**RĪGAS TEHNISKĀ UNIVERSITĀTE**

Datorzinātnes un informācijas tehnoloģijas fakultāte

Lietišķo datorsistēmu institūts

Mākslīgā intelekta un sistēmu inženierijas katedra

**Roberts Čīčis**

**Akadēmiskās** bakalaura studiju programmas „Datorsistēmas”

students, stud. apl. nr. 131RDB057

**CART bāzētu modeļu sarežģītības samazināšanas metožu novērtējums sintētiskiem un dabiskiem datiem**

Zinātniskā vadītāja

Dr.sc.ing., **V. Šakele**

Rīga 2021

Anotācija

Šis dokuments ir izmantots, lai aprakstītu bakalaura darba atskaiti pirmajam semestrim. Šajā atskaitē ir izveidots aptuvenais darba saturs, iezīmējot tā sadalījumu nodaļās. Ir norādīti arī plānotie termiņi bakalaura darba daļu izpildei. Dokumenta saturā ir sniegts ieskats dažās darba nodaļās un aprakstītas konkrētas lietas, kā arī ir norādīts sākotnējas izmantotās literatūras saraksts.

SATURA RĀDĪTĀJS

[1. Izskaidrojamu sistēmu 3](#__RefHeading___Toc50247_2182101473)

[1.1. Izskaidrojamības jēdziens 3](#__RefHeading___Toc50249_2182101473)

[1.2. Sarežģītības jēdziens 5](#__RefHeading___Toc50251_2182101473)

[1.3. Anskomba kvartets 7](#__RefHeading___Toc52025_2182101473)

[2. Klasifikācijas un Regresijas lēmumu koku algoritmi 11](#__RefHeading___Toc50255_2182101473)

[2.1. Lēmumu koki 11](#__RefHeading___Toc10994_3757879547)

[2.2. Lēmumu koku izlase 13](#__RefHeading___Toc10996_3757879547)

[2.3. Gradienta stiprinoša lēmumu koku izlase 14](#__RefHeading___Toc10998_3757879547)

[3. Globālas izskaidrojamības metodes 15](#__RefHeading___Toc50257_2182101473)

[3.1. Pīrsona korelācijas koeficients 15](#__RefHeading___Toc1232_1374917274)

[3.2. Spīrmena rangu korelācijas koeficients 17](#__RefHeading___Toc1234_1374917274)

[3.3. Kendala rangu korelācijas koeficients 18](#__RefHeading___Toc619_1448581708)

[3.4. Paredzošā spēka mērs 19](#__RefHeading___Toc1320_3444560930)

[3.5. Džinī netīrības mērs 20](#__RefHeading___Toc1083_3444560930)

[3.6. Permutācijas mainīgo svarīgums 21](#__RefHeading___Toc1322_3444560930)

[3.7. TreeSHAP mainīgo svarīgums 22](#__RefHeading___Toc1324_3444560930)

[3.8. Abpusējas informācijas mērs 25](#__RefHeading___Toc1326_3444560930)

[3.9. Viena faktora ANOVA un Determinācijas koeficients 26](#__RefHeading___Toc1328_3444560930)

[4. Eksperimenti modeļu sarežģītības noteikšanai 27](#__RefHeading___Toc50261_2182101473)

[4.1. Izmantotās datu kopas un to pirmsapstrāde 27](#__RefHeading___Toc50263_2182101473)

[4.2. Eksperiments ar dabiskiem datiem 29](#__RefHeading___Toc9610_667262781)

[4.3. Metožu novērtējums 32](#__RefHeading___Toc9612_667262781)

# **Izskaidrojamu sistēmu**

## Izskaidrojamības jēdziens

Confalonieri (Confalonieri et al., 2021, pp. 1-4) apgalvo, ka kopš 2020. gada izskaidrojamība (explainability) ir identificēta kā viens no galvenajiem faktoriem mākslīgā intelekta (tālāk MI) sistēmu ieviešanai dažādās nozarēs (Doshi‐Velez & Kim, 2017; Lipton, 2018; Ribeiro, Singh, & Guestrin, 2016). Diskusija, kas seko pieaugošai MI sistēmu ieviešanai tādās nozarēs kā autonoms transports, medicīna, apdrošināšana un finanšu pakalpojumi, atklāj, ka gadījumos, kad MI pieņem lēmumus, ir nozīmīgi sniegt šo lēmumu skaidrojumus lietotājiem, izstrādātājiem un regulētājiem praktisku, sociālu un legālu iemeslu dēļ.

Kā praktisku piemēru Confalonieri min GDPR (General Data Protection regulation) definēto tiesību iegūt “jēgpilnu informāciju par loģiku, kas tiek izmantota MI lēmumiem”. Šī definīcija bieži tiek interpretēta kā tiesības uz “izskaidrojumu”, kad lēmumu veic automātiska sistēma (Parliament and Council of the European Union, 2016/679).

Confalonieri apgalvo, ka lietotāju tiesību un tehnoloģiju pieņemšanas problemātika nav vienīgie iemesli inteliģentu sistēmu aprīkošanai ar izskaidrojamības spējām. Izskaidrojamība arī ir nepieciešama MI sistēmu izstrādātājiem, lai uzlabotu sistēmas izturību un iespējotu diagnostiku, kas palīdz nepieļaut aizspriedumus, negodīgumu un diskrimināciju MI lēmumos, kā arī palielinātu lietotāju uzticību MI lēmumiem. Tāpēc MI sistēmām spēja izskaidrot lēmumus ir kļuvusi par ļoti vēlamu un nepieciešamu īpašību. Izskaidrojumi ļauj lietotājiem izprast un efektīvi lietot sistēmu un palīdz atkļūdot MI pieņemtus lēmumu. Izskaidrojamība veicina konkrētā domēna izpratni par problēmu un veicina inovācijas tajā. Izskaidrojumiem ir tieša saistība ar lietotāju uzticību un spēju viņus pārliecināt par MI sistēmas efektivitāti.

Confalioneri spriež, ka vēsturiski pirmie centieni veidot izskaidrojamas sistēmas pieklusa 1980. gadu vidū, kad ekspertu sistēmas nespēja attaisnot augstās cerības (Buchanan & Shortliffe, 1984; Wick & Thompson, 1992) un interese par tām atdzima, pateicoties nesenajiem atklājumiem mašīnapmācības sfērā. Confalioneri secina, ka nav vispārpieņemtas definīcijas par to, kas ir izskaidrojamība un ko nozīmē izskaidrojamība. Tāpat izskaidrojamībai nav arī matemātiskas definīcijas.

Pat, ja nepastāv, skaidras un vienotas definīcijas izskaidrojamībai, iedvesmojoties no sociālajam zinātnēm Millers (Miller, 2019) piedāvā izskaidrojamības definīciju: “Izskaidrojamība ir pakāpe, ar kādu cilvēks var saprast cēloņus MI pieņemtam lēmumam” vai: “Izskaidrojamība ir cilvēka spēja patstāvīgi paredzēt modeļa prognozes.”

Taksonomiski Kristofs Molnars (Molnar, 2019, nodaļa 2.2) piedāvā izskaidrojamību iedalīt pēc šādiem kritērijiem:

1. Sarežģītību samazinošs vai modeļa analīzējošs
   1. Kritērijs raksturo vai izskaidrojamība tiek iegūta samazinot modeļa sarežģītību vai izmantojot metodes pēc modeļa apmācības. Sarežģītību samazinošas metodes piemērs ir sekli lēmuma koki. Pēc modeļa analizējošas metodes piemērs ir permutācijas mainīgo svarīgums (tālāk PIMP no angļ. **P**ermutation feature **IMP**ortance)
2. Interpretāciju metodes rezultāts
   1. Mainīga vispārīga mērs - Izskaidrojamības metodes, kas veido apkopojošu statistiku katram mainīgajam
   2. Modeļa iekšējais stāvoklis - Izskaidrojamu modēļu iekšējais stāvokli, piemēram, lineārās regresijas gadījumā mainīgo koeficienti, lēmumu koku vizuāla reprezentācija
   3. Piemēru bāzētas - Visas metodes, kas atgriež konkrētus datu piemērus (jau esošus vai uzģenerētus). Piemēri, šādām metodēm ir kontrastējošu un prototipsku vai kritizējošu piemēru atrašana (klasterizācija)
   4. Surogātveida izskaidrojams modelis – vienkāršāks, izskaidrojams modelis, kas aproksimē oriģinālā modeļa darbību
3. Modelim speficiska vai modeļa neatkarīga metode  
   Kritērijs raksturo vai metode ir pielietojama visiem modeļiem (modeļa neatkarīga) vai metode ir pielietojama tikai konkrētu modeļu izskaidrošanai
4. Lokāls izskaidrojums vai globāls izskaidrojums - Vai izskaidrojums raksturo tikai vienu piemēru no datu kopas vai visa modeļa (datu kopas) darbību?

