Univerzitet u Kragujevcu

Domaći zadatak

iz predmeta Sistem za podršku odlučivanju

Tema:

Teorijski izveštaj o regresionim algoritmima sa nadgledanim učenjem

mentori: Ognjen Pavić,
prof. dr. Tijana Geroski,
prof. dr. Nenad Filipović
student: Željko Simić 3vi/2023

Kragujevac 2024.

1 Uvod

1.1 Opis procesa učenja

Problem obuke obuhvata skupinu n uzoraka podataka i time se pokušava predvideti svojstvo nepoznatih podataka. Ako svaki uzorak ponaosob je više nego 1 unos, npr., multidimenzionalni unos, može se reći da ima više atributa ili features-a.[1]

Problemi obuke potpadaju u nekoliko kategorija:

- *Nenadgledajuće učenje (nije tema ovog rada)* - obuka podacima sastoji se u unosu više vektora x bez ikakvih uparenih ciljnih vrednosti. Cilj je otkriti pripadnost grupi po sličnosti naspram drugih vektora (klasterizacija), ili po raspodeli podataka unutar prostora ulaznih podataka (procena gustine), ili po projekciji podataka višedimenzionog prostora na manje dimenzioni shodnom za vizuelizaciju.
- *Delimično nadgledano učenje (nije tema ovog rada)[2]* - situacija gde skup podataka namenjen za obuku ima poneki uzorak koji nije uparen sa ciljanom vrednošću, gde se zarad adekvatnijeg previđanja boljom generalizacijom i distribucijom koriste ciljanom vrednošću nenaznačeni uzorci.
- *Podsticajuće učenje (nije tema ovog rada)[3]* - opšti okvir rada gde se vrši obuka *agenata* (tj. algoritma obuke) težeći da se obuče da vrše akcije u zadatom *okruženju* (tj. problema za rešavanje) zarad najboljeg mogućeg slučaja nagrađivanja.
- ***Nadgledano učenje*** - gde se skup podataka sa prikazanim primerom na slici 1.[4] prilaže sa dodatnim atributima po kojim će se predviđanje usmeravati (klasa za svaki od uzoraka, labela). Sastoji se od 2 načina rada (gde samo jednog obrađujemo):
  - *Klasifikacija (nije tema rada)* - uzorci mogu da pripadaju dvema (binarna klasifikacija) ili više klasa (višeklasna klasifikacija) i želimo ustanoviti kako bi se predviđanje klasa za nenaznačene podatke vršilo uz pomoć već klasama naznačenih podataka. Način da se sagleda klasifikacija je kao diskretni (suprotno kontinualnom) oblik nadgledane obuke gde od ograničenog broja kategorija za svaki od n uzoraka se namerava dodeliti tačna klasa ili kategorija.
  - ***Regresija*** - gde željeni izlaz u priloženom skupu podataka se izražava kao jedna ili više kontinualnih promenljivih, a zadatak predviđanja da nepoznatim ulazima uzorka se naziva regresijom.

1.1.1 Trening skup i test skup

Mašinsko učenje se sačinjava od poduhvata gde se vrši obuka uz svojstva jednog skupa podataka, a vrši testiranje svojstvima uz pomoć nekih drugih skupova podataka. Uobičajena praksa u mašinskom učenju je da se vrši evaluacija algoritma uz podelu skupa podataka na 2 dela. Gde je jedan deo namenjen za obuku (trening skup), a drugi za testiranje (test skup).

1.1.2 Multiclass vs. multilabel obuka

Kada se koriste multiklasni klasifikatori, zadaci obuka i predviđanja se izvršavaju tako da zavise od formata podataka ciljanih vrednosti. 1D niz višeklasna labela skupljenih svih uzoraka

age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	y
59	2	32.1	101.00	157	93.2	38.0	4.00	4.8598	87	151
48	1	21.6	87.00	183	103.2	70.0	3.00	3.8918	69	75
72	2	30.5	93.00	156	93.6	41.0	4.00	4.6728	85	141
24	1	25.3	84.00	198	131.4	40.0	5.00	4.8903	89	206
50	1	23.0	101.00	192	125.4	52.0	4.00	4.2905	80	135
23	1	22.6	89.00	139	64.8	61.0	2.00	4.1897	68	97
36	2	22.0	90.00	160	99.6	50.0	3.00	3.9512	82	138
66	2	26.2	114.00	255	185.0	56.0	4.55	4.2485	92	63
60	2	32.1	83.00	179	119.4	42.0	4.00	4.4773	94	110
29	1	30.0	85.00	180	93.4	43.0	4.00	5.3845	88	310
22	1	18.6	97.00	114	57.6	46.0	2.00	3.9512	83	101
56	2	28.0	85.00	184	144.8	32.0	6.00	3.5835	77	69
53	1	23.7	92.00	186	109.2	62.0	3.00	4.3041	81	179
50	2	26.2	97.00	186	105.4	49.0	4.00	5.0626	88	185
61	1	24.0	91.00	202	115.4	72.0	3.00	4.2905	73	118
34	2	24.7	118.00	254	184.2	39.0	7.00	5.0370	81	171
47	1	30.3	109.00	207	100.2	70.0	3.00	5.2149	98	166
68	2	27.5	111.00	214	147.0	39.0	5.00	4.9416	91	144
38	1	25.4	84.00	162	103.0	42.0	4.00	4.4427	87	97
41	1	24.7	83.00	187	108.2	60.0	3.00	4.5433	78	168

Slika 1: Primer skupa podataka - Diabetis, sa atributima `age,sex,bmi,bp,s1,s2,s3,s4,s5,s6` i targetom `y` u posebnih 442 uzoraka.

je moguće navesti da radi *multiklasna predviđanja*. Ciljane vrednosti je moguće mapirati tako da se konvertuju u binarni zapis, tj. niz sastojan od binarnih cifara tako da skup uzoraka ima ciljane vrednosti kao 2d niz i nakon obuke, za obavljana predviđanja smatra se da su *multilabel predviđanja*. Takođe je moguće da ciljana vrednost ima niz više labela (skupljeno po svim uzorcima ciljana vrednost biva 2d niz) za svaki uzorak pri obuci i kasnije naspram toga vršiti predviđanja.

