Trädbaserade maskininlärningsmetoder för analys av neuroendokrin cancerdata

Sebastian Uggla

Kandidatavhandling i datavetenskap

Handledare: Luigia Petre och Ion Petre i samarbete med

Valeriy Paramonov och Cecilia Sahlgren på The Cell Fate Group,

Cellbiologi, Åbo Akademi

Fakulteten för naturvetenskaper och teknik

Åbo Akademi

Våren 2019

Referat

Den mycket komplexa biologiska data som samlas i samband med biologisk forskning och medicinska undersökningar är ofta svåra att analysera. Maskininlärningsmetoder visar lovande resultat i analysen av denna typ av data och i denna avhandling beskrivs de vanligaste trädbaserade maskininlärningsmetoderna i detalj samt för och nackdelar med dessa. I analysdelen av denna avhandling används även den trädbaserade maskininlärningsmodellen slumpmässiga skogar för att analysera resultaten från histokemiska tester som utförts på neuroendokrina tumörer och kliniska data på de patienter som lidit av dessa tumörer. Resultaten från analysen jämförs med redan vedertagna biologiska fakta för att bevisa modellens effektivitet på denna typ av problem. Modellen synliggjorde en redan känd korrelation mellan två variabler och hittade ett par potentiellt intressanta korrelationer.

Innehåll

1. Inledning

I och med att att ny medicinsk teknologi har utvecklats har stora mängder cancerdata samlats in och gjorts tillgängligt för forskare [1]. Dessa data är mycket varierande och innefattar patientens kliniska data (blodtest, frågeformulär, kön), data från bilddiagnostik (tumörens storlek och tumörens volym), vårddata (kemoterapi, strålningsbehandling) och molekylära data (genetiska data, histokemiska test) [2]. För att utnyttja data inom läkemedelsutveckling och cancervård gäller det att analysera och tolka dem. Maskininlärningsmetoder har blivit populära inom cancerforskning för att man med hjälp av dem kan hitta mönster och korrelationer inom den mycket komplexa mängd data som samlats inom detta område [2]. Maskininlärning (ML) är en gren av artificiell intelligens [2] vars mål är att bygga datorsystem som automatiskt förbättrar sig själva och som lär sig av erfarenhet [3]. Tom Mitchell skriver i sin artikel The Discipline of Machine Learning, översatt till svenska, vi säger att en maskin lär sig med avseende på en viss uppgift T, prestandamått P, och erfarenhetstyp E, ifall systemet tillförlitligt förbättrar sin prestanda P för uppgift T, efter erfarenhet E. [3].

Med en ML-modell vill man alltså förutsäga en viss variabel, t.ex. överlevnadschansen för en cancerpatient genom att träna modellen på data från tidigare patienter. Modellen kan sedan användas för att hjälpa läkare att fatta vårdbeslut och dessutom kan modellen inspekteras för att urskilja vilka variabler som inverkar mest på överlevnadschansen och på detta sätt ge nya biologiska eller vårdrelaterade insikter.

Det finns två huvudtyper av maskininlärning, övervakat lärande och oövervakat lärande. Inom övervakat lärande finns det en viss variabel som modellen lär sig att förutspå, t.ex. överlevnadschans för cancerpatienter, medan det i oövervakat lärande inte finns någon sådan variabel utan modellen måste själv försöka hitta mönster och likheter mellan olika variabler i data [1]. Övervakat lärande kan i sin tur delas in i ytterligare två klasser, klassificering och regression. I klassificering är variabeln man

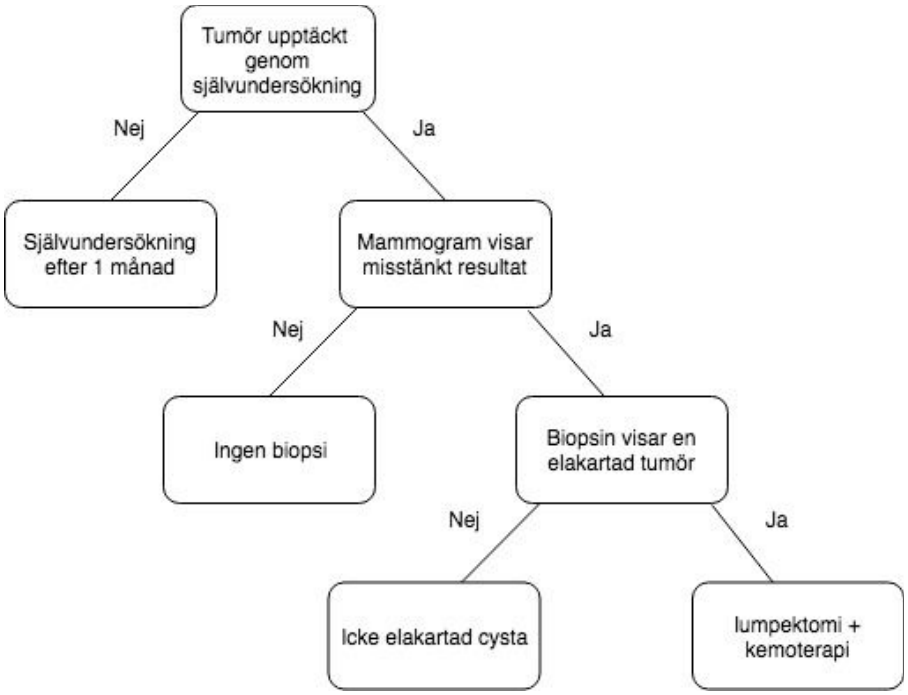
vill förutspå en konkret klass, exempelvis man eller kvinna, medan den i regression är en kontinuerlig variabel, exempelvis ålder [1].

I denna avhandling kommer jag att presentera ett par trädbaserade ML-modeller, beslutsträd och slumpmässiga skogar, som används inom cancerforskning. Jag kommer även att demonstrera slumpmässiga skogars användning på cancerdata för att hitta hittills okända korrelationer mellan variabler som kan användas som grund för fortsatt forskning. Modellens effektivitet kommer även att bevisas genom att hitta redan biologiskt vedertagna korrelationer i cancerdatan.

2. Beslutsträd

Beslutsträd utvecklades inom marknadsföring som ett verktyg för att segmentera potentiella kunder i målgrupper. Inom hälsovården kan de användas för att segmentera patienter i väldefinierade grupper vars medlemmar har sinsemellan liknande egenskaper [4].

Segmenteringen av patienter i subgrupper kan ge insikter om vilka egenskaper som skiljer dessa grupper åt samt göra vården av nya patienter mera målinriktad genom att man kan fatta vårdbeslut baserade på vilken subgrupp patienten placeras i. Ett exempel på detta ges i figur 1, när vårdbeslutet för en patient skall göras börjar man i början av figurens träd och ställer sig frågan i rutan, baserat på svaret fortsätter man nedåt i trädet tills man landar i ett vårdbeslut.