Šis darbs koncentrējas uz globālu izskaidrojumu (jeb mainīga svarīguma) metožu salīdzināšanu (4. kritērijas Kristofa izskaidrojamības taksonomijā).

## Sarežģītības jēdziens

Nodaļa 1.2 ir balstīta uz Höge zinātnisko rakstu (Marvin Höge et.al., 2018). Modeļu izvēle bieži tiek izskaidrota kā kompromisa atrašana starp modeļa spēju pielāgot datus un tam nepieciešamo modeļa sarežģītību. Modeļa apmācība nodrošina gan spēju reprezentēt esošos datus gan spēja precīzi vispārināt uz nākotnes datiem (Guthke, 2017; White, 2017). Pastāv skaidras definīcijas, kas raksturo modeļa apmācības kvalitāti, diemžēl modeļa sarežģītībai neeksistē skaidras definīcijas. Tāpēc nepieciešama dziļāka izpēte modeļa sarežģītības terminam.

Sarežģītība ir netverama (elusive) īpašība (Van Emden, 1971). Lai gan intuitīvi ir viegli izprast vai kaut kas ir sarežģīts vai nē, nepastāv skaidras definīcijas vai iespējas kvantitatīvi raksturot sarežģītību (Gell‐Mann, 1995a). Galvenie iemesli sarežģītības definēšanai un kvantificēšanai ir (Bialek et al., 2001a):

1. Pierādīt, ka evolūcijas (dabiskās evolūcijas vai sistēmas evolūcijas) gaitā lietas kļūst sarežģītākas
2. Kvantitatīvi izmērīt cik sarežģīts ir izskaidrojums, hipotēze vai modeļa fenomens
3. Novērtēt, cik grūti ir loģiski aprakstīt savstarpēji saistītas sistēmas stāvokli.

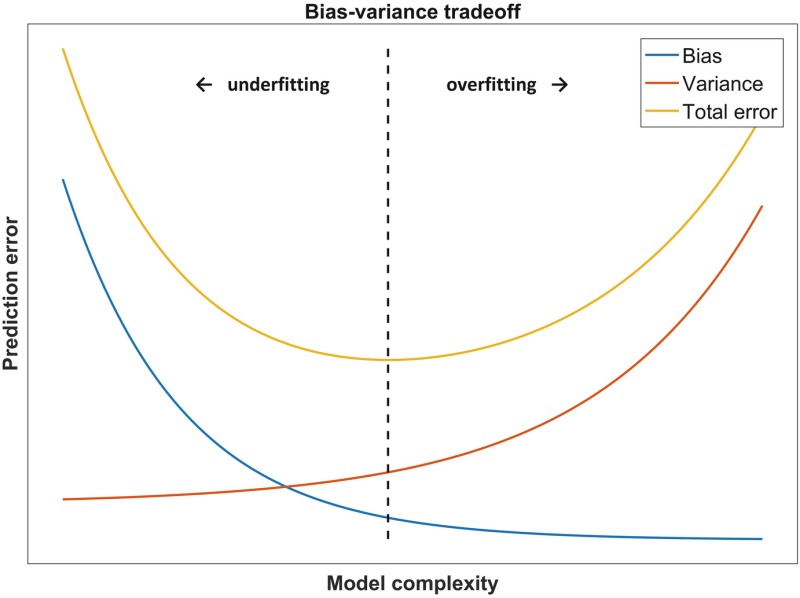
Centienos aprakstīt sarežģītību ir iegūta izpratne par īpašībām, kuras raksturo sarežģītību: sarežģītība ir maksimāla starp kārtību un nejaušību - gan pilnībā sakārtotu sistēmu, gan pilnīgi nejaušu sistēmu uzskata par nesarežģītu (Ladyman et al., 2012; Prokopenko et al., 2009; Rudnicki et al., 2016; Wiesner, 2015). Sarežģītība is saistīta ar emerģenci (angl. emergence). Emerģence ir īpašība, kad vairāku identisku objektu kopai piemīt lielāka sarežģītība nekā objektiem individuāli savstarpējas iedarbības rezultātā. Diskusiju par turpmākajām sarežģītības īpašībām var atrast, piemēram, Edmonds (1999, 2000). Šo īpašību konceptuālā izpratne ir novedusi pie daudziem sarežģītības rādītājiem, no kuriem daudzi ir apkopoti neizsmeļošā sarakstā (Du, 2016; Lloyd, 2001).

Modeļi ar vairāk parametriem un mainīgajiem ne vienmēr ir labāki par modeļiem ar mazāk parametriem un mainīgajiem (Perrin et al., 2001), ja vien papildu parametri un mainīgie nav labi izvēlēti.

Mainīgo pievienošana ir pieņemta kā sarežģītības palielināšana. Tomēr tas attiecas tikai uz modeļa parametrisko sarežģītību (Vanpaemel, 2009) un neietver matemātiskās funkcijas, kas veido modeli (piemēram, Pande et al., 2015). Turklāt tas neņem vērā parametru mijiedarbību un neatbild uz jautājumu, kurš modelis ir optimāli sarežģīts konkrētam mērķim (Claeskens, 2016). Optimālais modelis nav nedz pārāk sarežģīts (Myung, 2000; Orth et al., 2015; Warren & Seifert, 2011), nedz arī pārāk vienkāršs (Hunt et al., 2007; Mendoza et al., 2015) konkrētajam modelēšanas uzdevumam.

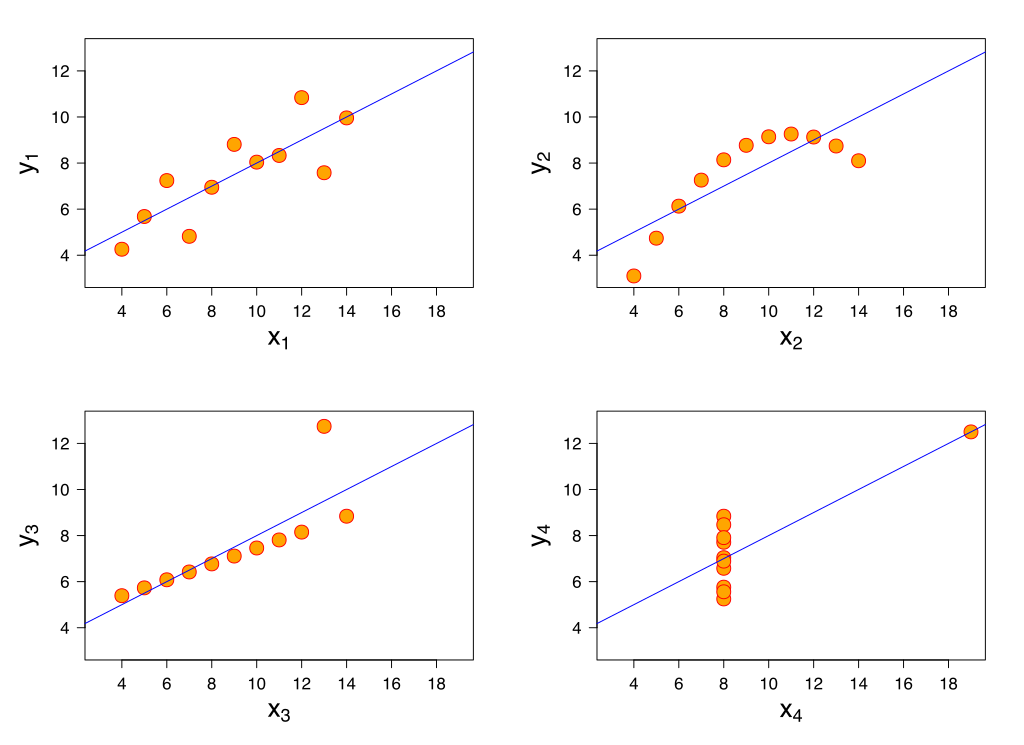
Pārāk sarežģīts modelis, ir pakļauts pārmērīgai uzstādīšanai. Tas nozīmē, ka tas cieš no pārāk augstas elastības un slikti identificējamiem parametriem, un tam ir lielas prognozes variācijas (Lever et al., 2016), izņemot, ja tiek izmantotas lielas datu kopas, lai ierobežotu modeļa parametrus.

Pretēji ir pārāk vienkāršiem modeļiem, kas ir nepietiekami. Tad dispersija ir maza, jo dažus modeļa parametrus var viegli ierobežot ar pieejamajiem datiem. Tomēr modelis parāda lielu novirzi starp tā prognozēm un datiem, jo ir pārāk vienkārši precīzi aprakstīt sistēmu un saskaņot datus. Novirze (bias) apraksta, cik labi modelis vidēji atbilst datiem, t.i., modeļa precizitāte. Dispersija (variance) atgādina atlikumu kvadrātā ap šo vidējo (kā nenoteiktības mēru), t.i., modeļa precizitāti. Abi ir vizualizēti 1.1 att.