1.2 Popularni algoritmi za regresiju

- **Jednostavna linearna regresija (SLR)** - statistički metod koji pomaže pri obuci nad srodnostima među dveju kvantitativnih nezavisnih promenljivih (ulazne `x` i izlazne `y`).
- **Višestruka linearna regresija (MLR)** - statistički metod kojim se koristi više raznolikih nekategoričkih nezavisnih promenljivih zarad predviđanja ishoda kontinualne, tj- nekategoričke zavisne promenljive.
- **Polinomijalna linearna regresija (PLR)** - specijalan slučaj MLR-a, obučavanja vrši nad nelinearnim srodnostima među nezavisnim ulaznim i zavisnim izlaznim vrednostima, a oslovljavanje "linearnom" je zbog linearnog odnosa koeficijenata **b** (koji su faktor evaluacije uz ulaznu **x** promenljivu).
- **Regresija stabla odlučivanja** - cepaju podatke u manje podskupove donesenim odlukama različitim upitima, u isto vreme stablo inkrementalno je razvijeno što ishoduje konačnim čvorovima odluke, ali i čvorovima listova. Radi i sa kategoričkim vrednostima podataka.
- **Regresija slučajne šume** - obuka ansambla koja je moćna tehnika pri unapređivanju modela, gde obuka ansambla podrazumeva kombinovanjem višestrukih evaluacija algoritama zarad formiranja veštog modela optimalnog predviđanja. Zadatak je da regresijom

niza više modela stabala odluka kao osnova budu izgrađeni, a pritom obuka bude obavljena agregacijama bootstrapping/bagging-a.

- **Regresija potpornim vektorima (SVR)** - drugačiji oblik mašina potpornih vektora (SVM) i mogu obavljati analize (ne)linearnih regresija na kontinualim vrednostima podataka umesto klasifikacije, koristeći se konceptima hiperravni i granične linije. Naspram drugih algoritama koji minimizuju stopu gubitka, ovi uklapaju gubitak po nekom kriterijumu praga.[5]

U sledećim sekcijama biće obrađivani pomenuti algoritmi.

2 Teorijske osnove i metodologija

2.1 Jednostavna linearna regresija (SLR)

Statistički metod koji pomaže pri obuci nad srodnostima među kvantitativnih promenljivih nezavisnih i zavisnih (ulazne `x` i izlazne `y`).

Primer primene SLR-a uključuje procenu plata, stanja tržišta nekretnina, stanja finansijskog portfolija, itd. Pri proceni kontinualnog izlaza na osnovu jednog ulaza, gde je linearna korelacija vrednosti ulaza i izlaza uzoraka, kao što je prikazano na slici 2. Što je veća vrednost ulaza, veća vrednost izlaza. Uzimaju se u obzir parametri:

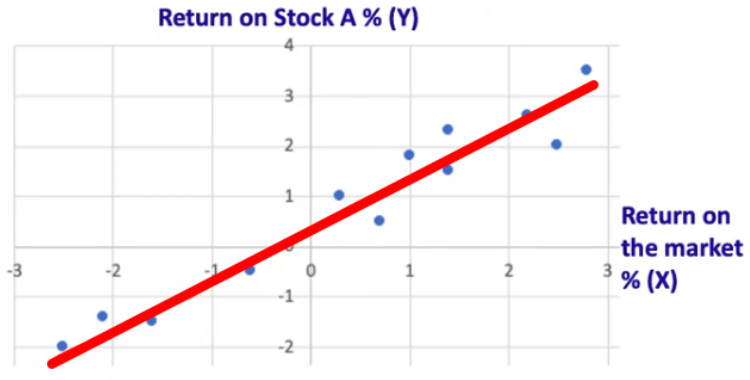
- m - broj uzoraka trening skupa
- x - ulazna promenljiva / feature / nezavisna promenljiva
- y - izlazna promenljiva / target / zavisna promenljiva
- (x, y) - opšti oblik uzorka trening podataka
- (x_i, y_i) - i-ti uzorak trening podataka
- $y_{predviđenog} - y_i$ - gubitak

Potom se uzima hipoteza SLR-a:

$$y = a * x + b$$

Target promenljiva y je linearna funkcija promeljive feature-a x i tako je ovo **univarijantna linearna regresija**.

Market return (%)	Stock A return (%)
2.2	2.6
1	1.8
-2.1	-1.4
-1.6	-1.5
1.4	2.3
0.7	0.5
-0.6	-0.5
2.5	2
1.4	1.5
-2.5	-2
0.3	1
2.8	3.5



Slika 2: Primer skupa podataka - berzanskih akcija, sa atributom `tržišni povrat na ulog` i targetom `akcijski povrat na ulog`, pa time određena linija linearne regresije.

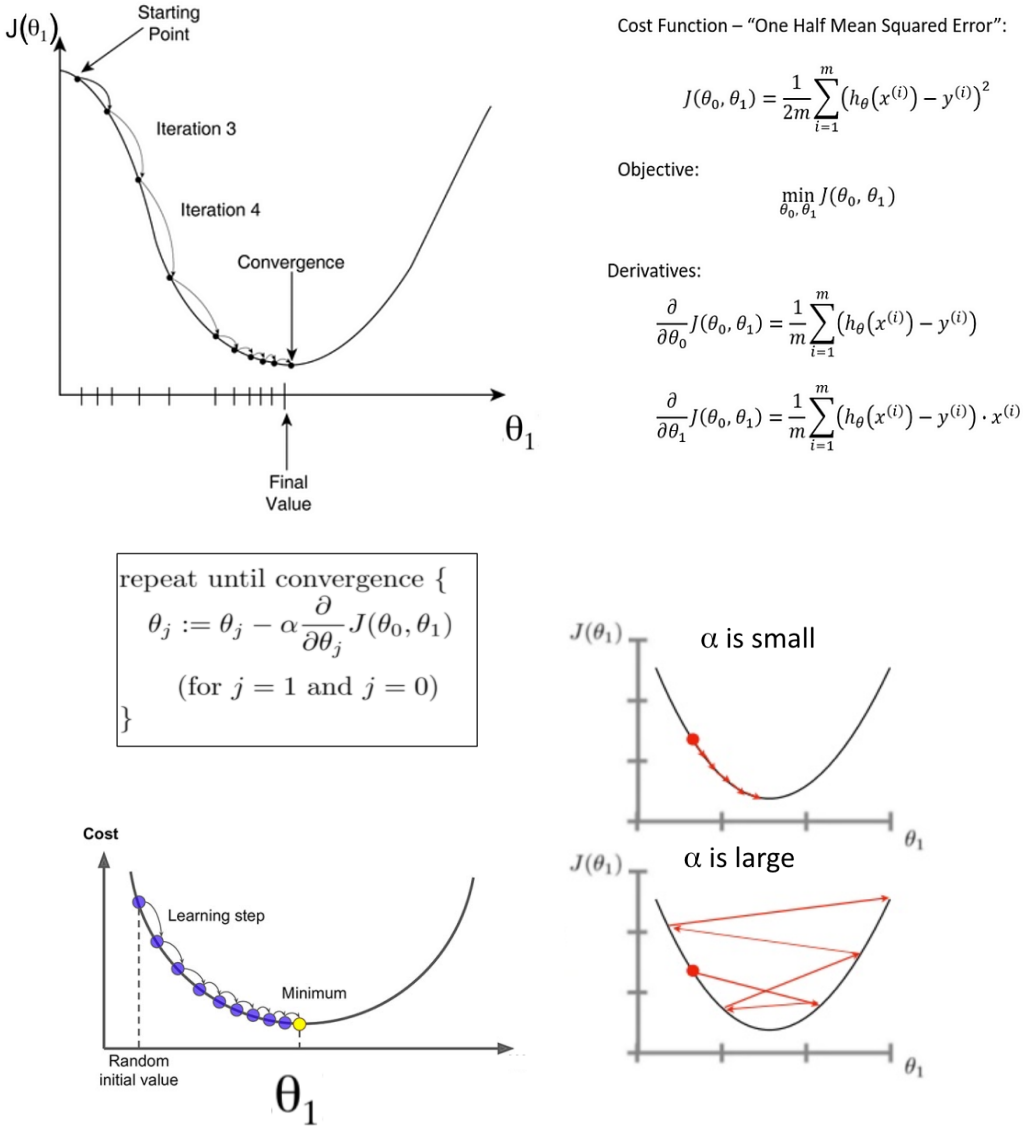
2.1.1 Funkcija ocenjivanja i gradijentni spust