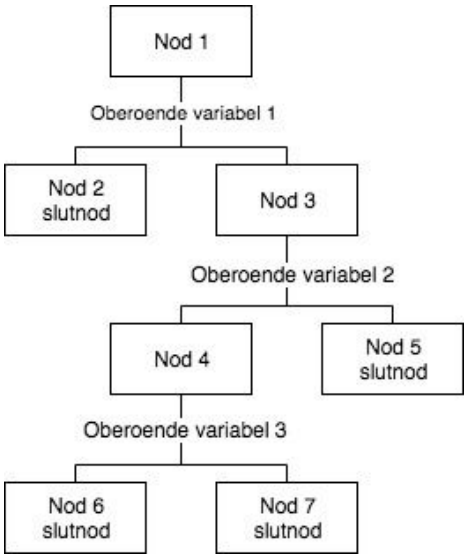


Figur 1. Exempel på beslutsträd som möjliggör vårdbeslut [5].

2.1 CART algoritmen

CART (Classification and regression tree) är en statistisk metod som använder beslutsträd för att lösa både klassificerings- och regressionsproblem. Den är icke-parametrisk vilket betyder att den inte gör statistiska antaganden om data. Med hjälp av CART bygger man ett beslutsträd som beskriver en beroende variabel, en variabel vars värde bestäms av andra variabler, som en funktion av ett flertal oberoende variabler, de variabler som tillsammans bestämmer den beroende variabelns värde. När den beroende variabeln är kategorisk bygges ett klassificeringsträd, medan ett regressionsträd byggs upp om den beroende variabeln är kontinuerlig [6]. Att en variabel är kategorisk innebär att de värden den antar faller inom diskreta klasser såsom ja/nej eller katt/hund, det finns inga mellanting hos diskreta variabler. Att en variabel är kontinuerlig innebär att den antar värden som är kontinuerliga såsom ålder eller huspris.

Vid uppbyggandet av ett beslutsträd börjar CART algoritmen med en nod som innehåller alla data, nod 1 i figur 2 [7]. Sedan undersöks alla variabler i noden för att hitta den oberoende variabel som delar in data i två sinsemellan så olika grupper som möjligt med avseende på den beroende variabeln. Noden förgrenas sedan in i två barn-noder, nod 2 och nod 3 i figuren, beroende på vilken variabel som valdes. Denna process upprepas rekursivt på vardera barn-nod, så att ett träd skapas och fortsätter tills varje barn-nod innehåller endast en instans av data, eller tills varje element i barn-noden har exakt samma värde på den beroende variabeln [4]. Dessa noder kallas för slutnoder och representeras av nod 2,5,6 och 7 i figuren. Man kan även stoppa delningen i ett tidigare skede genom att tillämpa någon regel, exempelvis att varje nod skall ha åtminstone ett visst minimiantal instanser i sig.



Figur 2. Exempel på uppbyggande av beslutsträd [7]

För att dela upp data i föräldra-noden i två så olika barn-noder som möjligt så används olika sorters delningskriterium. Gemensamt för alla delningskriterier är att de börjar med att definiera en nods orenhet, mängden variabilitet för den beroende variabeln hos de olika instanserna i noden. En nod utan orenhet skulle alltså inte heller ha någon variabilitet för den beroende variabeln [4]. Delningskriteriet är baserat på funktionen $\theta(s,\ t)$ som bestämmer hur bra en viss delning är, denna funktion definieras som:

$$\theta(s,\ t) = \phi(p)\ - PL * \phi(pL) - PR * \phi(pR)\ ,\ \textit{Ekvation 1}\ [8]$$

Där s är delningsvärdet som undersöks och det s som maximerar funktionens värde väljs. Värdet t är föräldra-noden, p är data i föräldra-noden, pL är data i den vänstra barn-noden och pR data i den högra barn-noden, PL och PR är proportionen av föräldra-nodens data som finns i respektive barn-nod, $\phi(p)$ är funktionen för en nods orenhet, varav den vanligast använda funktionen för detta ändamål är gini:

$$\phi(p) = \sum_j pj(1 - pj) = 1 - \sum_j pj^2$$

Ekvation 2 [6, 8].

Gini mäter hur ofta en slumpmässigt vald instans i nodens data skulle bli felaktigt klassificerad ifall den klassificerades enligt den beroende variabelns fördelning i nodens data. För regressionsträd mäts ofta orenhet med summan av avvikelserna från medelvärdet för den beroende variabeln för de olika instanserna i nodens data, så kallade MSE eller kvadratiska medelvärdet (eng. mean squared error).

$$\phi(p) = \sum_j (pj - \bar{p})^2$$

Ekvation 3 [6]

där $\bar{p}$ är medelvärdet för den beroende variabeln.

2.2 Styrkor och svagheter hos beslutsträdsmetoder

Beslutsträd är bättre än klassiska statistiska metoder på att upptäcka interaktioner och icke-linjäritet i data med många olika oberoende variabler [9]. De är lätta att förstå och lätta att tolka [10], jämfört med mera "ogenomskinliga" metoder som exempelvis neurala nätverk. Mjukvara för CART analys är även lätt att använda för personer utan en bakgrund i statistik [4]. Beslutsträdsmetoder har dock en del begränsningar jämfört med andra ML modeller. Eftersom beslutsträd gör så få antaganden om data de tränas på och således har en stor grad av frihet så finns det risk för att de överpassar på träningsdatan och inte klarar av att generalisera på nya instanser [10]. Överpassning (eng. overfitting) är ett centralt begrepp inom maskininlärning och ett av de största problemen för de flesta ML metoder. Överpassning innebär att en ML modell klarar av att förutspå instanser i träningsdata bättre än nya instanser som modellen inte

tränats på. Det finns metoder för att motverka överpassning för beslutsträd såsom att begränsa trädets djup och kräva en minimistorlek för noder. Beslutsträd är även känsliga för små variationer i träningsdata vilket betyder att en liten ändring i träningsdata kan leda till mycket annorlunda träd, denna instabilitet kan motverkas med ensemble metoder såsom slumpmässiga skogar (eng. Random Forests) som jämnar ut förutsägelser över många träd [10].

3. Ensemble lärande och slumpmässiga skogar

Ensemble lärande grundar sig på utnyttjandet av gruppens visdom, ifall man ställer en fråga till en stor grupp med människor så kommer medeltalet av deras svar ofta att vara mer korrekt än det mest korrekta enskilda svaret. Detta fenomen beror på att de kumulativa fel som individuella personers överskattningar och underskattningar ger upphov till närmar sig noll ifall gruppen är tillräckligt stor. Fenomenet gäller även för prediktorer såsom klassificeringsträd eller regressionsträd, medeltalet av flera prediktorers svar är ofta mera korrekt än den mest korrekta enskilda prediktorn. En grupp av prediktorer kallas för en ensemble och deras gemensamma förutsägelse/prediktion kallas för ensemble lärande [10].