1.1. att Novirzes-dispersijas kompromiss [todo: latvianize]

## Anskomba kvartets

Anskomba kvartets (Anscombe’s quartet) ir tādu datu kopu apkopojums, kurām piemīt identiski aprakstošie mēri (descriptive measures) – vidējā vērtība, dispersija (variance), Pīrsona korelācijas koeficients starp x un y, determinācijas koeficents – R2, taču tajā pat laikā ļoti atšķirīgi vizuālie sadalījumi. Anskomba kvartets ir attēlots att. 1.2

1.2 att Anskomba kvartets

1973. gadā statistiķis Francis Anskombs (Anscombe, 1973) izveidoju šo datu kopu, lai cīnītos ar tajos laikos izplatītu mītu statistiķu vidū, ka “skaitliskie aprēķini ir precīzi, bet grafiki ir aptuveni“. Anskomba kvartets parāda, cik svarīgi ir attēlot risinājumus vizuālā veidā, nevis paļauties uz redukcionisma paņēmieniem (aprakstošu statistiku), lai novērtētu risinājumus. Līdzīga pieeja ir vērojama mašīnapmācības praksē – modeļa optimizācijas funkcijas zaudējums un modeļu novērtēšans mēri tiek izmantoti ar redukcionistisku pieeju. Tas var novest pie neparedzētiem izaicinājumiem un pat neveiksmēm, mēģinot ieviest MI modeļus produkcijas vidē (D'Amour, A. et al., 2020).

Gadījumus, kad starpnieka mērs, nespēj precīzi attēlot produkcijas vidi vai īsto darbības effektivitāti raksturo Gudharta likums (Goodhart’s law). Gudharta orģinālā definīcija ir “jebkura statistiska likumsakarība mēdz sabrukt, kad uz to tiek izdarīts spiediens kontroles nolūkos” (Goodhart, 1984) Strathens (Strathern, 1997) piedāvā vispārināmāku definīciju “kad mērs kļūst par mērķi, tas pārstāj būt par labu mēru”.

Gudharta likumu taksonomiski var iedalīt 4 kategorijās (Manheim & Garrabrant, 2018)

1. Regresijas Gudharta likums
   1. Izvēloties aproksimējošo mēru, jūs izvēlaties ne tikai patieso mērķi, bet arī starpību starp aproksimējošo mēru un mērķi.
   2. Piemērs, augstums ir saistīts ar basketbola spējām un faktiski tieši palīdz, bet labākais spēlētājs ir tikai 190cm garš un nejauša 213 cm gara persona 20 gadu vecumā, visticamāk, nebūs tik laba
2. Ekstremāls Gudharta likums
   1. Telpas, kurās aproksimējošam mēram ir ekstremāla vērtība, var ļoti atšķirties no parastajām pasaulēm, kurās tiek novērota korelācija starp aproksimējošo mēru un mērķi.
   2. Piemērs, garākais cilvēks pasaulē Roberts Vadlovs bija 2,72m garš. Viņš nebūtu labs basketbola spēlētājs, jo hipofīzes dēļ viņam vajadzēja kāju iemauktus, lai spētu staigāt
3. Cēlonisks Gudharta likums
   1. Ja starp aproksimējošu mēru un mērķi pastāv cēloņsakarība, maiņa aproksimējošajā mēra var nemainīt mērķa mainīgo
   2. Piemērs, kāds, kurš vēlas būt garāks, varētu novērot, ka augums ir saistīts ar basketbola prasmēm, un nolemt sākt praktizēt basketbolu, lai kļūtu garāks.
4. Pretiniecisks Gudharta likums
   1. Optimizējot aproksimējošo mēru, pretincisikiem aģenti ir iniciatīvi korelēt savu mērķi ar jūsu pilnvaru, tādējādi iznīcinot korelāciju ar jūsu mērķi.
   2. Piemērs, Topošie basketbola spēlētāji var melot par savu augumu. Uzņēmuma darbinieki var sākt falsificēt vai negodīgi ziņot par datiem.

Datu vizualizācija un EDA (Exploratory Data Analysis) (Tukey, 1977) ir viens no veidiem kā cīnīties ar pārāk lielu sarežģītību un liekiem mainīgajiem modelī. Izmantojo tādas metodes kā dispersijas (variance), trūkstošo vērtību un augstas savstarpējās korelācijas filtrus ir iespējams samazināt modeļa sarežģītību atmetot liekus mainīgos vai mainīgos kuru iekļaušana modelī ir vairāk apgrūtinājums nekā ieguvums modelēšanas procesam. Tāpat ar EDA palīdzību ir iespējams vizuāli analizēt mainīgā lietderību datu kopā.

Manuālam EDA piemīt tādas problēmas kā Simpsona paradokss (Simpson’s paradox), kad attiecības kas piemērotas konkrētām datu apakšgrupām, pazūd, kad šīs apakšgrupas tiek apvienotas (Simpson, 1951). Simpsona paradoksu var arī attiecināt uz n-mainīgo modeli. Cilvēki var salīdzinoši vienkārši izskaidrot modeli, kuram ir tikai 1 mainīgais, piemēram, attēlojot attiecību starp mainīgo un mērķi izkliedes grafikā (scatterplot), bet, ja modelī ir vairāk nekā viens mainīgais, tad modeļa darbības izskaidrošana kļūst problemātiska. Pirmā problēma sarežģītu sistēmu interpretācijā ir tīri cilvēciska. Psihologs Džordžs Millers (G. A. Miller, 1956) aprakstīja cilvēku prāta spēju ierobežojumus, aprakstot slaveno Millera likumu, ka cilvēki īstermiņa atmiņā spēj rīkoties ar 5-9 objektiem (7+/-2). Otrā problēma ir emerģences īpašība (aprakstīta nodaļā 1.2), ja eksistē modelis y = f(x1, x2), tad ir salīdzinoši viegli izskaidrot kāda ir attiecība starp (y, x1) un (y, x2), piemēram, izmantojot izkliedes grafikus, bet, praktiski neiespējami paredzēt kā sistēma darbojas kā kopums y = f(x1, x2), t.i.

mainīgo kopai {x1, x2} savstarspējas iedarbības rezultātā modelējot y piemīt lielāka sarežģītība nekā mainīgajiem atsevišķi modelējot y (emerģences īpašība, apskatīta nodaļā 1.2).

Izskaidrojamības metodes var tikt izmantotas, lai izprastu, kādi modeļi no relatīvi līdzīgu modeļu apkopojuma ir vislabāk piemēroti konkrētajam uzdevumam. Praktisks piemērs izskaidrojamības nepieciešamībai ir COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) programmatūra, kura tiek lietota ASV tiesās, lai paredzētu noziegumu recidīvisma gadījumus. COMPAS sastāv no 137 jautājumu aptaujas; to izstrādāja privāta kompānija, tāpēc tā modeļa struktūra tiek turēta noslēpumā, un nav zināms, kā aptaujas jautājumi ļauj paredzēt cilvēku uzvedību. Dotā programmatūra tiek kritizēta; ProPublica žurnālistu pētījums (Angwin et al., 2016) ir apgalvojis, ka programmatūras modelim piemīt rasistiski aizspriedumi un tas nav drošs un precīzs veids, kā paredzēt vardarbīgas noziedzības un recidīvisma gadījumus.

Pētījums (Julia & Farid, 2018) atklāja, ka programmatūra COMPAS nav ievērojami precīzāka par “indivīdiem bez vai ar minimālu pieredzi kriminālistikā”. Salīdzinājumam: COMPAS precizitāte 65% pareiza pret kriminālistikā nepieredzējušu indivīdu precizitāti – 63% pareizi. Likumsakarīgi, ka tas raisīja bažas par šāda modeļa lietderību un drošību

(Rudin, et al., 2018a), it īpaši programmatūras COMPAS tieksme atzīmēt indivīdus ar garu un vērā ņemamu noziegumu vēsturi kā zema riska grupu un neskaidra privātuma politika.

(Rudin, 2018b) argumentē, ka “melnās kastes” modeļi nav labs risinājums “augsta riska” lēmumiem tā vietā vajadzētu izmantot sākotnēji izskaidrojamu modeļus. Sintija (Cynthia) skaidro, ka “melnās kastes” neizskaidrojamība tiek izmantota peļņas gūšanai no intelektuālā īpašuma. Programmatūras COMPAS alternatīva ir CORELS (Certifiably Optimal Rule Lists) (Angelino, et al., 2017) modelis, kas ir tikpat akurāts kā COMPAS, bet izmanto vien 3 likumu sistēmu un ir pilnībā caurskatāma, go:

JA indivīda vecums ir no 18 – 20 un dzimums ir vīriešu,

TAD tiek paredzēts arests (2 gadu laikā),

CITĀDI JA vecums ir no 21 – 23 un indivīdam bijuši 2 – 3 kriminālās sodāmības,

TAD tiek paredzēts arests,

CITĀDI JA indivīdam ir vairāk nekā 3 kriminālās sodāmības,

TAD tiek paredzēts arests, CITĀDI arests netiek paredzēts.