Stremnja linearne regresije je da nađe najbolju moguću vrednost za a (težina parametara u modelu) i b (bias modela) u pomenutoj jednačini.[6] **funkcija ocene** pomoći će da nađemo najbolji ishod tog problema i pritom uklopimo najbolju liniju linearne regresije za uzorke. Ovaj

problem se dovodi do potrebe za rešavanjem *problema minimalizacije* gde se minimalizuju gubici između predviđenih i tačnih vrednosti, tj. sa a i b će se *minimalizovati prosečna suma kvadrata grešaka*:

$$J_{(a,b)} = \frac{1}{2m}\sum_{i=1}^{m}(y_{predvi\bar{d}enog} - y_i)^2$$

Ovo je poznata *funkcija ocene kvadratnih grešaka*, iliti, *funkcija središnjih kvadratnih grešaka* koja pruža prosečne kvadratne greške nad svim uzorcima skupa podataka. *Gradijentni spust* je metod ažuriranja a i b zarad minimalizacije funkcije ocene $J(a, b)$. Počevši sa a i b postepeno umanjuje se funkcija ocenjivanja korišćenjem diskretnih količina koraka koja je takođe *stope obuke (learning rate) u gradijentnom spustu*. Ovim će biti odlučujuće koliko hipoteza je *brzo konvergira optimalnom minimumu*, prikazano na slici 3. U pythonu biblioteka koja je predodređena za ove `from sklearn.linear_model import LinearRegression`.

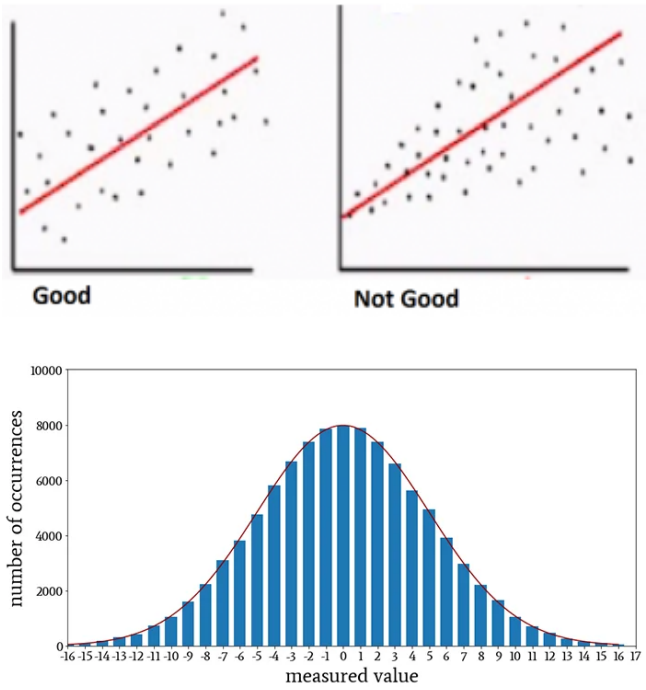


Slika 3: Određivanje funkcije ocenjivanja, kao i algoritamski postupak za slučaj kada je stopa obuke α veća ili mala.

2.2 Višestruka linearna regresija (MLR)

Ovo je najuobičajeniji oblik analize linearnom regresijom. Statistički metod kojim se koristi više raznolikih nekategoričkih nezavisnih promenljivih (features-a) zarad predviđanja ishoda kontinualne, tj- nekategoričke zavisne promenljive (target-a). Ako su features-i tipa kategoričkih vrednosti tada je neophodno pretvoriti ih u kontinualne promenljive pre korišćenja MLR-a. Stremi se da se nađe linearan odnos medju 2 ili više features-a (oslovljenim kao *prediktor promenljivama*, tj. *regresorima*) i jednom target promenljivom (tzv. *regresandima*). MLR pretpostavke:

1. Odstojanja od regresije bi trebalo biti **normalno distribuirane** (Gausovom distribucijom, središnja vrednost, specijalan slučaj očekivane vrednosti funkcije distribucije verovatnoća)[8][7][9], **homoscendastična** (središnja, tj. očekivana vrednost slučajne promenljive ima *istu* disperziju, tj. varijansu) i **aproksimirana u pravougaoni oblik**, prikazano na slici 4.



Slika 4: Homoscendastičnost ostvarena i neostvarena. Gausova distribucija, aproksimirana u pravougaoni oblik.

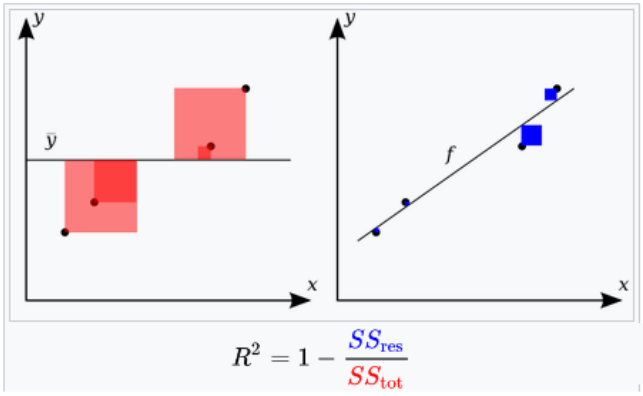
2. Linearan odnos se nagađa među promenljivama features-a i target promenljivoj.
3. Izostajanje višestruke kolinearnosti koja označava da features-i nisu blisko srodni međusobno.[10]
4. Dodavanje previše features-a će uvećati količinu *objašnjavajuće disperzije*

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

po promenljivi targeta (takođe postoji R2 ocena, koeficijent determinizma,

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

gde $\hat{y}_i$ je predviđena target vrednost, a $\bar{y}$ je očekivana vrednost, koja ukazuje na doslednost obuke i mera je koliko neuvaženih uzoraka će biti predviđeno modelom i predstavljena je na slici 5.)[11][12][13] i rezultovaće da model ode u stanje overfitting-a.

