Uvod

Ovaj master rad se bavi problemom nedostajućih vrednosti u podacima koji predstavljaju ulaz za algoritme mašinskog učenja. Ukoliko podaci koji služe za trening nisu potpuni, može se desiti da istreniran model nije dovoljno pouzdan i da kao rezultat donosi pogrešne zaključke.

Skup podataka koji će se koristiti u radu je pogodan za regresionu analizu. Pokazaće se kako različiti procenti

# 2. Експериментални подаци

## 2.1. Опис скупа података

Иако ће сам експеримент бити заступљен у другом делу рада, неопходно је да се сада објасни скуп података који ће се том приликом користити. Сви примери у раду ће се ослањати на овај скуп података (везани за екпсеримент или као помоћно средство за разумевање теорије)

Подаци се односе на пацијенте који болују од дијабетеса. . У табели (1) је визуелно представљен само подскуп података. Иницијални скуп података је садржао податке о 442 пацијента, што је за потребе овог рада подељено у два подскупа: први са 400 обсервација (тренинг подскуп) и други са 42 обсервације (тестни подскуп).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Старост | Пол | ИТМ\* | КП\*\* | Резултати серумских мерења | | | | | | резултат |
| Пацијент |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 59 | 2 | 32.1 | 101 | 157 | 93.2 | 38 | 4 | 4.9 | 87 | 151 |
| 2 | 48 | 1 | 21.6 | 87 | 183 | 103.2 | 70 | 3 | 3.9 | 69 | 75 |
| 3 | 72 | 2 | 30.5 | 93 | 156 | 93.6 | 41 | 4 | 4.7 | 85 | 141 |
| 4 | 24 | 1 | 25.3 | 84 | 198 | 131.4 | 40 | 5 | 4.9 | 89 | 206 |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 399 | 52 | 1 | 27.8 | 85 | 219 | 136 | 49 | 4 | 5.1 | 75 | 242 |
| 400 | 65 | 2 | 28.5 | 109 | 201 | 123 | 46 | 4 | 5.1 | 96 | 232 |

Табела Тренинг подскуп. Поред демографских података (старост и пол), на пацијентима је вршено 8 мерења који заједно утичу на ниво дијабетеса годину дана касније

\* ИТМ, индекс телесне масе

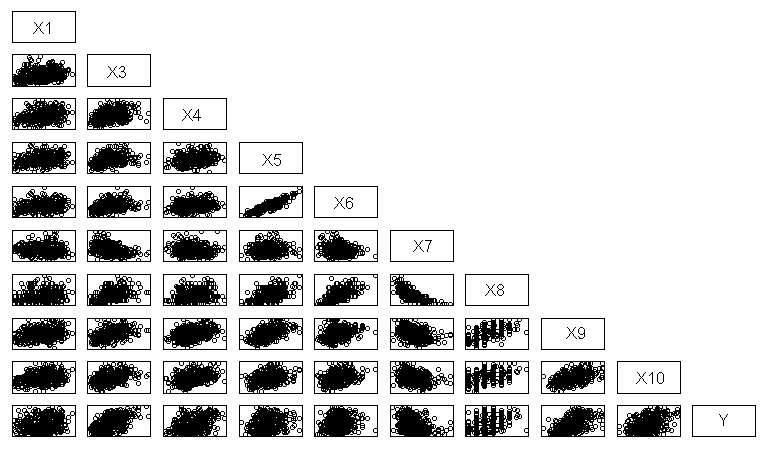
\*\* КП, крвни притисак

Скуп се састоји од 10 атрибута (варијабли) од који су 8 нумеричког типа () и једна номиналног карактера (). Резултат представља ниво дијабетеса код људи годину дана након мерења. [5] Такође, резултујућа колона је нумеричког карактера.

## 2.2. Корелациона матрица

Укључујући резултатску променљиву (колону), скуп података се сатоји од 10 колона нумеричког типа, што је одличан предуслов за анализу корелационе матрице. Корелациона матрица, са друге стране, може да укаже на линеаран однос између две колоне. Уколико је испостави да такве зависноти (лиенарне) постоје, у даљем раду ће линеарна регресија бити један од затупљенијих алгоритама машинског учења.

Свакако, пре даље анализе, одличан показатељ било које анализе може да донесе визуелизација самог скупа података (Слика 1).