En ensemble av beslutsträd där varje träd tränas på en delmängd av träningsdata och varje delning av noder i träden görs på en slumpmässigt vald delmängd av variablerna kallas för en slumpmässig skog (eng. Random Forest, här kallad SS) [10]. SS fungerar bra på data med få instanser men flera variabler [11]. Inom biovetenskaper, inklusive cancerforskning, är dessa sorters dataset vanliga på grund av svårigheten att hitta ett stort antal patienter med en specifik åkomma, från en specifik patient kan dock en stor mängd datapunkter/variabler samlas. SS kan liksom sina beståndsdelar, beslutsträd, även hantera korrelationer och interaktioner mellan variabler i data [11]. SS kan hantera dessa korrelationer i datat eftersom beslutsträden alltid väljer exakt en variabel att göra delningsbeslut på. En egenskap hos SS är lättheten med vilken de olika oberoende variablerna kan rangordnas baserat på hur mycket de påverkar den beroende variabeln. Denna egenskap kan ge biologiska insikter samt ytterligare förbättra SS modellens prediktioner genom att man kan ta bort variabler utan betydelse och därmed görs datat klarare [12].

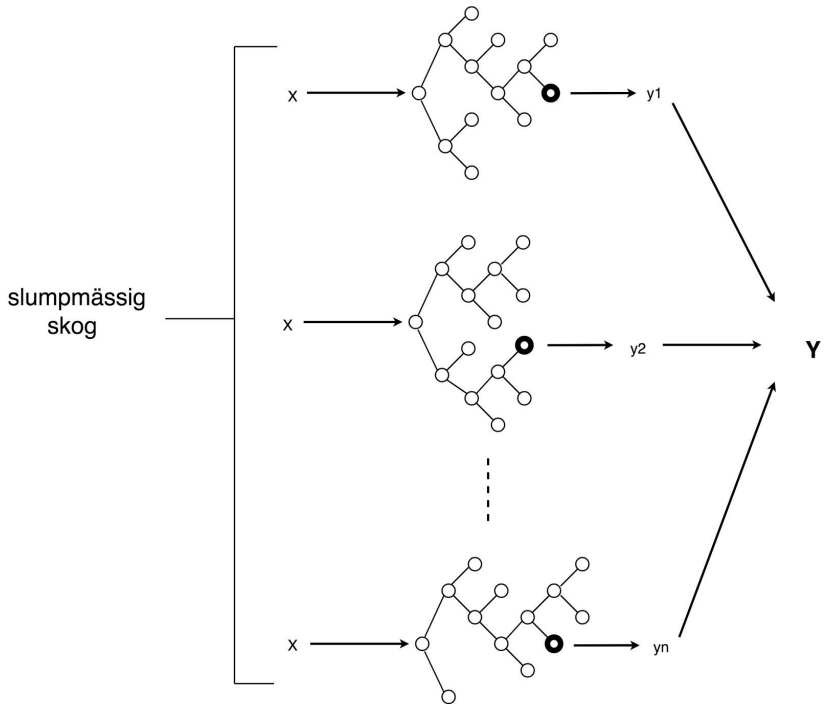
Konstruerandet av en SS modell kan sammanfattas i följande steg:

1) Bilda n stycken delmängder, så kallade “bags”, av storlek k av ursprungliga data. För bildandet av varje delmängd, välj slumpmässigt en instans från data och lägg till den i delmängden. Sätt sedan tillbaka instansen i data och upprepa denna procedur tills delmängden har storlek k. På detta sätt kommer varje delmängd att innehålla flera kopior av vissa instanser och sakna andra instanser från data helt, detta kallas bootstrapping.

2) Bygg ett beslutsträd för varje delmängd av data. För varje nod i varje träd, välj slumpmässigt m variabler att göra delningsbeslut på.

3) Samla ihop prediktionerna för varje träd, i figur 3 y1, y2, yn, från varje beslutsträd. Ifall den beroende variabeln är kontinuerlig, ta medeltalet av förutsägelserna. Ifall den är diskret, låt den klass som majoriteten av träden röstat på vara SS förutsägelse, i bilden Y. Denna förutsägelse kan sedan användas vid, exempelvis, fattandet av vårdbeslut.

4) Beräkna modellens noggrannhet genom att, för varje träd, beräkna hur noggrant trädet kan förutsäga datan som inte inkluderas i den delmängd som trädet tränades på. Dessa data kallas “OOB, out of bag” data [11].



Figur 3. Exempel på hur en slumpmässig skog gör en prediktion [13].

En SS byggs i allmänhet upp av en stor mängd träd, ofta 500 till 2000 stycken [9]. Vart och ett av dessa beslutsträd är enskilt en svagare prediktor än ett beslutsträd som byggts upp av all data på grund av att varje enskilt beslutsträd endast konstruerats av en delmängd av all data, tillsammans är de dock en stark prediktor [12]. SS ses ofta som en så kallad "svart låda" (eng. "black box") modell, alltså en modell som är svår att tolka och inspektera. SS är definitivt mer oklar än ett normalt beslutsträd då det inte är rimligt att undersöka tusentals beslutsträd individuellt för att komma fram till varför modellen gjorde en förutsägelse på ett specifikt sätt. Det finns dock ett antal metoder som kan ge insikter i hur SS gör beslut på ett visst dataset samt hur olika variabler är korrelerade med varandra.

3.1 Rangordning av variabler

Rangordning av de oberoende variablerna i ett dataset kan vara användbart för att tolka data och se vilka variabler som är relevanta. Viktiga biomarkörer kan exempelvis identifieras genom rangordning och icke-relevanta variabler kan elimineras från datamängden [12]. Det vanligaste måttet att rangordna variablerna i SS är enligt permutations betydelse (eng. permutation importance eller Breiman-Cutler importance) [11]. För att rangordna variablerna enligt permutations betydelse beräknas, för varje träd i SS, felen i prediktioner på OOB datan. Sedan permuteras (=ordnas slumpmässigt) den variabel man undersöker i OOB data för varje träd och felen i prediktioner på OOB data beräknas igen. Medeltalet på skillnaden av prediktionerna före och efter permuteringen av OOB data för varje träd är variabelns betydelse. Denna beräkning görs sedan för varje variabel i data så att de olika variablerna kan rangordnas [11]. Man undersöker alltså hur mycket sämre modellen blir på att förutsäga den beroende variabeln om varje oberoende variabel sätts ur spel, en åt gången.