# **Klasifikācijas un Regresijas lēmumu koku** algoritmi

## **Lēmumu koki**

Nodaļa 2.1 balstīta uz (Hastie, T. et al. 2009, pp. 305-321) grāmatu “Elements of Statistical Learning”. CART (angl. Classfication and regression trees) algoritms ir lēmumu koku bāzēta metode, lai sadalītu mainīgo “telpu” taisnstūros un taisntūru apakštelpai izveidotu ļoti vienkāršus modeļus (konstants modelis, apakštelpai paredz konstantu rezultātu). Tiek veikti rekursīvi bināri lēmumi alkatīgi (greedy) izvēloties mainīgo, kas maksimizē vai minimizē mūsu izvēlēto kritēriju (parasti, klasifikācijas uzdevumos - džinī netīrības mērs (gini impurity), regresijas uzdevumos - vidējā kvadrātiskā kļūda). Process turpinās līdz tiek sasniegts kāds apstādināšanas kritērijs (beidzas datu punkti kurus sašķelt vai arī pārsniegts kāds no modēļa ierobēžojumiem, piemēram, koka dzīļums).

Ieskatīsimies dziļāk principos kā darbojas lēmumu koks regresijas gadījumā. Formula 1.1. apraksta kā no mainīgajiem x tiek iegūts modeļa rezultāts y

(1.1)

kur

y – ir modeļa paredzējums

lbrace x in {R sub m} rbrace – ir identitātes funkcija, kas atgriež 1, ja x ir apakškopā R sub m, savādāk 0

c sub m – vidējā vērtība modeļa apakštelpa (konstants modelis)

R sub m – modeļa apakštelpa m

M – modeļa apakštelpas kopa

Atrast labāko bināro apakštelpu kopu minimizējot kvadrātisko summu ir, praktiski neiespējami izskaitļot, tāpēc tiek implementēts alkatīgs (greedy) algoritms, kurš izskata tikai nākamo līmeni (šķēlumu), nevis mēģina uzbūvēt globāli optimālus lēmumu kokus. Lai atrastu nākamo bināro šķēlumu tiek izvēlēts kvadratiskās kļūdas minimums no abām apakštelpām, tas nodrošina, ka ir iespējams ātri sarēķināt katra mainīgā šķēluma punktu visām mainīgā vērtībām, atrodot šķēluma punktu, kas naivā veidā minimizē kvadrātisko kļūdu. Atrodot labāko šķēluma punktu process tiek turpināts visām apakštelpām līdz sasniegts beigšanas kritērijs. Formulās bāzētu dziļāku procesa aprakstu iespējams atrast avotā (Hastie, T. et al. 2009, pp. 306)

Lai algoritms strādātu klasifikācijas lēmumu kokiem vienīgā izmaiņa, kas jāveic ir šķēlumu kritērija maiņa uz džinī netīrības mēru (gini impurity) definēts formulā 1.1:

(1.1)

kur

G – Džinī netīrības mērs

C – kopējais klašu skaits mērķim

p sub i – varbūtība izvēlēties klasi i datu punktam

Džinī netīrības mēra vērtības atrodas skalā no 0 līdz 0.5, kur vērtība 0 raksturo šķēlumu, kas spēj ideāli sadalīt klases un vērtība 0.5 raksturo šķēlumu, kurš ir maksimāli “netīrs” t.i. nespēj sadalīt klases vispār. Kā izmantot Džinī vai modeļa kritērija guvumu mainīgo svarīguma noteikšanai aprakstīts 3.1 nodaļā.

Viens no vislielākajiem trūkumiem lēmumu kokiem ir pārāk liela pielāgošanās (overfitting) mācību datiem, praktiski, “iegaumējot” datus nevis attēlojot vispārināmas saistības starp mainīgajiem. Šo problēmu risina lēmumu koku izlases, kuras aprakstītas nākamajā nodaļā.

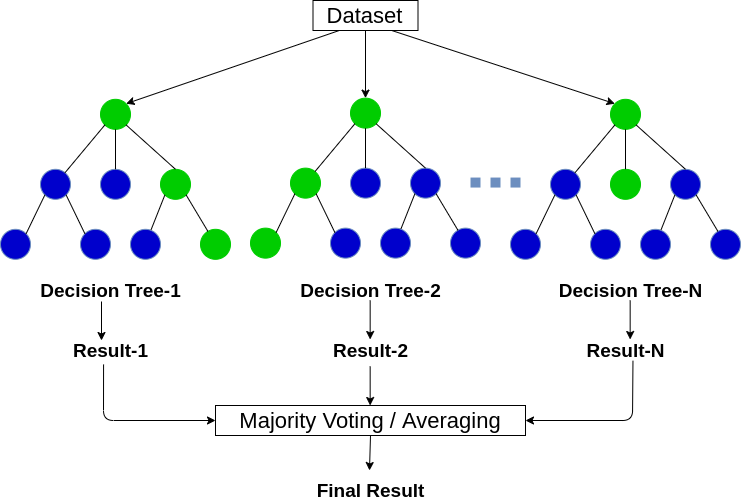
## **Lēmumu koku izlase**

Nodaļa balstīta uz Breimana (Breiman, 2004) zinātnisko rakstu.

Lēmumu koku izlases metode veido vairākus lēmumus kokus un veic agregācijas (ensemble) veida paredzējumu (klasifikācijas uzdevumos izvēloties vispopulārāko klasi, regresijas ņemot vidējo vērtību no visu koku paredzējumiem). Viena no galvenajām atšķirībām no lēmumu kokiem, kas ļauj lēmumu koku izlasēm sasniegt optimālu “sarezģītību”, nevis pārāk stipri pielāgoties datiem kā to dara lēmumu koki ir:

1. Lēmumu koku izlasēs apmācot lēmumu kokus izmanto nejauši izvēlētu mācību datu apakškopu nevis pilnu mācību datu kopu
2. Lēmumu koku izlasēs apmācot lēmumu kokus katram šķēlumam izvērtē labākos šķēlumus tikai mainīgo apakškopai nevis visiem mainīgajiem (parasti kvadrātsakne no mainīgo skaita)

Vizuāla reprezentācijas lēmumu koka izlasei ir redzama attēlā 1.1. Dziļākai izpratnei par matemātiskajiem principiem ir iespējams iegūt (Louppe, 2014) 4. nodaļā



Att 1.1 Lēmumu koku izlašu agregācijas process

## **Gradienta stiprinoša lēmumu koku izlase**

Nodaļa 2.3 balstīta uz (Natekin & Knoll, 2013) zinātnisko rakstu.

Līdzīgi kā lēmumu koku izlases metode, ir balstīta uz vairāku “vāju” modeļu agregāciju vienā lielā modelī. Gradienta stiprinoši lēmumu koki (tālāk GBDT, no angl. Gradient boosted decision trees) izmanto citādu koku “audzēšanas” stratēģiju. Galvenā ideja GBDT ir pievienot jaunus kokus agregētajam modelim sekvenciāli, mēģinot minimizēt iepriekšējā agregētā modeļa kļūdu (šajā kontekstā to sauc par atlikumu, angliski - residual). Šo metodi aprakstīja (Freund & Schapire, 1997) un (Friedman et al., 2000) (Friedman, 2001) kur tā tika saukta par gradienta stiprinošas mašīnas (Gradient boosting machines) metodi. GBDT metode ir guvusi popularitāti ar tādam programmatūras implementācijām kā CatBoost (Ostroumova et al., 2017), LightGBM (Guolin Ke et al., 2017) un XGBoost (Chen & Guestrin, 2016).

GBDT secīgi apmāca jaunus lēmuma koku modeļus, lai veidotu precīzāku aproksimāciju mērķa mainīgajam. Algoritms stradā tā, lai jaunie lēmuma koki būtu maksimāli korelēti ar negatīvo gradientu zaudējuma funkcijai.

Vispārīgu GBDT ir iespējams definēt šādi:

Ieejas parametri:

* ieejas dati (x, y)
* iterāciju skaits M
* zaudējuma funkcijai phi (y, f)
* modeļa izvēle h(x, theta)

Algoritms:

1. Inicializē sākuma modeli f sub 0 ar konstanti (vidējā vērtība regresijai, populārākā klase klasifikācijai)
2. līdz sasniegts iterāciju limits M darīt:
   1. aprēķini negatīvo gradienta vērtību
   2. apmāca jaunu izvēlēto modeli
   3. atrod labāko gradienta nolaišanas (gradient descent) soļa izmēru
   4. atjauno funkcijas aproksimāciju ņemot vērā jaunā izvēlētā modeļa atlikumu (residual)

Dziļāku izskaidrojumu iespējams atrast (Friedman, 2001).

# Globālas izskaidrojamības metodes

## Pīrsona korelācijas koeficients

*Nodaļa 3.1 balstīta uz grāmatu (Boslaugh & Watters, Statistics in a nutshell, 2008, pp. 182-186)*.