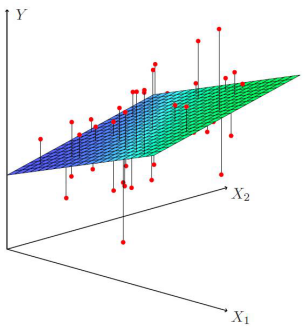


Slika 5: R2 skor vizuelizacija uticaja postupka.

MLR koristi se:

1. pri identifikaciji doprinosa features promenljivih na target promenljivu;
2. pri prognoziranju promene kod target promenljive sa promenama promljenivih features-a;
3. pri predviđanju trendova i budućih vrednosti na tržištu.

Formula MLR-a je $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$, gde postoji $n+1$ features promenljivih, y target promenljiva, $x_1, x_2, x_3, \ldots, x_n$ features-i, b_0 je odsečak na y-osi, $b_1, b_2, b_3, \ldots, b_n$ koeficijenti nagiba za svaku promenljivu features-a, kao na slici 6. MLR je primenjen da odredi



Slika 6: Višestruka linearna regresija.

linearni matematički odnos među nekolicini slučajnih promenljivih zarad formiranja linije prave koja najbolje aproksimira sve tačke uzoraka skupa podataka u multidimenzionalnom prostoru.

Za implementaciju u python-u koriste se moduli:

- `from sklearn.preprocessing import LabelEncoder, OneHotEncoder` za enkodiranje labele promenljive targeta i promenljivih features-a,
- a regresija je obavljena uz modul `from sklearn.linear_model import LinearRegression`.

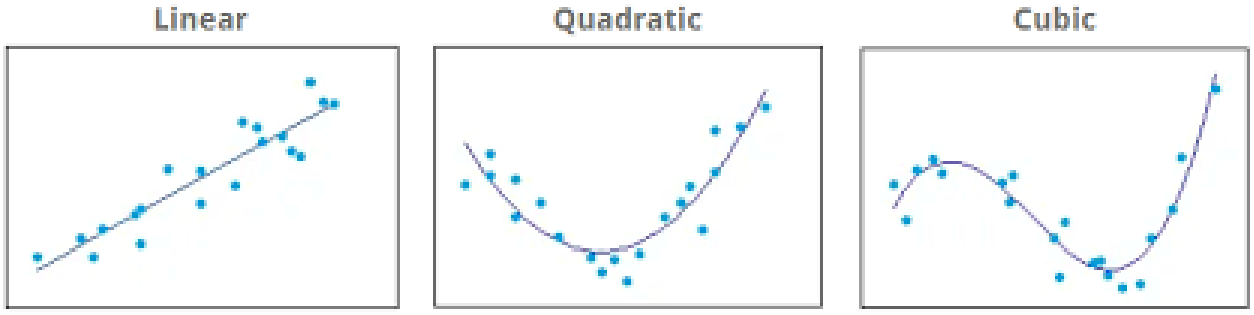
2.3 Polinomijalna linearna regresija (PLR)

Specijalan slučaj MLR-a, obučavanja vrši nad nelinearnim srodnostima među nezavisnim ulaznim i zavisnim izlaznim vrednostima, a oslovljavanje "linearnom" je zbog linearnog odnosa koeficijenata **b** (koji su faktor evaluacije uz **x**). Target promenljiva y je definisana kao n-ti stepen polinom po feature promenljivoj x. Teži se da se obuči tačkama uzoraka model sa najboljim mogućim vrednostima koeficijenata.

Hipoteza PLR-a je data formulom

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 + \cdots + b_n x_1^n,$$

gde je y target promenljiva, $x_1, x_1^2, x_1^3, \ldots, x_1^n$ feature promenljiva za član svakog stepena, $b_0, b_1, b_2, \ldots, b_n$ koeficijenti promenljive feature-a za član svakog stepena.



Slika 7: PLR za linearni, kvadtni i kubni slučaj.

PLR je korišćen kada linija prave iz SLR ili MLR nije skladna pri obuci tačaka uzoraka skupa podataka i potrebna je parabolička kriva 2. reda pri takvoj obuci. Polinomijalni član stepena 2. (kvadratnog) ili 3. (kubnog) pretvaraju linearni regresioni model u krivu 2. reda i slika 7. to prikazuje. Pri sagledanju tačaka uzoraka skupa podataka promenljivom features-a i target-a koje su razbacane ili u odnosu krivolinijskom najbolje je koristiti PLR s obzirom da na takvom tipu podataka time će se rezultovati više negativnim i pozitivnim odstojanjima.

Moguće je naići na problem overfittinga pri određivanju regresije uz povećavanje stepena polinomijalne regresije zarad dostizanja sve bolji ishod modela.[14] Tako po nastalom overfittingu nove tačke uzoraka skupa podataka ne bivaju obrađene.

Ovim povodom pri regresiji biće pokušaja da se penalizuju težine (koeficijenti) modela zarad regularizacije efekta problema overfitting-a. Tehnike regularizacije koje su dostupne kao metodologije su *Lasso regresija* i *Ridge regresija*.

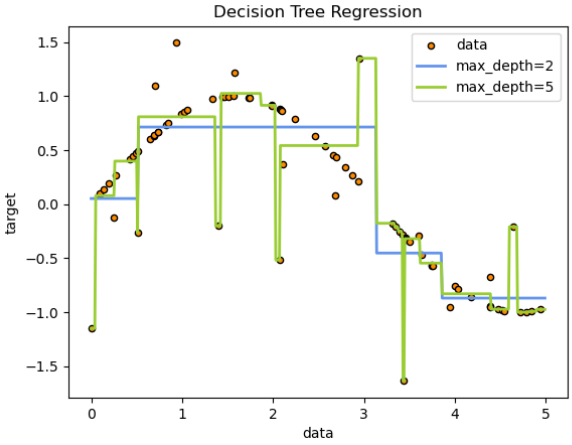
Generalizacija pristupa odlučivanja da li će se dati prednost bias-u ili disperziji da se zaobiđu problem underfitting-a i overfitting-a kojeg obrađujemo uz selekciju adekvatne vrednosti za stepen polinoma po kom se obuka nad podacima vrši. Stepen polinoma koji je povećan nakon početka rasta određenog nivoa razmaka trening i validacionih metrika.