3.2 Partiellt beroende

Visualisering är ett av de mest kraftfulla verktygen för att tolka en ML modell. Data med flera oberoende variabler är dock svåra att visualisera i två dimensioner. Att visualisera den partiella betydelsen för en eller ett fåtal variabler på den beroende variabeln är därför användbart [14]. En PDP (Partial Dependence Plot, swe. Partiellt Beroende Diagram) visar den marginella effekt en eller två oberoende variabler har på den beroende variabeln. Utifrån en PDP kan man se ifall förhållandet mellan en oberoende och en beroende variabel är linjärt, monotont eller icke-linjärt [15]. För regressionsproblem ges funktionen för partiellt beroende av:

$$\widehat{f}_{xs}(x_s) = \int \widehat{f}(x_s, x_c) dP(x_c)$$, *Ekvation 4* [15]

Där x_s är de oberoende variabler, ofta en eller två stycken, för vilka det partiella beroendet skall plottas och x_c är de övriga oberoende variablerna i modellen $\widehat{f}$ [15]. Varje delmängd av oberoende variabler s har en egen funktion för partiellt beroende $\widehat{f}_{xs}$ som ger medeltalet av funktionsvärdet $\widehat{f}$ när x_s är fixerat och x_c varierar över sin marginella fördelning. Eftersom varken $\widehat{f}$ eller $dP(x_c)$ är kända, uppskattas Ekvation 4 som:

$$\widehat{f}_{xs}(x_s) = \frac{1}{n} \sum_{i=1}^{n} \widehat{f}(x_s, x_c^{(i)})$$, *Ekvation 5* [15, 16].

Där $x_c^{(i)}$ är de olika värdena på x_c i data och n är antalet instanser i data. Ett antagande i PDP är att variablerna i x_s inte är korrelerade med variablerna i x_c. Ett problem med PDP är att den döljer heterogena detaljer i data eftersom den endast plottar den genomsnittliga marginella effekten. Ifall hälften av instanserna korrelerar starkt negativt och den andra hälften starkt positivt med den beroende variabeln så kan dessa förhållanden osynliggöras. För att lösa detta problem kan de olika instanserna i datan plottas åtskilda och då fås ett ICE (Individual Conditional Expectation) diagram [15]. I analysdelen kommer ett programvarubibliotek att användas som kombinerar PDP och ICE i ett diagram för att ge en helhetsbild av det partiella beroendet för en viss oberoende variabel på den beroende variabeln.

4. Analys av cancerdata

Metoden slumpmässiga skogar kommer att användas för att analysera data från patienter med endokrina tumörer i syfte att upptäcka mönster och korrelationer i data som kan vara till användning inom cancer forskning. Dessa data innefattar kliniska attribut från patienter såsom kön, ålder och lokalisering av metastaser samt resultat från histokemiska tester som gjorts på tumörer från patienterna i fråga. Först ges en kortfattad beskrivning av den biologiska bakgrunden till datan.

4.1 Endokrina tumörer

Tumörer som uppstår ur hormonproducerande celler i magen, tarmen eller bukspottkörteln är relativt sällsynta. Dessa så kallade neuroendokrina tumörer har en del kliniska egenskaper gemensamt och beter sig ofta biologiskt oförutsägbart och ovanligt [17]. Studier har påvisat att den genomsnittliga överlevnaden för patienter med metastaserad neuroendokrin tumör från tunntarmen är 56 månader medan den hos patienter med metastaserad tumör från bukspottskörteln är 24 månader [18]. Prevalensen av neuroendokrina tumörer har ökat under de senaste trettio åren antagligen på grund av förbättrad diagnostik, dock har ingen förbättring av överlevnad skett under samma tidsperiod [17]. Att belysa neuroendokrina tumörers biologiska funktioner är därför viktig för att förbättra överlevnadschansen för patienter som lider av dem.

En del upptäckter har gjorts gällande neuroendokrina tumörers biologi. Exempelvis somatostatin receptorerna SST 1-5 som har en hämmande effekt på neuroendokrina funktioner i kroppen. De minskar utsöndringen av hormoner samt blodtillförseln till magen och tarmen [18]. Hur dessa receptorer uttrycks i en tumör kan ge insikter i

tumörens biologi och sjukdomsprofil såsom chansen att metastaser kommer att bildas eller ej. Ett annat forskningsområde är NOTCH signaleringens relevans för neuroendokrina tumörers biologi. NOTCH signalering binder samman en cells öde med ödet hos en granncell genom interaktioner mellan NOTCH receptorn och ligander på granncellens membran. NOTCH signalering är viktig för en mängd olika vävnader i kroppen och kan i vissa fall leda till celldelning eller i andra fall celldöd. På grund av dess olika funktioner är det kanske inte oväntat att NOTCH i vissa fall kan fungera som en tumördämpare och i andra fall kan leda till cancer [19]. SST och NOTCH receptorernas uttryck i en tumör kan mätas genom histokemisk testning följt av visuell inspektion och betygsättning. Flera av variablerna i den kommande analysen är resultaten från sådana histokemiska tester och betygsättning av SST och NOTCH receptorers uttryck. Det finns även inkluderat andra biologiska markörer.

4.2 Analys

De data som denna analys grundar sig på är från patienter med endokrina tumörer med ursprung i tunntarmen och bukspottkörteln. Tumörerna utsattes för histokemiska test och resultaten från dessa test kombinerades med patientens kliniska data.

Förberedelse av data är ofta det första steget i en ML analys. Under förberedelsen struktureras data i en sådan form som kan användas som input i en ML algoritm. Instanser av data som saknar vissa variabler kan ersätta de variabler som saknas, med medeltalet av dem eller något annat ändamålsenligt värde. SS behöver fullständiga dataset för att konstruera en modell vilket betyder att varje instans måste ha ett värde på varje variabel. Det skulle inte vara ändamålsenligt ur ett biologiskt perspektiv att ersätta de saknade värdena med medeltalet av variabeln eller ett dylikt värde, de instanser med saknade värden togs helt enkelt bort från detta dataset.

Den första analysen kommer att göras på datan av de patienter vars tumör härstammar från tunntarmen. Varje klinisk variabel kommer i tur och ordning att undersökas som beroende variabel med de kemiska markörerna som oberoende variabler. Genom detta borde det kunna skönjas vilka kemiska markörer som mest påverkar varje klinisk egenskap såsom förekomst av hormonproducerande symptom eller storlek på tumören.

För denna analys kommer ett mjukvarubibliotek som kallas för scikit-Learn att användas [20]. scikit learn är det mest använda maskininlärningsbiblioteket för programmeringsspråket python. Mjukvarubiblioteken pandas [21] och pdpbox [22] kommer även att användas för datamanipulering och visualisering på grund av att de fungerar bra ihop med scikit-learn. Denna analys kommer sedan att göras i den interaktiva programmeringsmiljön jupyter notebooks [23]. Eftersom detta är ett klassificeringsproblem, den beroende variabeln tar antingen värdet 0 eller 1, så kommer en klassificeringsmodell att användas.