Слика Визуелни приказ скупа података

Са слике 1 се јасно види да неке променљиве имају изражену линеарну зависност, што је добар показатељ да корелациону матрицу треба креирати и да је могуће из ње извући одређене закључке. Приметимо да променљива није део графичког приказа јер номинални тип податка те променљиве није идеалан за визуелизацију.

За креирање корелационе матрице, потребно је израчунати 45 различитих вредности или:

|  |  |
| --- | --- |
|  | (1.1) |

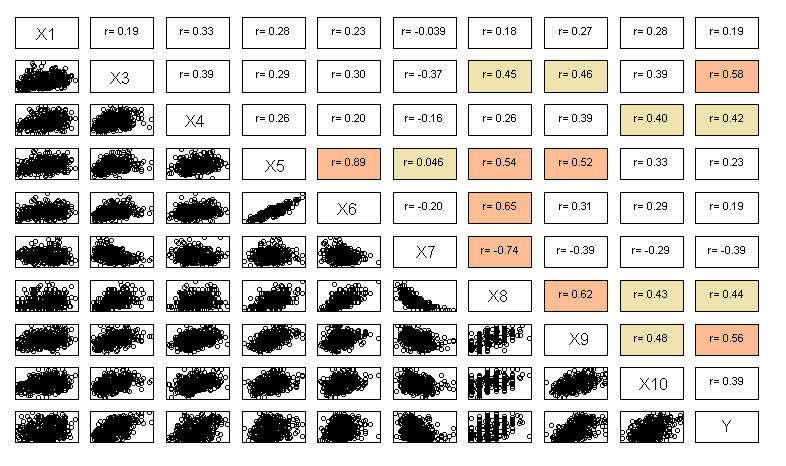
У једначини (1.1), представља број различитих вредности, док је укупан број променљивих у скупу података. У овом случају , јер колона пол није узета у обзир.

Корелација између две променљиве се рачуна по формули:

|  |  |
| --- | --- |
|  | (1.2) |

|  |  |
| --- | --- |
|  | (1.3) |

У једначинама изнад представља вредност корелације за променљиве и . Додатно коваријанса, , чија вредност не говори ништа о степену зависноти између и , се користи приликом рачунања корелације. Са друге стране корелација има вредности у опсегу , где екстремне врдности означавају јаку корелацију (јаку зависност), а вредности блиске нули значе да таква зависнот уопште не постоји.



Слика Визуелни приказ скупа података са корелационим вредностима

На слици 2 су приказане корелационе вредности где су поља у којима је апсолутна вредност корелације изнад 0.4 осенчена, тако да се лакше могу уочити променљиве које имају израженију међузависност.

По мом мишљењу, овакав графички приказ заједно са вредностима корелационе матрице су добар предуслов за регресиону анализу.

# 3. Регресиона анализа

## 3.1. Линеарна регресија

Регресиона анализа коцептуално представља једноставан метод проналажења функционалних зависности између променњљивих. [4] Та зависност је приказана у облику формуле у коме се са једне стране налази зависна променљива, а са друге стране скуп независних променљивих. Полази се од претпоставке да вредности независних променљивих утичу на вредности зависних променљивих.

Означимо ли зависну променљиву са *Y*, а остале променљиве *X1*, *X2*, … , *Xp* линеарну регресију можемо представити једначином.

|  |  |
| --- | --- |
|  | (2.1) |

Ознаком се означава грешка апроксимације. Уколико са означимо апроксимирану вредност (2.2), једначина (2.1) постаје:

|  |  |
| --- | --- |
|  | (2.2) |
|  | (2.3) |

Из једначине (2.3) је јасно да грешка представља разлику између очекиване и апроксимиране вредности, и пожељно је да та разлика буде што ближа нули[[1]](#footnote-1).

Овај рад ће се фокусирати на линерану регресију, и због тога се (2.1) се представља:

|  |  |
| --- | --- |
|  | (2.4) |

Или векторски:

|  |  |
| --- | --- |
|  | (2.5) |
|  | (2.6) |

Проналажењем параметара вектора , таквих да је вредност минимална, одређује се зависност између и . Величина вектора и је где представља број независних промељивих. Разлог за додавање вредности у вектор је једноставан. Како параметри тог вектора одређују апрксимирајућу функцију, она би без вредности засигурно пролазила кроз координатни почетак. Да би вредност увек била присутна, додата је вредност у вектор , и као што је приказано у једначини (1.4), она увек има исту вредност.

## 3.2. Примена

Линеарну регресију из претходног поглавља применићемо на скупу података описаном раније.