U python-u modul koji se koristi za ovaj model je:

- `from sklearn.preprocessing import PolynomialFeatures`,
- `from sklearn.linear_model import LinearRegression`.

2.4 Regresija stabla odlučivanja

Korišćena da previđaju target vrednost po uzetim za obuku pravilima odlučivanja naspram features-a. Stablo odlučivanja pri vršenju regresije nad modelom gradi strukturu stabla, kao što je prikazano na slici 8. Cepaju podatke u manje podskupove donesenim odlukama različitim upitima, u isto vreme stablo inkrementalno je razvijeno što ishoduje konačnim *čvorovi odluke*, ali i *čvorovi listova*. Radi i sa kategoričkim vrednostima podataka. Ovo je tehnika koja je nelinearni, nekontinualni regresioni model.



Slika 8: Regresija stablom odlučivanja.

Za čvor odluke koji može imati 2 grane gde svaka ističe testirane vrednosti atributa. Čvor lista ističe odluku na numeričku vrednost target promenljive. Najuzvišeniji čvor odluke u stablu ističe najboljeg predviđivača koji je tzv. *čvor korena* (prvi roditelj).

Stablo odlučivanja je konstruisano procesom *rekurzivnog particionisanja* počev od čvora korena. Svaki čvor može biti deljen naspram levog i desnog čvora potomka koji mogu se dalje i oni deliti. Ovi čvorovi će postati čvorovi roditelji za njihove ishodujuće potomke čvorove. Cilj ove procedure je da ispuni kriterijum da minimizuje ustanovljene lokacije za buduće podele koji je *središnja kvadratna greška (MSE, L2 greška), Poasonovu devijaciju*, *središnja apsolutna greška (MAE ili L1 greška)*.[15] MSE i Poasonova devijacija obe predviđaju vrednost čvorova listova za ustanovljenu središnju vrednost $\bar{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y$ za čvor m, gde je za MAE predviđena vrednost čvora lista na medianu $median(y)_m$.

- MSE se računa kao: $H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2$,
- Prepolovljena Poasonova devijacija kao: $H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y \log \frac{y}{\bar{y}_m} - y + \bar{y}_m)$. Poasonova devijacija je pogodna ako je target frekvencija ili količina, ali, u svakom slučaju, $y >= 0$ je uslov da bi se ona koristila kao kriterijum. Obučava sporije od MSE.
- $H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} |y - median(y)_m|$, gde je $median(y)_m = \underset{y \in Q_m}{\text{median}}(y)$. Obučava sporije od MSE.

Implementacija u python-u je vršena uz korišćenje modula:
`from sklearn.tree import DecisionTreeRegressor`.

2.5 Regresija slučajne šume

Obuka ansambla koja je moćna tehnika pri unapređivanju modela, gde obuka ansambla podrazumeva kombinovanjem višestrukih evaluacija algoritama zarad formiranja veštog modela optimalnog predviđanja.

Kao metod ansambla obavlja zadatke regresija kombinujući odluke iz niza više modela stabla odluka u osnovi. I njihovo matematičko tumačenje je da:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots + f_n(x),$$

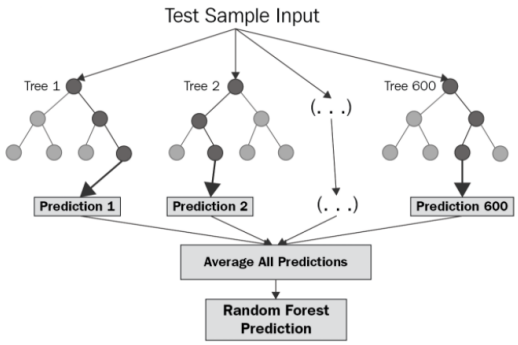
gde konačni ansamblov model regresije slučajnih šuma $g(x)$ je suma običnih modela stabala odlučivanja u osnovi $f_n(x)$. Posebno modeli stabla odlučivanja su konstruisani nezavisnim različitim poduzorcima trening skupa i ovaj proces treniranja svakog stabla odluke sa različitim uzorcima skupa podataka, gde je uzorkovanje urađeno sa zamenom, je poznato kao *Bagging/Bootstrap agregacija*.

Koristan je pri regresiji skupa podataka sa numeričkim i kategoričkim features-ima naspram drugih regresija. Za razliku od ostalih linearnih regresionih modela, regresije slučajnih šuma mogu da obuhvate nelinearnu interakciju među promenljivama features-a i promeljive target-a.

Ne radi efikasno sa raspršenim vrednosti promenljivih features-a, obično kategoričkog tipa višedimenziono. Neophodna je ili predobrada nad features-ima da se generišu numeričke vrednosti, ili primena linearnog modela.

Za izgradnju modela za regresiju slučajnih šuma (kao na slici 9.) prate se koraci:

1. Odabrati nasumičnih n uzoraka trening skupa;
2. Gradi se u osnovi model stabla odlučivanja uvaženih za n uzoraka skupa podataka;
3. Odabrati nekih N stabal odlučivanja koja će biti konstruisana;
4. Ponoviti korak 1. i korak 2. N puta;
5. Za novu tačku uzorka skupa podataka, napravi se N ustanovljenih predviđanja vrednosti od stabala odlučivanja i odrediti **prosečnu vrednost** svih vrednosti predviđenih izlaza nove tačke uzorka.



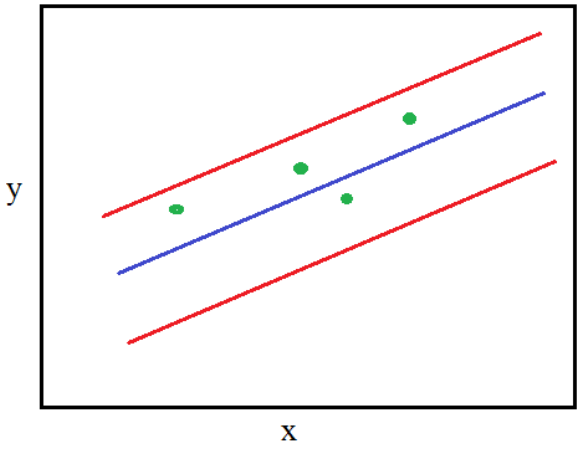
Slika 9: Regresija slučajnom šumom.