En SS modell konstrueras på följande sätt med scikit-learn:

```
classifier = RandomForestClassifier(
                    bootstrap=True,
                    criterion='gini',
                    n_estimators=2000,
                    n_jobs=-1,
                    oob_score=True,
                    random_state=0,
                    max_features=0.5
    )
```

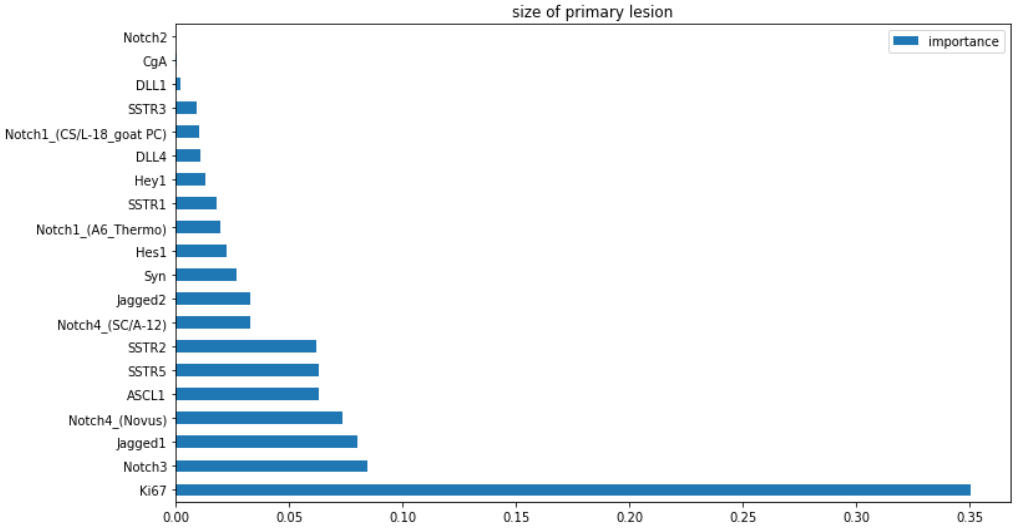
Denna kod ger direktiv för hur SS modellen skall se ut. Raden bootstrap=True betyder att bootstrappingmetoden för skapandet av delmängder skall användas. Raden

criterion=’gini’ innebär att splittningskriteriet som används är gini orenhet. Raden n_estimators=2000 innebär att det konstrueras 2000 enskilda beslutsträd från delmängderna. I litteraturen brukar det användas mellan 500 och 2000 beslutsträd men eftersom det dataset som används är litet och vi i detta fall inte behöver ta i beaktande träningstid eller energianvändning så valdes det stora antalet 2000 träd för att få en så korrekt modell som möjligt. Raden n_jobs=-1 innebär att alla processorkärnor som finns till förfogande får användas för beräkningen. Raden oob_score=True innebär att out-of-bad resultat skall beräknas, random_state=0 möjliggör att resultaten kan reproduceras och max_features=0.5 innebär att en slumpmässig hälft av alla variabler finns till förfogande varje gång ett delningsbeslut skall fattas i en nod på ett beslutsträd. Modellen classifier tränas sedan med koden:

```
classifier.fit(X, y)
```

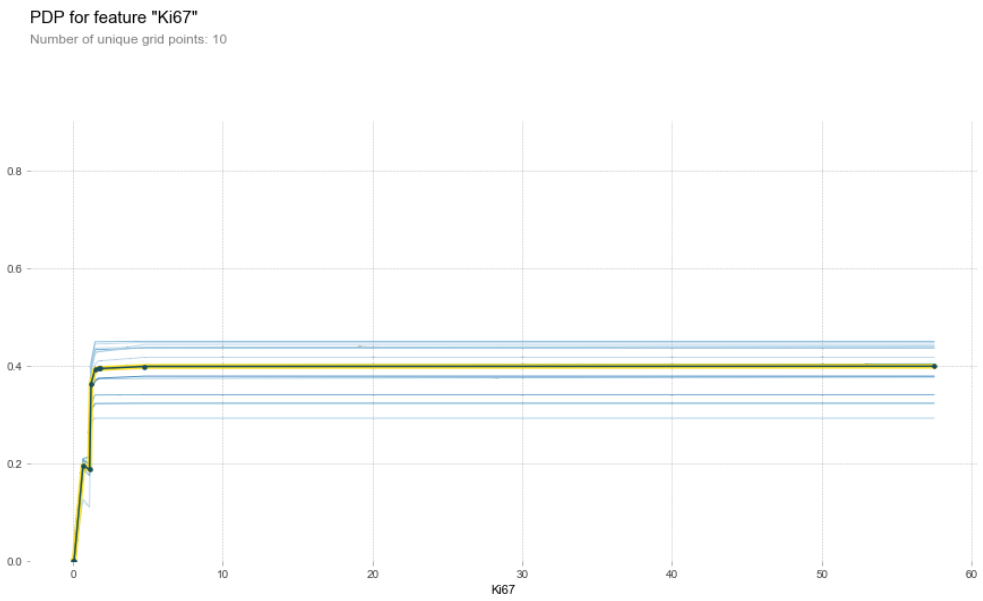
Där X är en tabell där varje rad är en instans och varje kolumn representerar en oberoende variabel, y är den beroende variabelns kolumn. Metodnamnet .fit innebär att modellen anpassas enligt input data. Modellen classifier beräknar automatiskt den relativa betydelsen för varje oberoende variabel, alltså hur viktiga de oberoende variablerna är för att bestämma den beroende variabeln.

Det mest intressanta resultatet av den första analysen är då den beroende variabeln är tumörens storlek. Denna variabel är binär och antar värdet 0 ifall tumören är mindre än 20 mm i diameter och 1 ifall den är större än 20 mm i diameter. Den relativa betydelsen för de olika oberoende variablerna ges i figur 4 där y-axeln visar de olika kemiska markörerna som undersökts och x-axeln visar dessa kemiska markörers relativa betydelse i bestämmandet av den beroende variabeln.



Figur 4. relativ betydelse för de olika oberoende variablerna i bestämmandet av tumörstorlek.

