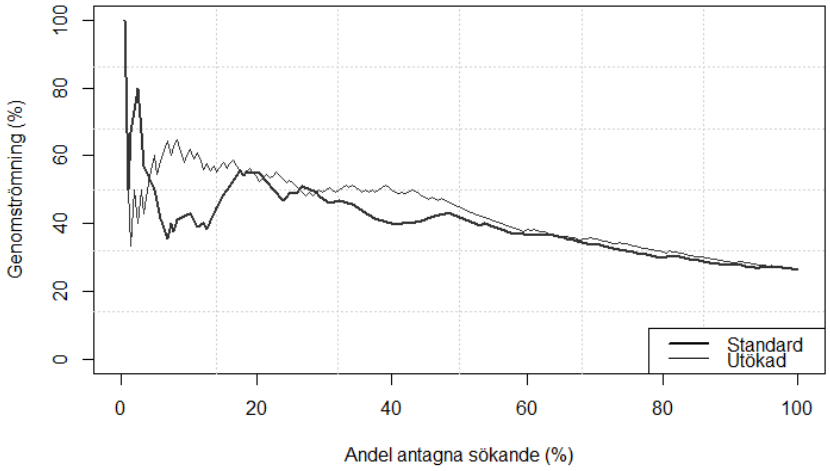
Tabell 5: Jämförelse av AUC, *algebra* (vänster) och *analys* (höger).

AUC	Standard	Skillnad (%)
Alla	0.7236	7.75
Yngre	0.7536	9.43
Äldre	0.6758	8.88

AUC	Standard	Skillnad (%)
Alla	0.7803	3.83
Yngre	0.8011	5.75
Äldre	0.7420	4.70



Figur 4: Jämförelse i ROC-kurva mellan framtagna (”utökade”) prediktionsmodeller och den normalt tillämpade urvalsmodellen, *algebra* (hela campus).



Figur 5: Jämförelse i genomströmning mellan framtagna (”utökade”) prediktionsmodeller och den normalt tillämpade urvalsmodellen, *algebra* (hela campus).

5 Diskussion

Vi har nu tillämpat några klassifikationsmodeller på genomströmningen i Matematik I och illustrerat/visualiserat några intressanta aspekter av dem. Vi har sett att vi med de givna variablerna faktiskt kan prediktera framgången i Matematik I på en godtagbar nivå. Vi har sett att logistisk regression och trädmodeller framstår som de mest lämpliga metoderna för detta ändamål, samt att klassifikationsträden även kunnat belysa intressanta underliggande hierarkiska strukturer i datan. Prediktionsförmågan är särskilt god sett till enbart yngre studenter, medan den är svagare för de äldre.

Vi har även jämfört våra modeller med den allmänt tillämpade urvalsmodellen, där man använder enbart medelbetyg från gymnasiet för att avgöra sannolikheten för studieframgång. Denna metod framstod också som godtagbar i sin prediktionsförmåga med AUC-värden kring 0.75, med relativt små skillnader i AUC när de övriga variablerna läggs till. Gymnasiesnittet framstod som ett särskilt bra urvalskriterie för momentet *analys*.

Det bör nämnas att en predikterande modell alltid behöver testas på ny data (som inte använts i själva modellanpassningen) innan man riktigt kan säga huruvida den klassificerar effektivt eller bara är överanpassad. Vår bedömning var att underlaget på ca 200 studenter var något för knapp för korsvalidering genom indelning i partitioner av träning- och testdata, och vi använde istället hela datamängden i anpassningen av modellerna. Förmodligen är dessa därför av begränsat praktiskt värde, men ger förhoppningsvis om inte annat en överskådlig bild och illustration av problematiken kring urval och studieunderlag till kursen.

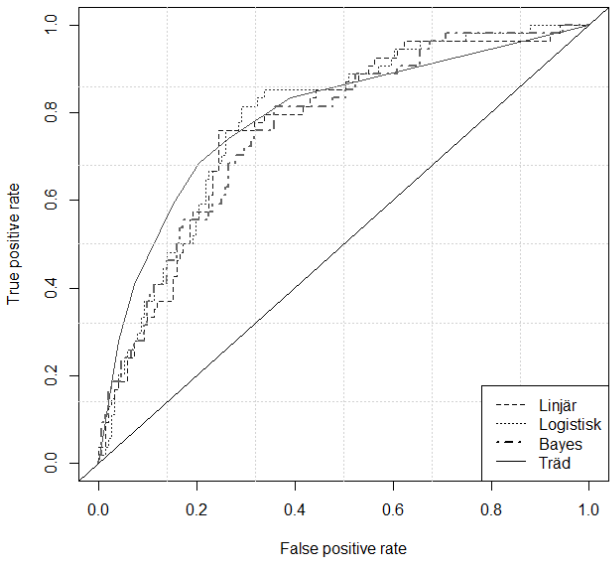
Appendix

Tabell A1: Andelen godkända nyregistrerade campusstudenter (%) i respektive kursmoment, samt andelen godkända i båda moment bland dem som var helfartsstudenter.
Med yngre studenter avses dem som är födda efter 1991.