Pīrsona korelācijas koeficients (dažreiz to dēvē par produkta un momenta korelācijas koeficientu no angļ. product-moment correlation coefficient) ir lineārā saistība starp diviem mainīgajiem. Pīrsona korelācijas koeficienta diapazons ir (−1, 1), ar 0 norāda, ka nav saistības starp mainīgajiem, un lielākās absolūtās vērtības norāda uz ciešāku saistību starp mainīgajiem. Pīrsona korelācijas koeficients var būt maldinošs, ja starp datiem ir nelineāras attiecības.

Matemātiski Pīrsona korelācijas koeficients ir definēts formulā 1.1. un 1.2.

, (1.1)

kur SSx - kvadrātu summa x

SSx - kvadrātu summa y

SSxy- kvadrātu summa x un y

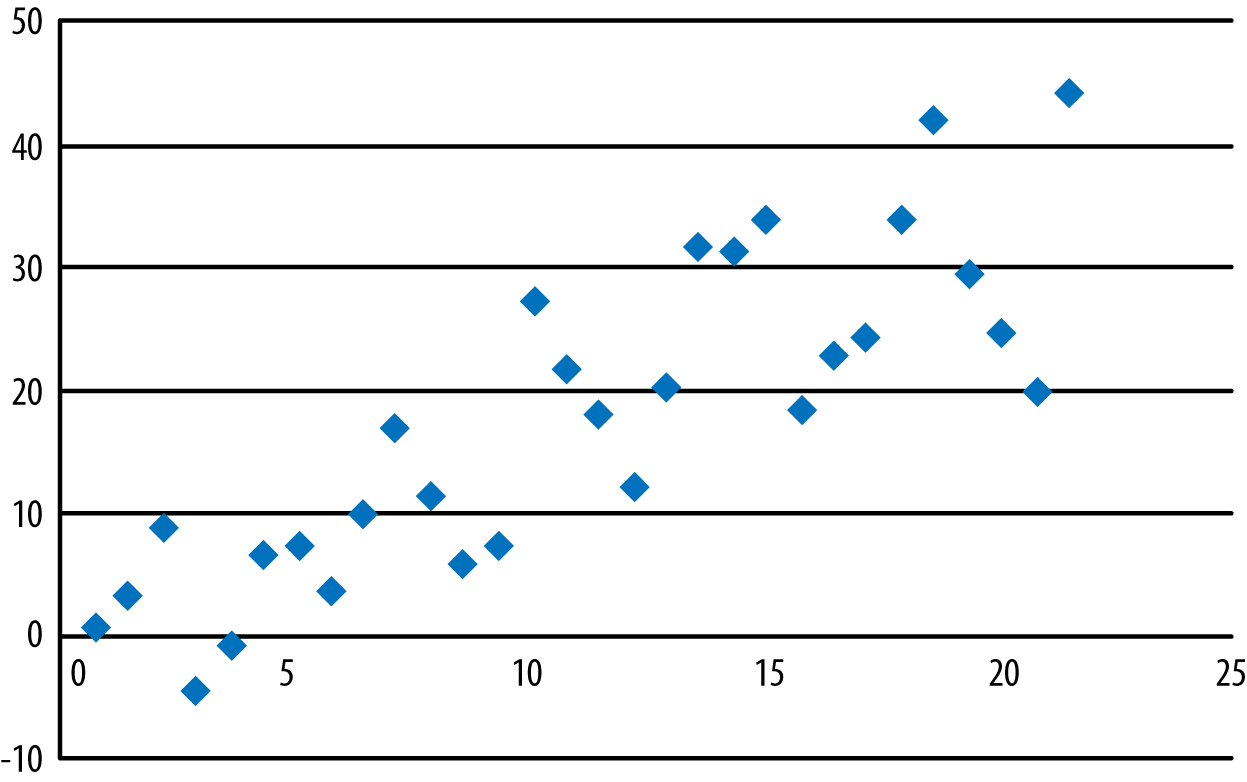
, (1.2)

kur xi – konkrēta x vērtība

– parauga vidējā vērtība

n – parauga izmērs

Praktisksu piemērus Pīrsona koeficientam var redzēt attēlā (3.1)

Att. 3.1. Izkliedes grafiks ar Pīrsona koeficientu 0.84

Kopumā Pīrsona koeficients tiek izmantots, lai izvērtētu lineāru monotonu attiecību spēku starp 2 mainīgajiem, koeficients ir simetrisks. Eksperimentā tiek izmantota absolutā vērtība starp katru mainīgo un mērķi, lai noteiktu, cik svarīgs ir mainīgais.

## **Spīrmena** **rangu** **korelācijas** koeficients

*Nodaļa 3.2 balstīta uz grāmatu (Corder & Foreman, Nonparametric Statistics for Non-Statisticians A Step-by-Step Approach, 2009, pp. 124-131).*

Spīrmena rangu korelācijas koeficients piemīt praktiski identiskas īpašības kā Pīrsona korelācijas koeficientam, ar vienu būtisku atšķirību, Spīrmena koeficients spēj precīzi attēlot nelineāras attiecības un ir balstīts uz datu vērtību ranžēšanu nevis datu vērtībām.

Matemātiski Spīrmena koeficients ir definēts formulā 1.1

, (1.1)

kur

, (1.2)

, (1.3)

n – ranga pāru skaits

Di – starpība starp rangu pāri

g – vienādu grupu skaits datos

ti – vienādu vērtību skaits grupā

Eksperimentā tiek izmantota absolutā vērtība starp katru mainīgo un mērķi, lai noteiktu, cik svarīgs ir mainīgais.

## **Kendala rangu korelācijas koeficients**

*Nodaļa balstīta uz grāmatu (Prokohorov, 2001, pp. 18-27)*

Kendala koeficients izmēra divu datu kopu līdzību salīdzinot ranžētu objektu pārus. Koeficients atkarīgs no pāru saskanības un nesaskanības.

Matemātiski Kendala koeficients ir definēts formulā 1.1:

, (1.1)

kur

, (1.2)

ri  rangs y no (x, y) pāra, kuram x rangs ir vienāds ar i

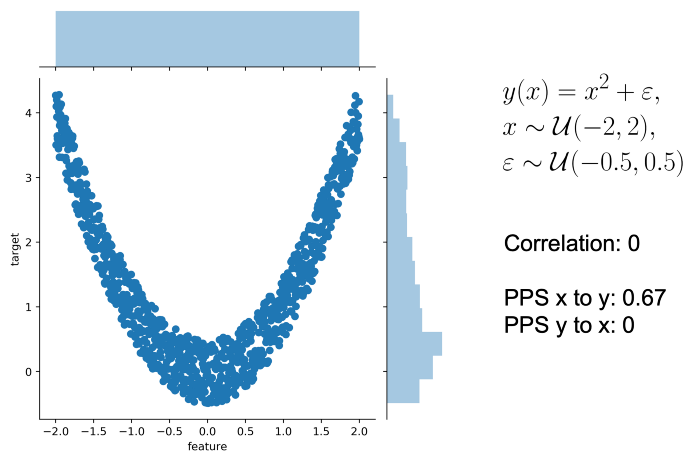
N – elementu skaits kur j > i un rj > ri

Eksperimentā tiek izmantota absolutā vērtība starp katru mainīgo un mērķi, lai noteiktu, cik svarīgs ir mainīgais.

## **Paredzošā spēka mērs**

Paredzošā spēka mērs (angl. Predictive power score, tālāk PPS) (Wetschoreck et al., 2020)ir asimetriska, datu tipa neatkarīga metode, kas spēj noteikt lineāras vai nelineāras attiecības starp diviem mainīgajiem. Rezultāts ir no 0 (nespēj prognozēt) līdz 1 (perfekti prognozē). Galvenā ideja PPS ir veidot lēmumu kokus ar 1 mainīgo, lai paredzētu mērķi, izmantojot 4-kārtīga krusteniskās validāciju klasifikācijas gadījumā nejauši stratificējot (stratification) testa un treniņa datus pēc mērķa mainīgā, lai pārbaudītu rezultātu uzticamību. Izskaidrojums pieejams: <https://www.sr-sv.com/the-predictive-power-score/>

Viena no interesantajām īpašībām PPS ir, ka tas spēj attēlot asimetriskas attiecības.

Att 1.1 PPS aprēķina piemērs

Attēlā redzams, ka x, y korelācija ir 0.67, bet y, x korelācija ir 0. X var paredzēt y labi, jo ir deterministiska kvadrātiska attiecība, bet Y nespēj paredzēt X, jo, teiksim, kad Y ir 4 ir grūti pateikt vai X ir aptuveni -2 vai 2 (vienam Y punktam pieder 2 X punkti, kas izraisa nepārliecību par Y vērtību). Šī īpašība nav svarīga eksperimentā, bet var būt lietderīga manuālā datu analīzē.

## **Džinī netīrības mērs**

*Nodaļa balstīta uz grāmatu (Trevor Hastie, The Elements of Statistical learning, 2009, pp. 593-594)*

Viens no veidiem kā noteikt mainīgo globālo svarīgumu ir izmantot koku modeļa katra šķēluma kritērija vērtības guvumu. Šo guvumu apkopo pa visiem kokiem un sarēķina katram mainīgajam atsevišķi.