Као што је напоменуто раније, скуп података се састоји од 10 улаза и излаза (ниво дијабатеса). Другим речима, линеарна регресија би требало да процени ниво дијабетеса годину дана након мерења. За ту потребу, пронађени су параметри коришћењем тренинг подскупа, и затим је цео модел проверен тестним подскупом. Резултати се налазе у табели 2.

|  |  |
| --- | --- |
|  | Вредност |
|  | -321.14 |
|  | 0.02 |
|  | -22.73 |
|  | 5.62 |
|  | 1.03 |
|  | -1.03 |
|  | 0.70 |
|  | 0.31 |
|  | 6.85 |
|  | 64.40 |
|  | 0.37 |
| Средња квадратна грешка (MSE) | 1668.71 |
| Корен средње квадратне грешке | 40.85 |

Табела Резултати линеарне регресије

Горњи део табеле 2 показује све вредности вектора . Очигледно је да је параметар додат са разлогом јер има вредност другачију од нуле, што значи да има утицаја на предвиђењу вредност. Доњи део табеле показује колико истренирани модел добро предвиђа.

|  |  |
| --- | --- |
|  | (1.5) |

Разлика између предвиђене вредности и тражене вредности се квадрира, а затим се израчуна просечна вредност такве разлике међу свим пацијентима. Добијена вредност се зове средња квадратна грешка (MSE). Додатно се рачуна и корен ове грешке јер он даје бољу слику колико модел греши. У овом случају, модел приликом процене нивоа дијабетеса за следећу годину греши за око 40 јединица. Важно је приметити да средња квадратна решка не даје никакве информације о знаку, тј. не зна се да ли модел предвиђа више вредности или ниже од очекиваних.

Овај рад се неће даље бавити линераном регресијом, али ће ови резултати служити као мерило (референтне вредности) у будућим експериментима.

# 4. Недостајуће вредности

## 4.1. Механизми недостајућих вредности

Технике машинског учења као сто су надгледано и ненадгледано учење се могу посматрати као системи у ком су улази представљени као подаци, а излази су истренирани модели (алгоритми). Из овог угла посматрања, квалитет података директно утиче на каснију тачност предвиђања истренираног алгоритма. Међутим, улазни подаци често нису комплетни и као такви често онемогућавају тренинг алгоритама.

Некомплетан скуп података се означава и као скуп података са недостајућим вредностима. У таквим скуповима вредности недостају по једном од три могућа механизма, а механизам представља математички однос између података и недостастка. [1]. Механизми могу бити:

**MCAR** – Missing Completely At Random – Nедостајуће вредности су присутне без икакве законитости. Уколико скуп података има 2 променљиве (колоне), непостојаност податка у првој колони нема никакву повезаност са вредностима из обе колоне. Овај случај је веома чест, обзиром да углавном настаје људском ненамерном грешком. На пример, испитаник је случајно превидео одређено питање и оно је остало неодговорено. [3]

**MAR** – Missing At Random – недостајуће вредности присутне унутар једне променљиве немају никакву законитост (повезаност) са том променљивом. Уколико посматрамо исти скуп података (2 колоне), непостојаност податка у првој колони не зависи од вредности те колоне, али зависи вредности из друге колоне(а). На пример, прва колона садржи податке о просечној оцени током студија, а друга колона резултате теста приликом запослења. Испитаници (редови у скупу података) са ниском просечном оценом неће бити ни узети у разматрање, па њихова оцена са теста је ирелевентана, и не садржи се у скупу података.

**MNAR** – Missing Not At Random – недостајуће вредности једне променљиве су директно зависне од посматране променљиве. У скупу са две колоне, недостајуће вредности прве колоне недостају због ње саме, и немају никакве повезаности са другом колоном. На примеру скупа података који садржи резултате теста као променљиву (назив колоне), подаци те колоне могу да недостају у свим редовима где је резултат теста мањи од одређене оцене.

За потребе експеримената у овом раду посматраће се искључиво MCAR механизам недостајућих вредности. Разлог за ту одлуку је могућност добијања недостајућих вредности синтетичким путем. Скуп података из 1.2 (Примена регресионе анализе) ће бити ''пробушен'' више пута насумично и том приликом ће проценат недостајућих вредности бити различит. На тај начин од првобитног комплетног скупа података добиће се више некомплетних, и као додатна погодност знаће се иницијални вредности (касније ће се те почетне вредности користити за евалуацију технике уметања података).