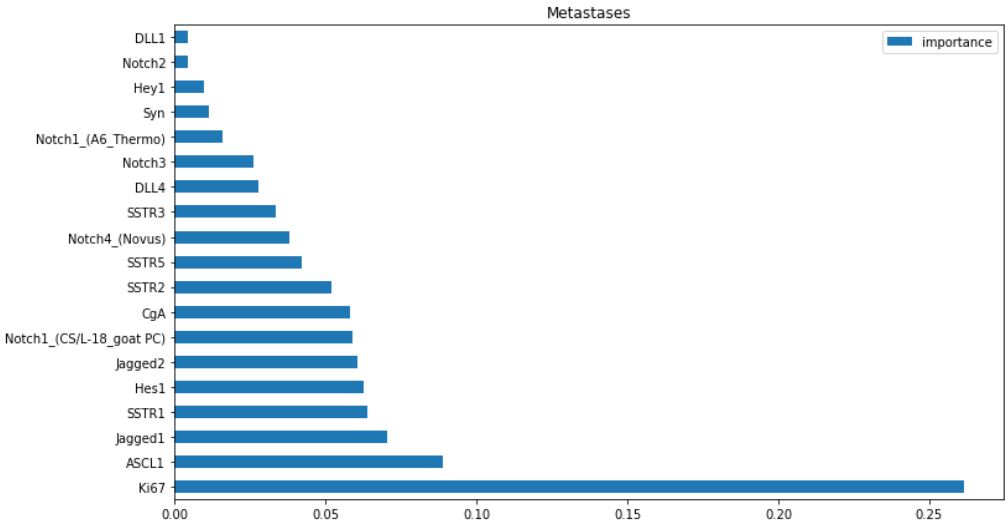
Den kemiska markören Ki67 är enligt figur 4 den absolut viktigaste faktorn i bestämmandet av tumörstorlek. Med hjälp av pdp diagrammet i figur 5 blir relationen mellan Ki67 och tumörstorlek tydligare.



Figur 5. pdp diagram för relationen mellan Ki67 och tumörstorlek.

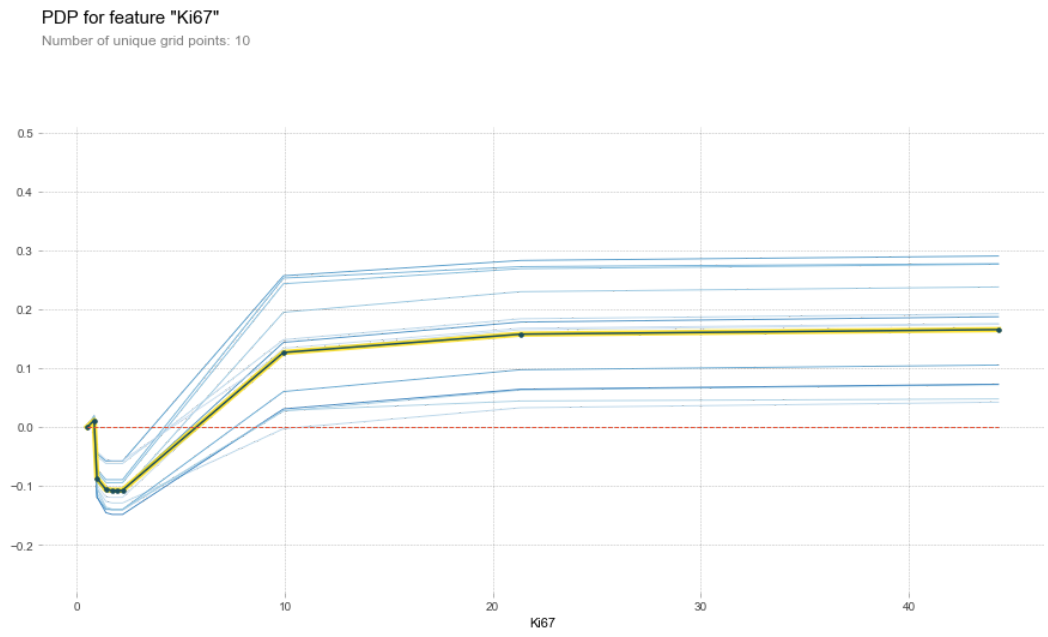
Vi kan se i figur 5 hur chansen att en tumör är större än 20 mm stiger drastiskt då värdet på Ki67 stiger. Ki67 är ett protein i celler som produceras då cellerna gör sig redo att dela på sig, kemiska tester kan mäta procenten av cellerna i en tumör som producerar Ki67 och det är denna procent som är värdet på Ki67 i detta dataset. En vedertagen biologisk uppfattning är att ett högt värde på Ki67 tyder på en snabbt växande och aggressiv cancer [24]. Det faktum att SS modellen gav så mycket betydelse åt Ki67 kan ses som en validering av modellen. Modellens noggrannhet på OOB datan är 0.923 vilket innebär att modellen förutspådde ifall patienterna hade tumörstorlek över eller under 20 mm med ungefär 92 procents noggrannhet. Resultatet av denna analys är kanske inte helt tillförlitligt eftersom det fanns endast tre stycken negativa instanser (under 20 mm i diameter) och tio stycken positiva instanser (20+ mm i diameter).

I följande analys, gjord på datan av de patienter vars tumör härstammar från bukspottkörteln, stod även Ki67 ut som den viktigaste oberoende variabeln i bestämmandet av ifall en patient har metastaser eller ej. Figur 6 visar de oberoende variablernas relativa betydelse.



Figur 6. relativ betydelse för de olika variablerna i bestämmandet av förekomsten av metastaser.

Pdp diagrammet för Ki67 i relation till förekomsten av metastaser i figur 7 är mera svårtolkat än det föregående pdp diagrammet.



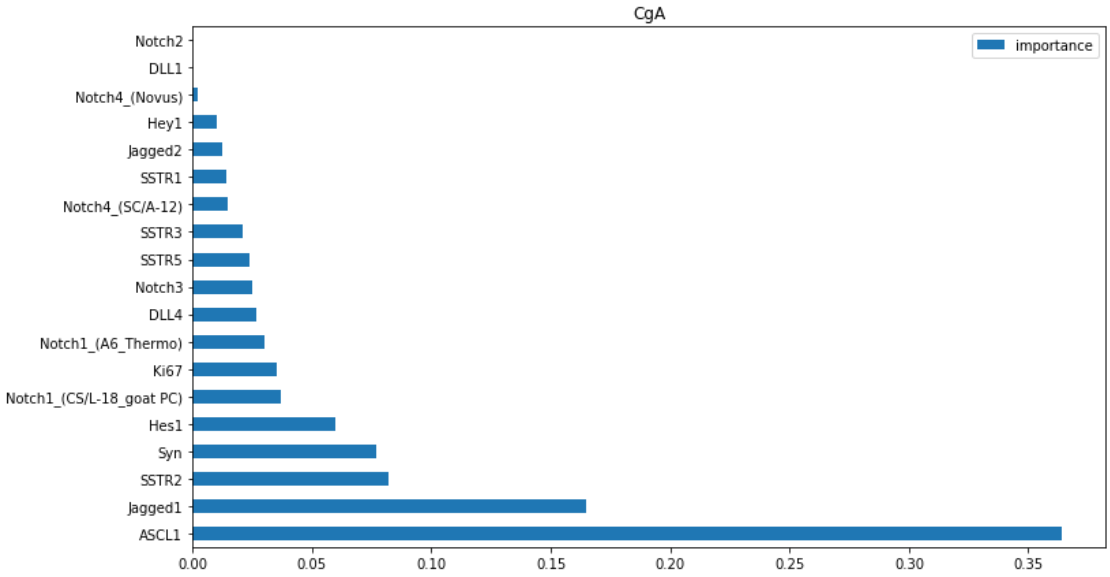
Figur 7. pdp diagram för relationen mellan Ki67 och förekomsten av metastaser.