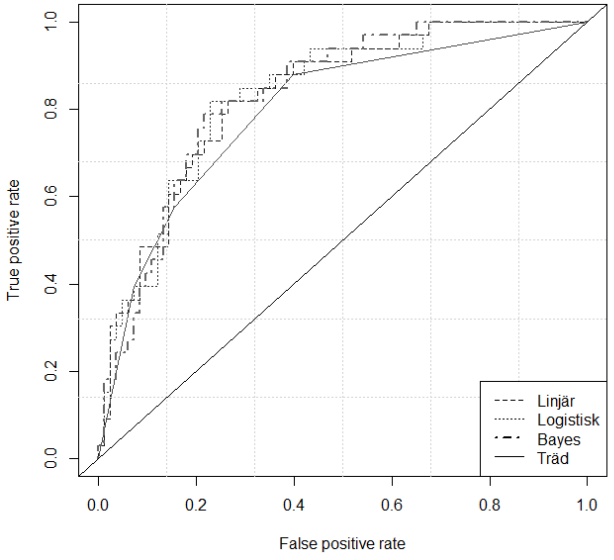
	Algebra	Analys	Båda
Campus	26	24	20
Yngre	29	27	25
Äldre	23	17	15

Tabell A2: Fördelningen av Matematik I-studenter på olika kandidatprogram (nyregistrerade campus-studenter).

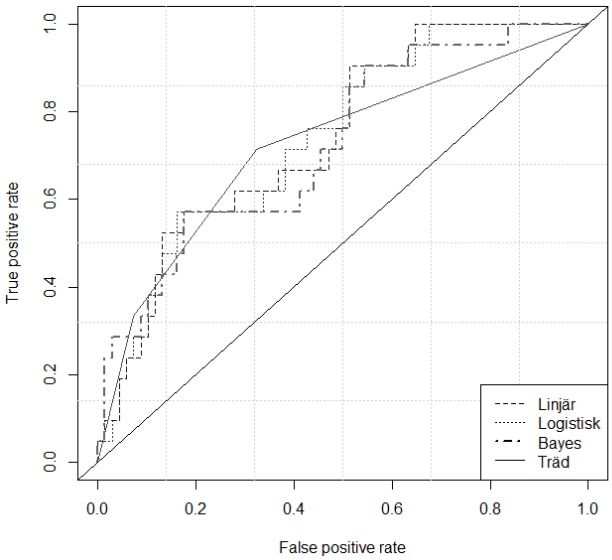
Program	Andel (%)
Fysik	13.56
Sjukhusfysik	4.24
Biofysik	0.85
Astronomi	12.29
Meteorologi	0.85
Matematik	25.42
Biomatematik	2.12
Matematik-filosofi	4.24
Matematik-ekonomi	0.42
Datalogi	9.75
Lärarprogram	0.42
Fristående	25.85



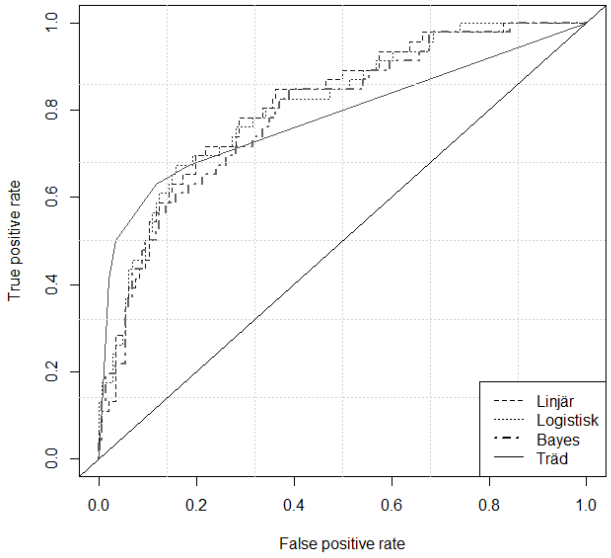
Figur A1: ROC-kurvor, *algebra* (hela campus).