Za implementaciju korišćeni su moduli:

- `from sklearn.ensemble import RandomForestRegressor`
- `from sklearn.preprocessing import LabelEncoder`

2.6 Regresija potpornim vektorima (SVR)

Drugačiji oblik mašina potpornih vektora (SVM) i mogu obavljati analize (ne)linearnih regresija na kontinualim vrednostima podataka umesto klasifikacije, koristeći se konceptima **hiperravni** i **granične linije**, uzimajući u obzir dijagrama na slici 10. (hiperravan je plava, a granične linije su crvene). Naspram drugih algoritama koji minimizuju stopu gubitka, ovi uklapaju gubitak po nekom kriterijumu praga.



Slika 10: Regresija potpornim vektorom.

SVR podržava (ne)linearne regresije i tako pokušava obučiti sa što više uzoraka skupa podataka moguće koja je ostvarena usput plavim i crvenim linijama ograničavajući margine prekršavanja. SVR vrši regresiju u višedimenzionom prostoru i svaki posebno uzorak predstavlja svoju sopstvenu dimenziju. Širina među crvenim linijama i plavom linijom je kontrolisana je uz **hiperparametar** ϵ. Stremnja SVR modela je da uzme u obzir uzorke unutar oblasti između crvenih linija i najbolja obuka je ostvarena na plavoj hiperravni koja će imati maksimalni broj uzoraka.

Kada se vrši evaluacija kernela među uzorcima trening skupa i uzorcima test skupa, rezultujuća vrednost daje koordinatu uzorka test skupa u toj dimenziji. Vektor k proizveden kada tačka uzorka test skupa je evaluirana za sve tačke uzoraka trening skupa i tumači se kao tačka test skupa u višedimenzionim prostorima. Ovaj vektor može biti korišćen da obavi linearnu regresiju. Vektori najbliži tački test skupa su oslovljeni kao potporni vektori.

Da bi se obavila obuka SVR modela, moramo imati trening skup, koji pokriva oblast interesa i podstaknut je rešenjima u ovoj oblasti. Posao SVR-a je da aproksimira funkciju koju koristimo da generišemo obuku trening skupom. Nakon pribavljanja trening skupa, moramo odabrati **kernel**, njegove parametre i neki potrebna **regularizacija** C-om. Onda, moramo ostvariti **matricu korelacija** i obučiti model da dobavi **kontrakcione koeficijente**. Korišćenjem ovih koeficijenata, moguće je generisati poseban procenjivač.

Za implementaciju u python-u koristi se modul `from sklearn.svm import SVR`.

3 Diskusija o algoritmima

3.1 Prednosti i mane

3.1.1 Jednostavna linearna regresija

Prednosti jednostavne linearne regresije:[16]

- Linearna regresija je relativno jednostavan algoritam što ga čini lakim za razumevanje i implementiranje;
- Koeficijentima modela linearne regresije može biti protumačena promena u target promenljivoj za jednu jediničnu vrednosnu promenu u feature promenljivoj što daje uvid u korelaciju među promenljivama;
- Linearna regresija je računski efikasna i može rukovati ogromnim skupovima podataka efikasno;
- Može biti obučena brzo na velikim skupovima podataka, što je čini adekvatnom za primene u realnom vremenu;
- Linearna regresija je relativno robustna što se tiče outliers-a u poređenju sa algoritmima mašinskog učenja. Outliers-i mogu imati manji uticaj na sveukupne performase modela;
- Linearna regresija često služi dobrim baseline model (bez ikakvih inicijalnih konfiguracija parametara) za poređenje više složenijih algoritama mašinskog učenja;
- Linearna regresija je dobro ustanovljen algoritam sa bogatom istorijom i široko je dostupna u raznovrsnim softverskim bibliotekama mašinskog učenja.

Mane jednostavne linearne regresije:

- Linearna agresija nagađa linearnu srodnost između target i feature promenljivih. Ako odnos je nelinearan, model se neće dobro pokazati pri obavljanju posla;
- Linearna regresija nagađa da feature je i inače u pogodnom obliku za model;
- Neophodno je neko prilagođavanje feature-a tako da se obavlja preoblikovanje feature-a u format koji može biti efektivno korišćen od modela;
- Linearna regresija je podložna da ode u stanje overffiting-a i underfitting-a. Overfitting nastaje kada model uči iz trening skupa toliko detaljno da ne uspeva da izopšti nesagledane podatke. Underfitting nastaje kada model je jednostavan da obuhvati uspostavljene odnose u skupu podataka;
- Linearna regresija ustupa ograničenu moć pojašnjavanja složenih odnosa među promenljivama. Što su tehnike mašinskog učenja naprednije to su neophodni produbljeni uviđaji.

3.1.2 Višestruka linearna regresija

Prednosti višestruka linearna regresija:[17]

- Mogućnost zaključivanja uticaja 1 ili više feature promenljivih na target vrednost.
- Identifikuju outliers-e ili anomalije.

Mane višestruke linearne regresije:

- Linearna regresija je osetljiva na multikolinearnost, što se ističe kada je visoka korelacija među feature promenljivama;
- Multikolinearnost može uvećati disperziju koeficijenata i dovesti do nestabilnih predviđanja modela;
- Mogućnost loših performansi pri nedostajućim vrednostima.

3.1.3 Polinomijalna linearna regresija

Prednosti polinomijalna linearna regresija:[18]

- Širok opseg funkcija može biti korišćen obuku;
- Obično obuku vrši nad širokim opsegom krivih 2. reda;
- Prilaže najbolje aproksimacije srodnosti među target i nezavisnim promenljivama.

Mane višestruke linearne regresije:

- Preosetljive na outliers-e;
- Prisutnost 1-2 outliers-a u skupu podataka mogu ozbiljno uticati na rezultate nelinearne analize;
- Pristuno je mali broj alata za validaciju za detekciju outliers-a u nelinearnoj regresiji nego što ih ima za linearnu regresiju.

3.1.4 Regresija stabla odlučivanja

Prednosti regresija stabla odlučivanja:[19]

- Jasnost pravila odlučivanja bivaju učena od strane algoritma, omogućuju razumljivost i vizuelizaciju, pored toga ustupaju moć lakog pojašnjavanja nestručnim licima;
- Nelinearni odnosi između features-a i target promenljive koji su obuhvaćeni stablom odlučivanja; Za razliku od linearnih modela, stabla odlučivanja mogu reprezentovati složene granice odluke, njihovim uspostavljanjem prilagođeni su za skupove podataka sa složenijim obrascima.
- Skaliranje (normalizacija ili standardizacija) features-a nije potrebno pošto stabla odlučivanja nisu osetljiva na raspone features-a.