Fastän detta diagram inte är lika tydligt som det föregående så kan man ändå skönja att en ökning i värdet på Ki67 leder till en ökad chans för metastasbildning. Det finns dokumenterat biologiskt belägg för denna typ av relation mellan Ki67 och förekomst av metastaser i andra typer av tumörer, exempelvis för prostatacancer [25]. Denna analys måste även den betraktas kritiskt då de positiva instanserna, där patienterna i fråga har metastaser, endast är fyra stycken och de negativa är nio stycken. Dessutom visade sig modellens prediktionsförmåga inte vara så stark i denna analys med en noggrannhet av 0.692 på OOB datan.

I en tredje serie med analyser, denna gång på samma data av patienter vars tumör härstammar från bukspottkörteln, gjordes analysen endast på de kemiska markörerna. En efter en valdes en kemisk markör som beroende variabel och de övriga markörerna som oberoende variabler och en SS modell tränades för varje sådan situation. Eftersom värdet på de kemiska markörerna är kontinuerligt (från 0 - 3) så användes regressionsmodellen:

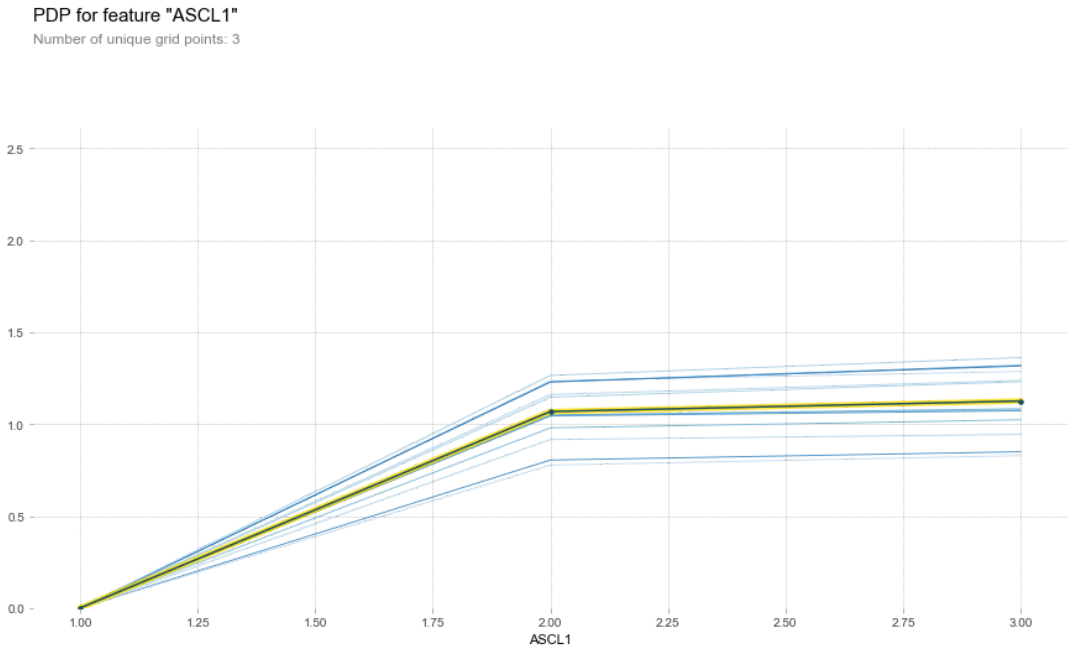
```
regr = RandomForestRegressor(
                    bootstrap=True,
                    criterion='mse',
                    n_estimators=2000,
                    n_jobs=-1,
                    oob_score=True,
                    random_state=0,
                    max_features=0.5
    )
```

De flesta parametrarna i denna modell är desamma som i den tidigare använda klassificeringsmodellen utom delningskriteriet som i detta fall sätts till kvadratiskt medelvärde (eng. MSE eller mean squared error). Ett par av analyserna stod ut genom att det i dem fanns en oberoende variabel som var tydligt viktigast istället för flera medelmåttiga variabler. Rangordningen från en av dessa analyser ges i figur 8.



Figur 8. relativ betydelse för de olika kemiska markörerna i bestämmandet av den kemiska markören Cga.

Pdp diagrammet för relationen mellan ASCL1 och CgA ges i figur 9.



Figur 9. pdp diagram för relationen mellan ASCL1 och CgA .

I pdp diagrammet kan man skönja att en ökning i värdet på ASCL1 leder till en ökning i värdet på CgA. Relationer som denna kunde vara till hjälp i fortsatt forskning men eftersom det inte finns någon vedertagen biologisk kunskap om relationen mellan dessa olika variabler i tumörer så är det svårt att veta ifall denna relation återspeglar ett verkligt biologiskt fenomen eller bara är ett resultat av slumpmässighet i data. Resultaten av denna analys tar inte heller i beaktande individuella skillnader mellan patienter i hur de olika kemiska markörerna inverkar på den beroende variabeln. För vissa patienter kan ett visst värde på en kemisk markör vara signifikant medan den för andra inte är det.

Det faktum att en del kemiska markörer enligt modellen har en låg betydelse måste även det betraktas kritiskt. Eftersom viktiga interaktioner mellan dessa markörer och en beroende variabel kan vara viktiga i en subgrupp av patienterna medan de överskuggas av andra interaktioner då alla patienter tas i beaktande. Ytterligare experiment och analyser krävs alltså för att bekräfta de relationer som hittats.

5. Sammanfattning och slutsats

I denna avhandling har beskrivits varför maskininlärningsmetoder kan vara till nytta vid analys av biologiska data. De kan hitta korrelationer och mönster som med traditionella metoder inte skulle vara så lätta att hitta. Beslutsträd och slumpmässiga skogar, trädbaserade maskininlärningsmetoder, har även beskrivits i detalj och deras för och nackdelar tydliggjorts.

I analysdelen av denna avhandling har sedan en serie analyser gjorts med hjälp av färdig programvara som implementerat slumpmässig-skog-algoritmen. Denna analys har gjorts på kliniska data från patienter och resultatet av histokemiska tester på de tumörer som dessa patienter lidit av. Detta dataset är, på grund av hur sällsynt denna typ av tumör är, inte av en sådan storlek att definitiva slutsatser kan dras med hjälp av analysens resultat. Dock kan de korrelationer som framkommit i analysen vara riktgivande för framtida forskning och, tillsammans med övriga statistiska tester, styrka hypoteser som kan göras om relationer mellan olika variabler i datan och deras biologiska betydelse.