Breimans (Breiman, 2001) definējā kā izvērtēt mainīgo svarīgumu lēmumu kokos.

Mainīgajam Xj paredzot Y summējot netīrības mazinājumu p(t) delta i(st, t) katram koka zaram, kur Xj ir izmantots, ņemot vidējo vērtību pa visiem kokiem mežu izlasē (random forest). Matemātiska definīcija formulā 1.1

, (1.1)

kur

p(t) – proporcija Nt / N paraugu skaits, kas sasniedz t

jt – indetifikators mainigajam, kas izmantots, lai sadalītu koku zaru t

## **Permutācijas mainīgo svarīgums**

Breimans (Breiman, 2001) piedāvāja arī alternatīvu veidu kā aprēķināt mainīgo svarīgumu koku modeļos. Permutācijas mainīgo svarīgums (Permutation feature importance), saukts arī par Vidējo precizitātes zaudējumu (Mean Decrease Accuracy) vai Vidējo kļūdas palielinājumu (Mean Increase Error) balstās uz modeļa precizitātes zaudējuma (vai kļūdas palielinājuma) mērīšanu nejauši sajaucot mainīgā vērtības.

(Louppe, 2014, pp. 123-125)

(1.1)

kur

xj – mainīgais j

π sub j (%LAMBDA) – replika xj vērtībām, kur veikta nejauša sajaukšana

mk1...mkM-1 – indeksi kokiem, kas izveidoti no atkārtotas paraugu ņemšanas (bootstrap), kurā nav (xi, yi)

## **TreeSHAP mainīgo svarīgums**

Nodaļa 3.7 balstīta uz grāmatas (Molnar, “Interpretable machine learning”, 2019) *5.9 – 5.10* nodaļām

Lai izprastu TreeSHAP algoritma darbību nepieciešams raksturot:

1. Šaplija vertības
2. Summāra mainīgo attiecinājuma (attribution) metode
3. TreeSHAP darbības principu
4. TreeSHAP mainīgo svarīgumu aprēķinu

Šaplija vērtības – metode no koalīciju spēļu teorijas – ļauj “taisnīgi” sadalīt “izmaksu” starp spēlētājiem (mainīgajiem) (Shapley, 1951). Modeļa lēmumus var izskaidrot, pieņemot, ka katrs mainīgais ir “spēlētājs”, kur katra spēlētāja “izmaksa” ir modeļa paredzējums. Matemātiski Šaplija vērtības aprēķinu ir definēts formulā 1.1

(1.1)

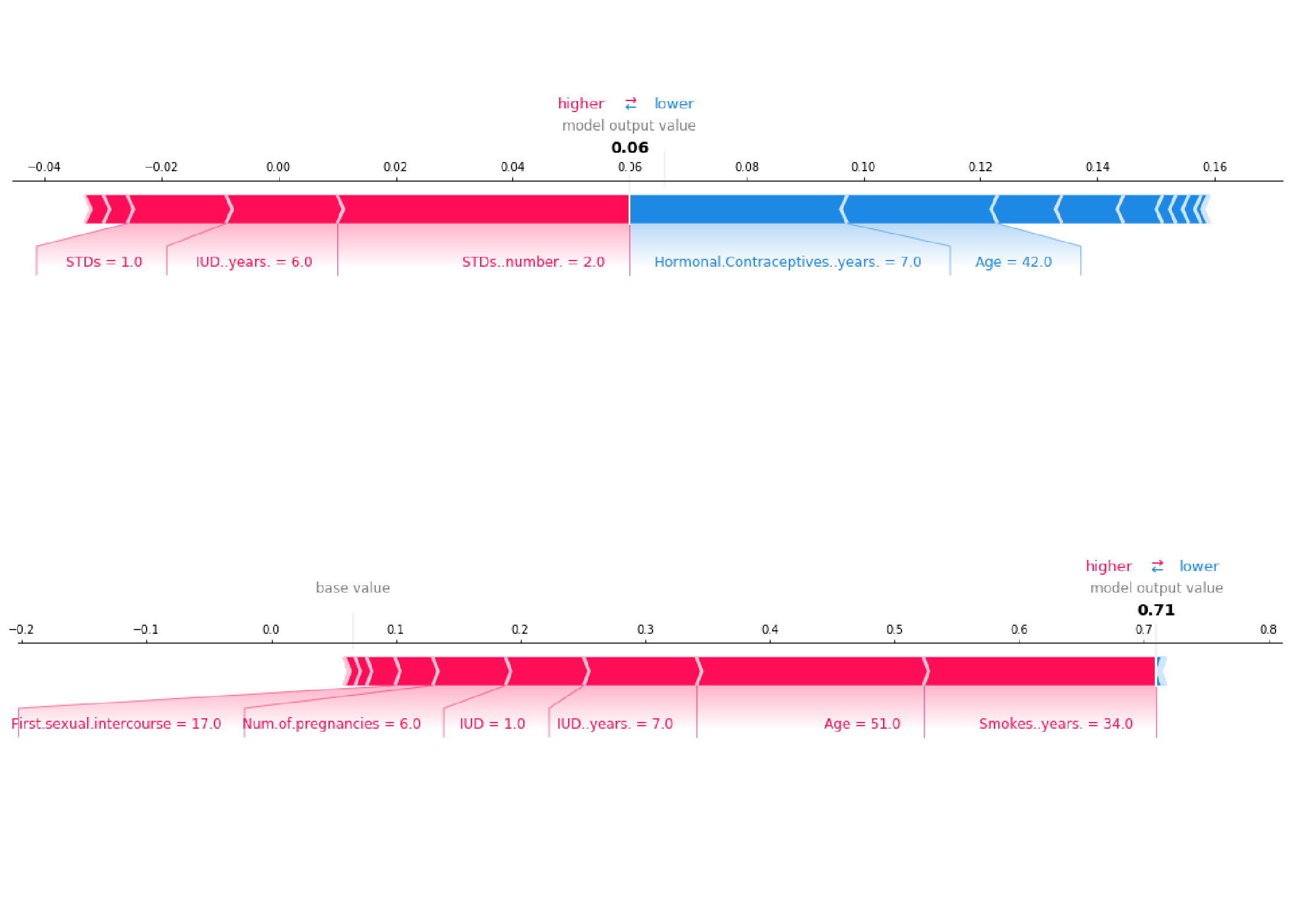
kur

phi i – Šaplija vertībā spēlētājam i

S – spēlētāju koalīcija

fx(s) – koalīcijas vērtība spēlētājiem sadarbojoties

Summāra mainīgo attiecinājuma (attribution) metodes mērķis ir izskaidrot katras instances paredzējumu x aprēķinot mainīgā “ieguldījumu” (contribution) paredzējumā. Izskaidrošanas metode aprēķina Šaplija vērtības no spēļu teorijas un katras instances mainīgo vērtības uzvedas kā spēlētāji koalīcijā. Šaplijas vērtības izskaidrojums tiek reprezentēts kā summārs mainīgo attiecinājums vizuāla reprezentācija principam 1.1.att.

1.1 att Summāra mainīgo attiecinājuma metodes vizualizācija [todo: latvianize]

Attēlā redzams kā dzemdes kakla vēža paredzēšanas modelis ar summāru mainīgo attiecinājuma metodi raksturo mainīgo vērtību ietekmi uz dzemdes kakla vēža varbūtību, kā redzams 1.1. att risku mazina faktori zilā krāsa un palielina faktori sarkanā krāsā, sasumējot visus risku palielinošus un samazinošus faktorus iegūst modeļa paredzējumu.

Matemātiski šo procesu definē formula 1.1:

(1.1)

kur

g - izskaidrojuma modelis

z’ {0, 1}M – koalīcijas vektors

M – maksimālais koalīcijas izmērs

phi j – Šaplija vērtībās mainīgā j vērtībām

phi o – modeļa paredzējums ar tukšu koalīciju (visi mainīgo vērtības ir “trūkstošas”)

Galvenā motivācija TreeSHAP lietošana ir ka, Šaplija vērtību aprēķins ir nepraktisks, jo tā sliktākā gadījuma skaitļošana sarezģītība ir 2N (eksponenciāla). Dziļāku izpratni par to kā Šaplija vērtības tiek aprēķinātas koālīcijās var iegūt (Grabisch & Roubens, 1999).