- Rukuje i sa numeričkim i kategoričkim podacima (bez ikakvih one-hot enkodiranja ili drugih tehnika predobrada). Ovo ih čini pogodnim za skupove podataka sa izmešanim tipovima podataka.
- Robustni na outliers-e u podacima. Počeviši od njihovog particionisanja feature prostora u regione zasnovane na vrednostima features-a, outliers-i streme da imaju minimalni uticaj na sveukupnu performansu modela.

Mane višestruke linearne regresije:

- Overfitting-u podložno, pogotovo pri produbljenosti stabla ili je skup podataka raspršen.
- Visok nivo disperzije je prisutan, jer male izmene u trening skupu mogu rezultovati značajno različitim stablima;
- Nestabilnost pošto je osetljivost velika za male promene u podacima, što vodi u različite podele i posledično drugačija stabla, a to ih čini manje pouzdanim naspram drugih algoritama;
- Bias naspram features-a, tj. za visok nivo (tj. visok broj uzoraka) može dovesti do pristrasnosti naspram features-a sa manjim nivoima u podelama stabla odlučivanja;
- Težina u prikupljanju linearnih srodnosti, iako je shodan uhvatiti nelinearne srodnosti, stablo odlučivanja ima ovu problematiku za srodnosti među features-ima i target promenljom.

3.1.5 Regresija slučajnih šuma

Prednosti regresija slučajnih šuma:[20]

- Lako je koristiti i manje je osetljiva naspram trening skupa;
- Tačniji za razliku od stabala odlučivanja;
- Efektivniji u baratanju ogromnim skupovima podataka koji imaju više atributa;
- Barataju nedostajućim podacima, outliers-ima, šumovima u features-ima.

Mane višestruke linearne regresije:

- Teško rastumačiv;
- Zahteva neka poznavanja u stučnoj oblasti zarad odabira odgovarajućih parametara kao što su: broj stabala odlučivanja, maksimalna dubina drveta, broj features-a za uzimanje u obzir svakom podelom;
- Računski je skup, pogotovo za ogromne skupove podataka;
- Ispašta po pitanju overfitting-a ako model je previše složen ili broj slučajnih stabala je previsok.

3.1.6 Regresija potpornim vektorima

Prednosti regresija potpornim vektorima:[21]

- Robustni na outliers-e;
- Model odlučivanja se lako ažurira;
- Odlična pogodnost generalizovanja, sa visokom tačnosti predviđanja;
- Laka za implementaciju.

Mane višestruke linearne regresije:

- Neprikladne za ogromne skupove podataka;
- Slučajevima gde broj features-a za svaki uzorak nadmašuje broj uzoraka trening skupa, rad SVR-a ispašta;
- Model odlučivanja ne obavlja baš valjan rad kada skup podataka ima šumova, tj. target klase se prepliću.

3.2 Sveukupni pregled svojstava algoritama

U tabeli 1. će biti dat neki ukupan pregled svih algoritama jedan naspram drugog.[22]

Model regresije	**Prednosti**	**Mane**
Jednostavna linearna regresija	• Lak za implementaciju;	• Ograničen na samo jednu feature promenljivu.
Višestruka linearna regresija	• Dobro radi bez obzira na veličinu skupa podataka; • Daje informacije o srodnosti features-a.	• Pristrasnosti linearne regresije.
Polinomijalna linearna regresija	• Dobro radi bez obzira na veličinu skupa podataka; • Obavlja dobar posao naspram nelinearnih problema.	• Moramo odabrati pravi polinomijalni stepen zarad dobrog odnosa bias-a naspram disperzije.
SVR	• Lako prilagodljiv; • Radi dobro sa nelinearnim problemima; • Nije bias-ovan naspram outliers-a.	• Zahteva da feature promenljiva skupa podataka bude skalirana; • Nije poznat; • Težak za razumevanje.
Regresija stabala odluka	• Jednostavno tumačenje; • Radi dobro sa nelinearnim problemima; • Nema potreba za primena skaliranja raspona feature-a.	• Loši rezultati na manjim skupovima podataka; • Overfitting može lako se ispoljiti.
Regresija sluičajnih šuma	• Vlada širokim rasponom mogućnosti; • Daje tačne rezultate; • Dobre performanse na velikom broju problema, uključujući nelinearne.	• Teško rastumačiv; • Overfitting može lako se ispoljiti. • Mora biti unapred biti dat broj stabala koji će biti korišćen.

Tabela 1: Celokupno poređenje po prednostima i manama.

4 Zaključak

Na početku su pokušana izlaganja objašnjenja za stukturu skupa podataka uz razjašćanjavanje (šta je učenje, šta se najčešće koristi, šta su zahtevi rada pri nadgledanom učenju) i postepeno razgraničenje pojmova sadržanih (uzorak, atribut/feature, labela po brojnosti, klasa po brojnosti, trening i test skup), načina obuke sve do klasifikacija.

U teorijskim osnovama i metodologijama redom su istaknute teme jednostavne, višestruke i polinomijalne linearne regresije, regresija stabla odlučivanja, slučajne šume, potpornih vektora uz pominjanje potrebnih alatki pri implementaciji u python okruženju.

Jednostavna linearna regresija u sebi sadrži koncepte rada sa kontinualnim vrednostima, linearnih korelacija, gubitka, hipoteze linearne fukcije kao jednačine prave koja je tzv. univarijantna linearna regresija. Zatim se pominju funkcija ocenjivanja za vrednosti težine parametara i bias-a, pa i gradijentni spust se pominje. Sagledaju se problemi minimalizacije gubitaka između tačnih i predviđenih vrednosti, kao i minimalizacija prosične sume kvadrata grešaka, što vodi u pojmove funkcije kvadratnih grešaka, funkcija središnjih kvadratnih grešaka i njeno postepeno umenjivanje gradijentnim spustom, uz neku stopu obuke što vodi u brzo konvergiranje prema optimalnom minimumu.

Višestruka linearna regresija radi sa nekategoričkim, ali i ranije kategoričkim, pa zatim konvertovanim u numeričke vrednostima zarad nalaženja kontinualnih izlaza. Nalazi linearan odnos za izlaz sa 2 ili više features-a. Koncepti koji se podrazumevaju i uzeti kao pretpostavljeno omogućeni su normalna distributivnost odstojanja uzoraka naspram regresije, homoscendastičnosti, oblika pravougaone aproksimiranosti, linearnog odnosa features-a i target-a, izostajanja višestruke kolinearnosti, većom dimezionalnošću uticaja na objašnjavajuće disperzije (R^2). MLR-om detektuju se doprinosi svakog od feature-a na target, prognoziraju se promene kod vrednosti targeta, predviđaju se trendovi. Rezultuje se nalaženjem odsečka i nagiba ovim algoritmom.