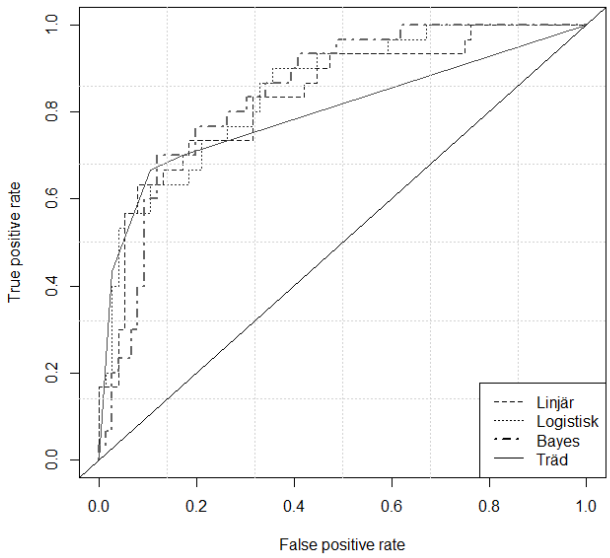
Figur A2: ROC-kurvor, *algebra* (yngre studenter).



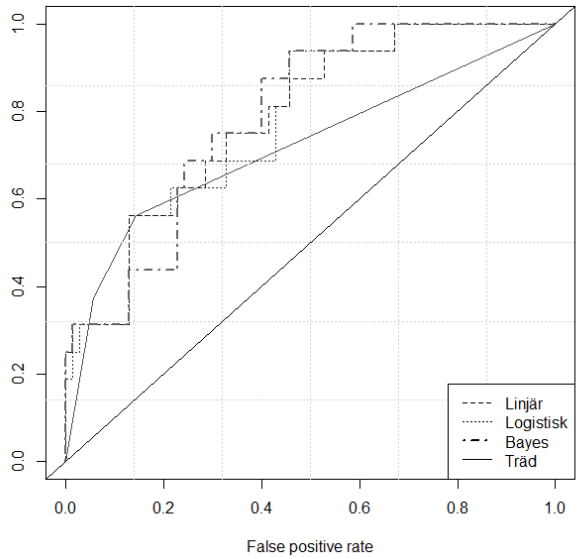
Figur A3: ROC-kurvor, *algebra* (äldre studenter).



Figur A4: ROC-kurvor, *analys* (hela campus).



Figur A5: ROC-kurvor, *analys* (yngre studenter).



Figur A6: ROC-kurvor, *analys* (äldre studenter).

Tabell A3: Logistisk regression av *algebra*-studenter.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.5855	1.3451	-3.41	0.0007	***
Kön	-0.4055	0.4427	-0.92	0.3597	
Födelseår	-0.0297	0.0624	-0.48	0.6344	
Betygssnitt	0.1438	0.0866	1.66	0.0966	.
Andel MVG	1.3479	0.5629	2.39	0.0166	*
Matte E	0.7996	0.4193	1.91	0.0565	.
Förberedande kurs	1.0187	0.5763	1.77	0.0771	.
Högskoleprov 1.5	-0.0419	0.4978	-0.08	0.9329	
Studieuppehåll matematik	0.0322	0.0845	0.38	0.7033	

Tabell A4: Logistisk regression av *algebra*-studenter (yngre).

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.3205	2.3187	-2.73	0.0064	**
Kön	0.1433	0.5812	0.25	0.8053	
Födelseår	-0.0046	0.4165	-0.01	0.9912	
Betygssnitt	0.1996	0.1255	1.59	0.1116	
Andel MVG	1.4453	0.7312	1.98	0.0481	*
Matte E	1.5406	0.7956	1.94	0.0528	.
Förberedande kurs	2.4569	1.0572	2.32	0.0201	*
Högskoleprov 1.5	-0.5287	0.7685	-0.69	0.4915	
Studieuppehåll matematik	-0.1805	0.5687	-0.32	0.7510	

Tabell A5: Logistisk regression av *algebra*-studenter (äldre).

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.8687	1.9290	-2.01	0.0449	*
Kön	-0.8370	0.7688	-1.09	0.2763	
Födelseår	-0.0678	0.0775	-0.87	0.3817	
Betygssnitt	0.1179	0.1276	0.92	0.3556	
Andel MVG	1.1979	0.9363	1.28	0.2007	
Matte E	0.2071	0.5917	0.35	0.7263	
Förberedande kurs	0.3223	0.8100	0.40	0.6907	
Högskoleprov 1.5	0.1306	0.7191	0.18	0.8558	
Studieuppehåll matematik	0.0289	0.0912	0.32	0.7510	

Tabell A6: Logistisk regression av *analys*-studenter.