En del resultat som framkom i analysen var förenligt med redan vedertagna biologiska uppfattningar. Dessa resultat handlade om markören Ki67 och dess relation till en tumörs tillväxt och aggressivitet, vilket klassificeringsmodellen som användes kunde urskilja. Eftersom relationen mellan Ki67 och tillväxt/aggressivitet är den enda relation som är känd sedan tidigare är det svårt att dra slutsatser om resultatet av de övriga analyserna. Dock kan det faktum att klassificeringsmodellen lyckades urskilja relationen mellan Ki67 och tillväxt/aggressivitet ses som ett bevis på att trädbaserade maskininlärningsmetoder kan hitta insikter i denna typ av data och därmed ge kredibilitet åt de övriga resultaten av analysen.

[1] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015. [Online] Tillgänglig vid https://www.sciencedirect.com/science/article/pii/S2001037014000464 [hämtat 14.03.2019]

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015. [Online] Tillgänglig vid https://www.sciencedirect.com/science/article/pii/S0304383516303469 [hämtat 14.03.2019]

[3] Mitchell, Tom Michael. *The discipline of machine learning*. Vol. 9. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006. [Online] Tillgänglig vid http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf [hämtat 14.03.2019]

[4] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski, "Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression," *Annals of Behavioral Medicine*, vol. 26, no. 3, pp. 172–181, 2003. [Online] Tillgänglig vid https://link.springer.com/content/pdf/10.1207%2FS15324796ABM2603_02.pdf [hämtat 14.03.2019]

[5] egen figur, översatt från figur i:
J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," *Cancer Informatics*, vol. 2, p. 117693510600200, 2006. [Online] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/ [hämtat 14-03.2019]

[6] E. Deconinck, T. Hancock, D. Coomans, D. Massart, and Y. V. Heyden, "Classification of drugs in absorption classes using the classification and regression trees (CART) methodology," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 39, no. 1-2, pp. 91–103, 2005. [Online] Tillgänglig vid https://www.sciencedirect.com/science/article/pii/S0731708505002116 [hämtat 14.03.2019]

[7] egen figur: inspiration från figur i artikel [4]

[8] L. Breiman, "Technical note: Some properties of splitting criteria," *Machine Learning*, vol. 24, no. 1, pp. 41–47, 1996. [Online] Tillgänglig vid https://link.springer.com/content/pdf/10.1023%2FA%3A1018094028462.pdf [hämtat 14.04-2019]

[9] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006. [Online] https://link.springer.com/content/pdf/10.1007%2Fs10021-005-0054-1.pdf [hämtat 14.03.2019]

[10] A. Géron, "Hands-On Machine Learning with Scikit-Learn & TensorFlow" O'Reilly, 2017

[11] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012. [Online] Tillgänglig vid https://www.sciencedirect.com/science/article/pii/S0888754312000626 [hämtat 14.03.2019]

[12] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. F. T. Van Hijum, "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?," *Briefings in Bioinformatics*, vol. 14, no. 3, pp. 315–326, 2012. [Online] Tillgänglig vid https://academic.oup.com/bib/article/14/3/315/255469 [hämtat 14.03.2019]

[13] egen figur, inspiration från figur i artikeln: D. B. Araya, K. Grolinger, H. F. Elyamany, M. A. Capretz, and G. Bitsuamlak, "An ensemble learning framework for anomaly detection in building energy consumption," *Energy and Buildings*, vol. 144, pp. 191–206, 2017. [Online] Tillgänglig vid https://ir.lib.uwo.ca/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1153&context=electricalpub [hämtat 14.03.2019]

[14] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29, no. 5 (2001): 1189-232. [Online] Tillgänglig vid https://www.jstor.org/stable/2699986?origin=JSTOR-pdf&seq=1#metadata_info_tab_contents [hämtat 14.03.2019]

[15] C. Molnar *"Interpretable machine learning: a guide for making black box models explainable"* [Online] Tillgänglig vid https://christophm.github.io/interpretable-ml-book/pdp.html [hämtat 14.03.2019]

[16] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015. [Online] Tillgänglig vid https://arxiv.org/pdf/1309.6392.pdf [hämtat 14.03.2019]

[17] I. M. Modlin, K. Oberg, D. C. Chung, R. T. Jensen, W. W. D. Herder, R. V. Thakker, M. Caplin, G. D. Fave, G. A. Kaltsas, E. P. Krenning, S. F. Moss, O. Nilsson, G. Rindi, R. Salazar, P. Ruszniewski, and A. Sundin, "Gastroenteropancreatic neuroendocrine tumours," *The Lancet Oncology*, vol. 9, no. 1, pp. 61–72, 2008. [Online] Tillgänglig vid https://www.sciencedirect.com/science/article/pii/S1470204507704102 [hämtat 14.03.2019]

[18] P. L. Kunz, "Carcinoid and Neuroendocrine Tumors: Building on Success," *Journal of Clinical Oncology*, vol. 33, no. 16, pp. 1855–1863, 2015. [Online] Tillgänglig vid http://ascopubs.org/doi/full/10.1200/JCO.2014.60.2532 [hämtat 14.03.2019]

[19] K. Hori, A. Sen, and S. Artavanis-Tsakonas, "Notch signaling at a glance," *Journal of Cell Science*, vol. 126, no. 10, pp. 2135–2140, 2013. [Online] Tillgänglig vid http://jcs.biologists.org/content/126/10/2135.eLetters [hämtat 14.03.2019]

[20] Scikit-Learn https://scikit-learn.org/stable/ [använt 14.03.2019]

[21] Pandas https://pandas.pydata.org/ [använt 14.04.2019]

[22] PDPBox https://github.com/SauceCat/PDPbox/blob/master/docs/index.rst [använt 14.03.2019]

[23] Jupyter https://jupyter.org/ [använt 14.03.2019]

[24] breastcancer.org https://www.breastcancer.org/symptoms/diagnosis/rate_grade [hämtat 14.3.2019]

[25] W. J. Green, G. Ball, G. Hulman, C. Johnson, G. V. Schalwyk, H. L. Ratan, D. Soria, J. M. Garibaldi, R. Parkinson, J. Hulman, R. Rees, and D. G. Powe, "KI67 and DLX2 predict increased risk of metastasis formation in prostate cancer–a targeted molecular approach," *British Journal of Cancer*, vol. 115, no. 2, pp. 236–242, 2016. [Online] Tillgänglig vid https://www.nature.com/articles/bjc2016169 [hämtat 14.03.2019]