Lai spētu praktiski pielietot Šaplija vērtības modeļu darbības izskaidrošanā tās tiek deterministiski aproksimētas ar lēmumu koku bāzētos modeļos ar TreeSHAP algoritma palīdzību, kas samazina sliktākā gadījuma skaitļošana sarezģītību uz polinomiālu laiku (eksponenciāla vietā). Precīzs TreeSHAP algoritma skaitļošanas laiks raksturots formulā 1.1:

O(TLD2) (1.1)

kur

T ir lēmumu koku skaits,

L ir maksimālais koka galotņu skaits (leaves) jebkurā kokā,

D ir maksimālais dziļums jebkuram kokam

Pseidoalgoritma aprakstu TreeSHAP ir iespējams atrast (Lundberg, 2018) 3.2 nodaļā

TreeSHAP mainīgā svarīgums tiek aprēķinats izmantojot formulu 1.1:

(1.1)

kur

Ij – mainīgā j svarīguma vērtība

phi j pow i – mainīgā j Šaplija vērtības i mainīgā vērtībai

Būtībā aprēķins ir ļoti vienkāršs – TreeSHAP algoritma visas iegūtās absolūtās Šaplija vērtības tiek sasumētas katram mainīgajam un tas rada mainīgā svarīguma mēru.

## **Abpusējas informācijas mērs**

*Nodaļa balstīta uz (Cover & Thomas, 1991, pp. 19-22)*

Abpusējas informācijas mērs (mutual information) mēra “informācijas” daudzumu

kurš pieder vienam mainīgajam par otru mainīgo. To arī var aprakstīt kā nenoteiktības (uncertainty) mazinājumu vienam mainīgajam par otru zināšanu dēļ.

Matemātiski abpusējas informācijas mēru aprakstu formula 1.1:

(1.1)

Viena no galvenajām priekšrocībām abpusējas informācijas mēram ir, ka tas spēj modelēt nemonotonas, nelineāras attiecības (atšķirībā no Pīrsona korelācijas koeficienta, kas spēj modelēt tikai monotonas lineāras attiecības)

## **Viena faktora ANOVA un Determinācijas koeficients**

*Nodaļa balstīta uz avotu (McDonald, 2014, pp. 145-156)*

Klasifikācijas uzdevumos ir iespējams izmantot viena faktora ANOVA (angl. **AN**alysis **O**f **Va**riance) metodi, lai noteiktu cik labi mainīgais spēj atdalīt starp mērķa klasēm. Galvenā ideja ir aprēķināt vidējo vērtību katrā grupā un salīdzināt katras grupas vidējās vērtības dispersiju ar vidējo dispersiju katrā grupā. Matemātiski process ir definēts formulās 1.1-1.1:

(1.1)

(1.1)

(1.1)

kur,

F – viena faktora ANOVA F vērtības

MST – vidējā kvadrātiskā kļūda starp grupām

MSE – vidējā kvadrātiskā kļūda

Yij – ij datu punkts (novērojums)

Ti – grupas kopējā vērtība

G – ir datu punktu (novērojumu) kopējais skaits

ni – punktu skaits grupā i

n – kopējais datu punktu skaits

k – kopējais grupu skaits

~~Determinācijas koeficients ir Pīrsona korelācijas koeficients kvadrātā (apskatīts nodaļā 3.1), lai gan tam nav atšķirības ranžējot mainīgo svarīgumu kā metode tā ir iekļauta, lai būtu iespējams salīdzināt rezultātus klasifikācijas un regresijas uzdevumiem agregētos kopskatos.~~

# Eksperimenti modeļu sarežģītības noteikšanai

## Izmantotās datu kopas un to pirmsapstrāde

Datu kopas tiek apstrādātas ar šādu loģiku:

* Visus kategoriskos (categorical) mainīgos kodē ar apzīmējuma kodējumu (label encode)
* Nepārtrauktos mainīgos ar trūkstošām vērtībām pārvēršs kategoriskajos mainīgajos ar papildus klasi – trūkstošo vērtību klasi
* Nepārtrauktie mainīgie bez trūkstošām vērtībām tiek izmantoti modelī bez papildus darbībām
* Dažās datu kopas neinformatīvas vai atkļudošanas informācijas kolonas ir izņemtas skatīt tabulu 4.1, lai iegūtu detalizētāku informāciju.

Tabula 4.1 Izmantotās datu kopas un to pirmsapstrāde

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Datu kopas vārds | Mainīgo skaits | Mainīgo skaits pēc apstrādes | Instanču skaits | Instanču skaits pēc apstrādes | Uzdevums | Mērķis |
| Diabetes | 10 | 10 | 442 | 442 | Regresija | Kvantitatīvs slimības progresijas mērs |
| Boston housing | 13 | 13 | 506 | 506 | Regresija | Mājas vērtība |
| Crime | 128 | 122 | 1994 | 1994 | Regresija | Vardarbīgi noziegumi uz iedzīvotāju |
| Ames housing | 82 | 79 | 2930 | 2925 | Regresija | Mājas vērtība |
| Wine | 13 | 13 | 178 | 178 | Vairāku klašu klasifikācija | Vīnu klases |
| Breast cancer | 30 | 30 | 569 | 569 | Bināra klasifikācija | Krūts vēža esamība |
| Phishing | 30 | 30 | 11055 | 11055 | Bināra klasifikācija | Vai mājaslapa pīkšķerē |
| Mushrooms | 22 | 22 | 8124 | 8124 | Bināra klasifikācija | Sēņu indīgums |

## **Eksperiments ar dabiskiem datiem**

Eksperimentā datu kopas tiek nejauši sadalītas apmācības (70%) un testa datos (30%) . Regresijas datu kopas tiek vienkārši nejauši sadalīti, bet klasifikācijas dati tiek nejauši sadalīti stratificēti (stratified) saglabājot klašu balansu starp apmācības un testa datiem. Abos gadījumos nejauša datu sadalīšana tiek deterministiski kontrolēta izlases veidā (random seed).

Katrā iterācijā tiek:

1. Apmācīts bāzes modelis izlases veidā (random seed) N reizes ar visiem mainīgajiem, veidojot bāzes precizitāti.
2. Sarēķināts mainīgo svarīgums izmantojot konkrētu mainīgo svarīguma metodi un apmācības datus
3. No mainīgo svarīguma aprēķina izvēlēts mainīgais ar vismazāko svarīgumu
4. Visnesvarīgakais mainīgais tiek izņemts no apmācības un testa datu kopas
5. Modelis tiek pārmācīts izlases veidā (random seed) N reizes bez visnesvarīgākā mainīgā
6. Process turpinās līdz vairs nav mainīgo kurus atmest no datu kopas

Katrā iterācijā tiek saglabāts kāda datu kopa izmantota, kurš mainīgais tika atmests, kāda ir modeļa precizitāte, kādi izlases parametri izmantoti 1) datu nejauša sadalīšanai 2) modeļa nejaušai apmācībai. Process notiek visām mainīgo samazināšanas metodēm (9), visiem modeļu paveidiem (3), visām datu kopām (8), visiem mainīgajiem (atkarībā no datu kopas 10-128), 10 nejaušām izlases veida datu sadalīšanām, 10 nejaušām modeļu izlases veida apmācīšanām.

Regresijas uzdevumu precizitātes izvērtēšanai tiek izmantota vidējā kvadrātiskā kļūda, klasifikācijas uzdevumu izvērtēšanai Matteja korelācijas koeficients (Matthew correlation coefficient). Kvadrātiskā kļūda jau ir apskatīta nodaļā 3.1. Matteja korelācijas koeficientu raksturo formula 4.1:

(4.1)

kur

MCC - Matteja korelācijas koeficients

TP – patiesi pozitīvi klasificēti dati (true positives)

TN – patiesi negatīvi klasificēti dati (true negatives)

FP - nepatiesi pozitīvi klasificēti dati (false positives)

FN – nepatiesi negatīvi klasificēti dati (false negatives)

Matteja korelācijas koeficients izvēlēts kā klasifikācijas modēlu precizitātes mērs, jo tas spēj labi attēlot nebalansētu klasifikācijas uzdevumu modelēšanas precizitāti. Dziļāku izskaidrojumu iespējams atrast (Boughorbel et al., 2017).

Modeļi tiek apmācīti ar bāzes hiperparametriem. Tiek izmantota programmēšanas valodas Python modulis Sklearn, lai apmācītu modeļus – lēmumu koks, lēmumu koku izlase, un Python implementācija XGBoost modulim – gradienta stiprinoša lēmumu koku izlases modelim.