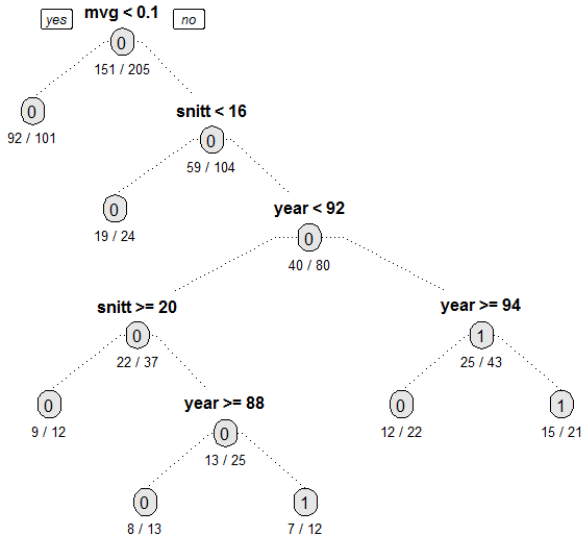
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.4870	1.5455	-3.55	0.0004	***
Kön	0.6726	0.4422	1.52	0.1283	
Födelseår	0.0150	0.0774	0.19	0.8465	
Betygssnitt	0.1848	0.0975	1.89	0.0582	.
Andel MVG	1.7547	0.6182	2.84	0.0045	**
Matte E	0.4291	0.4598	0.93	0.3506	
Förberedande kurs	0.4557	0.6401	0.71	0.4766	
Högskoleprov 1.5	0.0563	0.5282	0.11	0.9151	
Studieuppehåll matematik	-0.0204	0.1032	-0.20	0.8432	

Tabell A7: Logistisk regression av *analys*-studenter (yngre).

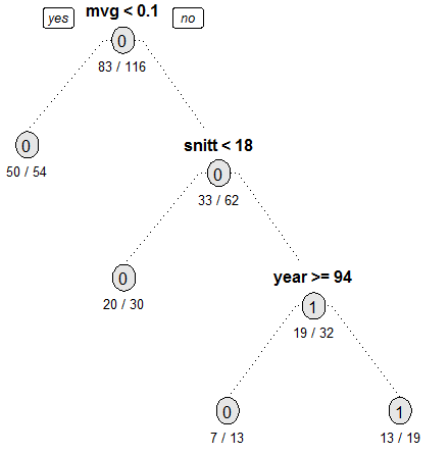
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.8993	2.7216	-2.90	0.0037	**
Kön	1.4760	0.6546	2.25	0.0242	*
Födelseår	0.0903	0.4461	0.20	0.8395	
Betygssnitt	0.2947	0.1430	2.06	0.0394	*
Andel MVG	1.7151	0.8065	2.13	0.0335	*
Matte E	0.8149	0.8022	1.02	0.3097	
Förberedande kurs	0.9872	0.9893	1.00	0.3183	
Högskoleprov 1.5	-0.2394	0.8370	-0.29	0.7749	
Studieuppehåll matematik	-0.6828	0.6569	-1.04	0.2986	

Tabell A8: Logistisk regression av *analys*-studenter (äldre).

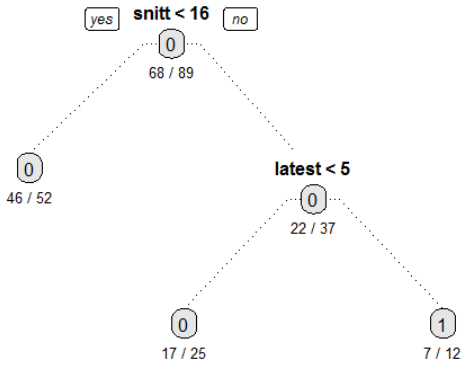
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.3848	2.2234	-1.97	0.0486	*
Kön	0.1502	0.6974	0.22	0.8295	
Födelseår	-0.0098	0.0908	-0.11	0.9138	
Betygssnitt	0.1287	0.1439	0.89	0.3714	
Andel MVG	1.6376	1.0145	1.61	0.1065	
Matte E	0.2377	0.6376	0.37	0.7092	
Förberedande kurs	0.1452	0.9108	0.16	0.8733	
Högskoleprov 1.5	0.2581	0.7388	0.35	0.7268	
Studieuppehåll matematik	-0.0140	0.1070	-0.13	0.8957	