Polinomijalna linearna regresija obučavanja vrši nad nelinearnim odnosima između features-a i target-a, a linearna je, jer radi sa koeficijentima članova stepena polinoma hipoteze PLR-a koji su međusobno linearni, pritom težeći da vrednosti koeficijenata budu najbolje izvedene. Uočilo se da neprikladnost SLR i MLR obuka pri radu sa feature-a člana stepena polinoma koji gradi paraboličkі oblik regresije, ovo pogodno i pri raspršenosti uzoraka u skupu podataka. Pomenut je proces regulacije nastajanja overfitting-a i tehnika kojim je moguće zaobići ove nepovoljnosti. Navodi se proces generalizacije pristupa odlučivanja između bias-a ili disperzije pri malopre navedenih problema, kao i važnost odabira stepena polinoma u datim slučajevima obuke.

Regresijom stabla odlučivanja obuka se vrši pravilima odlučivanja, gradi se struktura stabla, cepaju se podaci u manje podskupove donesenim odlukama različitim upitima, inkrementalno simultano ishoduje konačim čvorovima odluke i listova. Pominje se pojam rekurzivnog particionisanja, minimalizacija zarad naspram MSE (L2), Poasonova devijacija, MAE (L1) greškama.

Regresija slučajne šume je obuka ansambla kojom se vrši kombinovanje višestrukih evalu-

acija algoritama (tj. modela stabala odlučivanja) sve do optimalnog predviđanja nezavisnim različitim poduzorcima trening skupa sve u čast bootstrap/bagging agregacija. Moguć rad i sa kategoričkim i sa numeričkim features-ima, i ističe se moć obuhvatanja nelinearnih interakcija među features-ima i target promenljivama. Naznačava se da nije pogodan za rad sa raspršenim skupovima podataka, i da mu je neophodna predobrada, određivanja broja stabala i broja uzoraka sa kojim će se baratati i kako je sveukupna procena prosečna vrednost skupljenih rezultata modela unutar ansambla.

Regresija potpornim vektorima se stara o analizama (ne)linearnih regresija na kontinualnim vrednostima. Pominju se koncepti hiperravni, granične linije i kako se oslanjaju na uklapanja gubitka po nekom kriterijumu praga. Navodi se ograničavanja margina prekršavanja, hiperparametra ϵ, slučaja najbolje ostvarene obuke na hiperravni zbog maksimalnog broja uzoraka, evaluacije kernela što rezultuje udeljivanjem koordinate uzorka test skupa u dimenziji i isticanjem vektora k. Data je ideja potpornih vektora, potreba za posedovanjem test skupova koji pokriva oblast interesa, ciljanog zadatka SVR-a da aproksimira funkciju, odabira kernela, parametara, ideja o regularizaciji.

U sekciji 3. se diskutuje prednostima i manama svakog algoritma ponaosob, uzimajući u obzir koncepte lakoće razumevanja, lakoće implementacija, uvida u korelaciju među promenljivama, računske efikasnosti, brzine rada na velikim skupovima podataka, robustnost po pitanju outliers-a, baseline stanja, implicitnog podrazumevanja nekih karakteristika, zahtevanja za predobradama, isticajnosti naspram overfitting-a i underfitting-a.

Literatura

[1] An introduction to machine learning with scikit-learn, https://scikit-learn.org/stable/tutorial/basic/tutorial.html, Datum poslednjeg pristupa: 14. april 2024.

[2] Semi-supervised learning, https://scikit-learn.org/stable/modules/semi_supervised.html, Datum poslednjeg pristupa: 14. april 2024.

[3] Introduction to RL and Deep Q Networks, https://www.tensorflow.org/agents/tutorials/0_intro_rl, Datum poslednjeg pristupa: 14. april 2024.

[4] Toy datasets: Diabetes dataset, https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset, Datum poslednjeg pristupa: 14. april 2024.

[5] D. Sumon, Machine Learning – Part 3 – Regression, https://www.sumondey.com/machine-learning-part-3-regression/, Datum poslednjeg pristupa: 14. april 2024.

[6] Types of Regression Techniques in ML, https://www.geeksforgeeks.org/types-of-regression-techniques/, Datum poslednjeg pristupa: 14. april 2024.

[7] Mean And Variance Of Random Variable, https://byjus.com/maths/mean-variance-random-variable/, Datum poslednjeg pristupa: 14. april 2024.

[8] 4.2 Mean or Expected Value and Standard Deviation, https://openstax.org/books/statistics/pages/4-2-mean-or-expected-value-and-standard-deviation, Datum poslednjeg pristupa: 14. april 2024.

[9] Normal Distribution: What It Is, Uses, and Formula, https://www.investopedia.com/terms/n/normaldistribution.asp, Datum poslednjeg pristupa: 14. april 2024.

[10] Multicollinearity: Meaning, Examples, and FAQs, https://www.investopedia.com/terms/m/multicollinearity.asp, Datum poslednjeg pristupa: 14. april 2024.

[11] 3.3.4.8. Explained variance score, https://scikit-learn.org/stable/modules/model_evaluation.html#explained-variance-score, Datum poslednjeg pristupa: 14. april 2024.

[12] 3.3.4.1. R^2 score, the coefficient of determination, https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score, Datum poslednjeg pristupa: 14. april 2024.

[13] Y Hat: Definition, https://www.statisticshowto.com/y-hat-definition/, Datum poslednjeg pristupa: 14. april 2024.

[14] Implementation of Polynomial Regression, https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/, Datum poslednjeg pristupa: 14. april 2024.

[15] 1.10.7.2. Regression criteria, https://scikit-learn.org/stable/modules/tree.html#regression-criteria, Datum poslednjeg pristupa: 14. april 2024.

[16] Linear Regression in Machine learning, https://www.geeksforgeeks.org/ml-linear-regression/, Datum poslednjeg pristupa: 14. april 2024.

[17] The Advantages & Disadvantages of a Multiple Regression Model, https://sciencing.com/difference-between-bivariate-multivariate-analyses-8667797.html, Datum poslednjeg pristupa: 14. april 2024.

[18] Implementation of Polynomial Regression, https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/, Datum poslednjeg pristupa: 14. april 2024.

[19] Pros and Cons of Decision Tree Regression in Machine Learning, https://www.geeksforgeeks.org/pros-and-cons-of-decision-tree-regression-in-machine-learning/, Datum poslednjeg pristupa: 14. april 2024.

[20] Random Forest Regression in Python, https://www.geeksforgeeks.org/random-forest-regression-in-python/, Datum poslednjeg pristupa: 14. april 2024.

[21] Unlocking the True Power of Support Vector Regression, https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0, Datum poslednjeg pristupa: 14. april 2024.

[22] Advantages and Disadvantages of different Regression models, https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-regression-models/, Datum poslednjeg pristupa: 14. april 2024.