Eksperimenti tika veikti uz AMD Ryzen 3200g procesora ar 4 kodoliem. Eksperimenti aizņēma ~80 stundas. Kopēju iterāciju skaitu iespējams redzēt apkopojuma tabulā 4.2

Tabula 4.2 Eksperimenta iterāciju skaits

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Datu kopas vārds | Mainīgo skaits | Datu sadalīšanas izlašu skaits | Modeļu apmācības izlāšu skaits | Mainīgo svarīguma metožu skaits | Iterāciju skaits\* | Procenti no kopējā iterāciju skaita |
| Diabetes | 10 | 10 | 10 | 9 | 9000 | 3.05% |
| Boston housing | 13 | 10 | 10 | 9 | 11700 | 3.96% |
| Crime | 128 | 10 | 10 | 9 | 115200 | 39.02% |
| Ames housing | 82 | 10 | 10 | 9 | 73800 | 25.00% |
| Wine | 13 | 10 | 10 | 9 | 11700 | 3.96% |
| Breast cancer | 30 | 10 | 10 | 9 | 27000 | 9.15% |
| Phishing | 30 | 10 | 10 | 9 | 27000 | 9.15% |
| Mushrooms | 22 | 10 | 10 | 9 | 19800 | 6.71% |
| Kopā |  |  |  |  | 295200 | 100.00% |

\*Iterāciju skaits ir iegūts sareizinot mainīgo skaitu, datu sadalīšanas izlāšu skaitu, modeļu apmācības izlašu mainīgo skaitu, mainīgo svarīguma metožu skaitu

## **Metožu novērtējums**

Izmantotā Literatūra

*Confalonieri, R. et al. “A historical perspective of explainable Artificial Intelligence.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11, (2021),.*

*Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable MachineLearning.arXiv: Machine Learning.*

*Lipton, Z.C. (2018). The Mythos of Model Interpretability. Queue, 16, 31 – 57.*

*Ribeiro, Marco Tulio et al. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).*

*EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, OJ 2016 L 119/1, Pieejams: https://eur-lex.europa.eu/eli/reg/2016/679/oj*

*Buchanan, B. and E. Shortliffe. “Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley series in artificial intelligence).” (1984).*

*Wick, Michael R. and W. B. Thompson. “Reconstructive Expert System Explanation.” Artif. Intell. 54 (1992): 33-70.*

*Miller, T.. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” Artif. Intell. 267 (2019): 1-38.*

*Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.*

*Höge, M. et al. “A Primer for Model Selection: The Decisive Role of Model Complexity.” Water Resources Research 54 (2018): 1688-1715.*

*Guthke, Anneli. “Defensible Model Complexity: A Call for Data-Based and Goal-Oriented Model Choice.” Ground water 55 5 (2017): 646-650 .*

*White, J. T.. “Forecast First: An Argument for Groundwater Modeling in Reverse.” Ground water 55 5 (2017): 660-664 .*

*Emden, Van. “An analysis of complexity.” (1971).*

*Mann, G.. “The Quark and the Jaguar: adventures in the simple and the complex.” (1995a).*

*Mann, M.. “What is complexity? Remarks on simplicity and complexity by the Nobel Prize-winning author of The Quark and the Jaguar.” Complex. 1 (1995b): 16-19.*

*Bialek, W. et al. “Complexity through nonextensivity.” Physica A-statistical Mechanics and Its Applications 302 (2001): 89-99.*

*Ladyman, James et al. “What is a complex system?” European Journal for Philosophy of Science 3 (2013): 33-67.*

*Prokopenko, M. et al. “An information-theoretic primer on complexity, self-organization, and emergence.” Complexity 15 (2009): 11-28.*

*Rudnicki, Lukasz et al. “Monotone measures of statistical complexity.” Physics Letters A 380 (2016): 377-380.*

*Wiesner, K.. “Complexity Measures and Physical Principles.” (2015).*

*Edmonds, B. (1999). What is complexity?—The philosophy of complexity per se with application to some examples in evolution. In The evolution of complexity.*

*Edmonds, B. (2000). Complexity and scientific modelling. Foundations of Science, 5(3), 379–390.*

*Du, J. (2016). The “weight” of models and complexity. Complexity, 21(3), 21–35.*

*Lloyd, S. (2001). Measures of complexity: A nonexhaustive list. IEEE Control Systems Magazine, 21(4), 7–8.*

*Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. Journal of Hydrology, 242(3), 275–301.*

*Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. In Y. Bengio et al. (Eds.), Advances in neural information processing systems (Vol. 22, pp. 1919–1927). Red Hook, NY: Curran Associates.*

*Pande, S. et al. “Hydrological model parameter dimensionality is a weak measure of prediction uncertainty.” Hydrology and Earth System Sciences Discussions 12 (2014): 3945-4004.*

*Claeskens, G. (2016). Statistical model choice. Annual Review of Statistics and Its Application, 3, 233–256.*

*Myung, I. J. (2000). The importance of complexity in model selection. Journal of Mathematical Psychology, 44(1), 190–204.*

*Orth, R. et al. “Does model performance improve with complexity? : A case study with three hydrological models.” Journal of Hydrology 523 (2015): 147-159.*

*Warren, D. and S. Seifert. “Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria.” Ecological applications : a publication of the Ecological Society of America 21 2 (2011): 335-42 .*

*Hunt, R. et al. “Are models too simple? Arguments for increased parameterization.” Ground water 45 3 (2007): 254-62.*

*Mendoza, P. A. et al. “Are we unnecessarily constraining the agility of complex process-based models?” Water Resources Research 51 (2015): 716-728.*

*Lever, Jake et al. “Points of Significance: Model selection and overfitting.” Nature Methods 13 (2016): 703-704.*

*D'Amour, A. et al. “Underspecification Presents Challenges for Credibility in Modern Machine Learning.” ArXiv abs/2011.03395 (2020).*

*Goodhart, C.. “Monetary Theory and Practice.” (1984).*

*Strathern, Marilyn (1997). "'Improving ratings': audit in the British University system". European Review. John Wiley & Sons. 5 (3): 305–321.*

*Manheim, David and Scott Garrabrant. “Categorizing Variants of Goodhart's Law.” ArXiv abs/1803.04585 (2018)*

*Tukey, J.. “Exploratory data analysis.” Addison-Wesley series in behavioral science : quantitative methods (1977).*

*Simpson, E.. “The Interpretation of Interaction in Contingency Tables.” Journal of the royal statistical society series b-methodological 13 (1951): 238-241.*

*Miller, G. A.. “The magical number seven plus or minus two: some limits on our capacity for processing information.” Psychological review 63 2 (1956): 81-97.*

*Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. ProPublica; 2016. Pieejams : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*

*Dressel, Julia and H. Farid. “The accuracy, fairness, and limits of predicting recidivism.” Science Advances 4 (2018).*

*Rudin, C. et al. “The age of secrecy and unfairness in recidivism prediction.” ArXiv abs/1811.00731 (2018a).*

*Rudin, C.. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” Nature Machine Intelligence 1 (2018b): 206-215.*

*Angelino, E. et al. “Learning Certifiably Optimal Rule Lists for Categorical Data.” J. Mach. Learn. Res. 18 (2017): 234:1-234:78.*

*Hastie, T. et al. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition.” Springer Series in Statistics (2009).*

*Breiman, L.. “Random Forests.” Machine Learning 45 (2004): 5-32.*

*Louppe, Gilles. “Understanding Random Forests: From Theory to Practice.” arXiv: Machine Learning (2014).*

*Natekin, Alexey and A. Knoll. “Gradient boosting machines, a tutorial.” Frontiers in Neurorobotics 7 (2013).*

*Freund, Y. and R. Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” COLT 1997 (1997).*

*Friedman, J.. “Special Invited Paper-Additive logistic regression: A statistical view of boosting.” Annals of Statistics 28 (2000): 374-376.*

*Friedman, J.. “Greedy function approximation: A gradient boosting machine.” Annals of Statistics 29 (2001): 1189-1232.*

*Ostroumova, L. et al. “CatBoost: unbiased boosting with categorical features.” NeurIPS (2018).*

*Ke, Guolin et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” NIPS (2017).*

*Chen, T. and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).*

*Corder, Gregory W. and D. Foreman. “Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach.” (2009).*

*Prokhorov, A.V. (2001), "Kendall coefficient of rank correlation", Encyclopedia of Mathematics, EMS Press.*

*Wetschoreck, F., Krabel, T., & Krishnamurthy, S. (2020). 8080labs/ppscore: zenodo release (1.1.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.4091345*

*Breiman, L.. “Random Forests.” Machine Learning 45 (2004): 5-32.*

*Shapley, L. S.. “Notes on the n-Person Game — II: The Value of an n-Person Game.” (1951).*

*Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019, Pieejams: https://christophm.github.io/interpretable-ml-book/.*

*Grabisch, M. and M. Roubens. “An axiomatic approach to the concept of interaction among players in cooperative games.” International Journal of Game Theory 28 (1999): 547-565.*

*Lundberg, Scott M. et al. “Consistent Individualized Feature Attribution for Tree Ensembles.” ArXiv abs/1802.03888 (2018),*

*Cover, T. and J. Thomas. “Elements of Information Theory.” (1991).*

*McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing,*

*Rudin, C. et al. “The age of secrecy and unfairness in recidivism prediction.” ArXiv abs/1811.00731 (2018a).*

*Rudin, C.. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” Nature Machine Intelligence 1 (2018b): 206-215.*

*Boslaugh, S. and P. Watters. “Statistics in a nutshell.” (2008), O'Reilly Media.*

*Anscombe, F.. “Graphs in Statistical Analysis.” The American Statistician 27 (1973): 17-21.*

*Breiman, L.. “CONSISTENCY FOR A SIMPLE MODEL OF RANDOM FORESTS.” (2004).*

*Boughorbel, S. et al. “Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric.” PLoS ONE 12 (2017).*