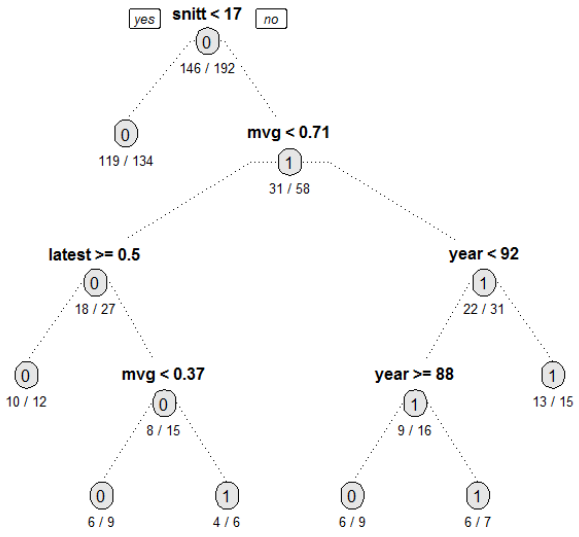
Figur A7: Träddiagram, *algebra* (hela campus).



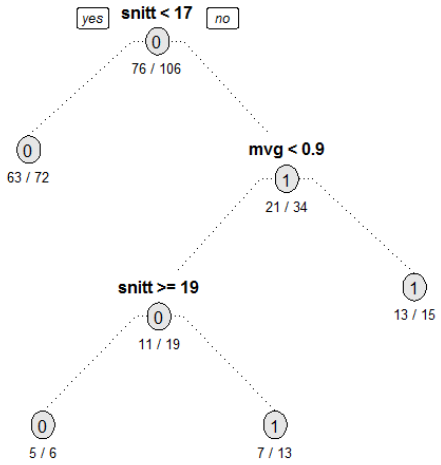
Figur A8: Träddiagram, *algebra* (yngre studenter).



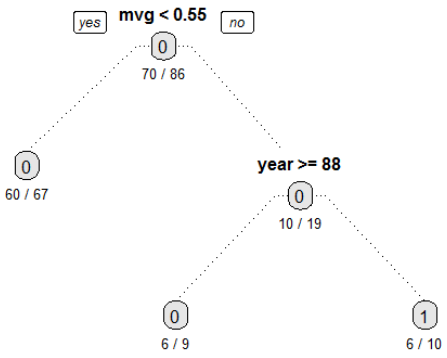
Figur A9: Träddiagram, *algebra* (äldre studenter).



Figur A10: Träddiagram, *analys* (hela campus).



Figur A11: Träddiagram, *analys* (yngre studenter).



Figur A12: Träddiagram, *analys* (äldre studenter).

Referenser

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons, 3rd edition.

Alkhasawneh, R. (2011). *Developing a Hybrid Model to Predict Student First Year Retention and Academic Success in STEM Disciplines Using Neural Networks*. VCU Theses and Dissertations, Paper 2570.

Chong, H. Y., DiGangi, S., Jannasch-Pennell, A., and Kaprolet, C. (2010). *A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year*. Journal of Data Science, 8, 2010.

Dryler, H. (2013). *Social bakgrund och genomströmning i högskolan - En studie av långa och medellånga yrkesexamenprogram*. UK-ämbetet, Rapport 2013:4.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978.

Eriksson, L. (2010). *Orsaker till studieavbrott*. Högskoleverkets rapportserie 2010:23 R.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, 2nd edition.

Herzog, S. (2006). *Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression*. New Directions for Institutional Research, 2006(131), 17-33.

Hosmer, D. W. and Lemeshow, S. (2013). *Applied Logistic Regression*. John Wiley & Sons, 3rd edition.

Maindonald, J. and Braun, W. J. (2010). *Data Analysis and Graphics Using R*. Cambridge, 3rd edition.

Pugh, C. M. and Lowther, S. (2004). *College Math Performance and Last High School Math Course*. Annual conference of the southern Association for Institutional Research, Biloxi, Mississippi.

Steyerberg, E. W., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. (2009). *Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures*. Epidemiology (Cambridge, Mass.).

Wikström, C. and Wikström, M. (2012). *Urval till högre utbildning - Påverkas betygens prediktionsvärde av ålder?* IFAU, Rapport 2012:21.

Zhang, H. (2004). *The Optimality of Naive Bayes*. Proceedings of the 17th FLAIRS Conference, The AAAI Press.