## 4.2. Технике уметања података

У случају недостајућих података, понекад је најлакше одбацити обсервације које нису потпуне. На пример, уколико скуп садржи 100 обсервација (редова) и 10 атрибута (колона) , и од тога 10 различитих редова има тачно једну недостајућу вредност (тачно једну једну колону непопуњену), овом јендноставном техником остало би 90 редова (инстанци) као улаз за алгоритам машинског учења.

Описан скуп података се састоји од 100 редова и 10 колона. Матрица података потенцијално садржи 1000 вредности (потенцијално јер неке вредности нису присутне). Уколико 10 вредности недостаје, ова матрица ће садржати 990 ненедостајућих вредности, што представља 99%. Уколико применимо технику одбацивања обсервација, избацићемо 10 редова, односно укупно 100 вредности, и коначна матрица ће садржати само 90 редова (900 вредности, 90%).

Јасно се види да се оваквим приступом због 1% недостајућих података, може елиминистаи чак 10% укупних вредности. У сваком случају, избацивање података може касније довести до већих грешака предвиђања, јер се скуп података који служи за тренинг драстично смањује [6].

Стога, у овом делу ће бити описане само технике уметања података, где ће се поља која недостају у матрици података заменити (највероватнијом) вредношћу.

### 4.2.1. Уметање средње вредности

Уметање средње вредности је вероватно најједноставнија метода. Она подразумева замену недостајуће вредносте за сваку променљиву (колону) са средњом вредности познатих обсервација у посматрној колони. Овај приступ може бити погодан у случају када мало података недостаје и то по MCAR механизму. У сваком случају смањује се варијанса међу подацима као и корелација између променљивих. [7] Логично је да се варијанса смањује јер до сада непознату вредност замењујемо ''очекиваном'' вредности, и самим тим смањујемо одступање од те ''очекиване'' вредности. Када се каже да ће корелација бити мања мисли се на корелацију између варијабли (колона). Како овим приступом покушавамо да пронађемо везу између свих обсервација (редова) унутарт једне колоне, логично је да смањујемо корелацију између колона. На пример, скуп података садржи две колоне (висина и тежина), и 5 обсервације. (Табела3)

|  |  |
| --- | --- |
| Висина | Тежина |
| 160 | 67 |
| 165 | 65 |
| 158 | 59 |
| ? | 98 |
| 191 | 98 |

Табела Уметање средње вредности

Очигледно је да постоји веза између прве и друге колоне, и да би недостајућа вредност требало бити замењена са 191 (или неком вредношћу блику кој). Међутим, овом техником се та веза (корелација) занемарује и недостајућа вредност ће постати 168.

Уметање средњих вредности се због описаних недостатака неће даље разматрати у раду. Екпериментални скуп података садржи вредности о особама, где неке променљиве (колоне) имају велику варијансу, па свакако није добра идеја сманњивати ту варијансу. Такоће, испитује се утицај разних фактора на ниво дијабетеса, па је корелација један од главних предуслова.

### 4.2.2. Преношење задњег запажања

Преношење задњег запажања је још једна техника која захтева MCAR механизам недостајућих вредности. Веома је популарна у ситуацијама где се посматрају одређене појаве (субјекти) кроз време. Изузетно може бити знимљива у истраживањимакоје садрже номиналне типове атрибута (вредности колона). На пример, посматрајмо медицинско истраживање које прати пацијента (субјекат) кроз време и бележи да ли је узео терапију или није. Другим речима, постоји номинална променљива са могућим вредностима ДА/НЕ. Уколико је пацијент одређеног дана заборавио да унесе да ли је узео лек или није, та вредност ће се попунити са вредношћу из претходног дана.

Међутим, за променљиве нумеричког типа, ова техника није препоручљива јер повећава пристраснот модела, и такође измњује средњу вредност и варијансу (по променљивој) [8]. Ни ова теника неће бити коришћења у експериментима у овом раду, јер се скуп података за тренинг сасатоји углавном од нумеричких типова, и притом павијенти се не посматрају кроз време.

### 4.2.3. Уметање података коришћењем линеарне регресије

Ова техника захтева одређени ниво корелације између променљивих. Идеја је једноставна; у скупу података са две улазне променљиве и једном излазном неке вредности недостају. Потребно је да попунимо све вредности у нпр. користећи вредности из и . Оно што је компликовано у овом примеру је то што приликом уметања података у , очекивано је да се деси да неке од вредности из , такође нису познате. Због тога потребно је направити 3 једначине линеарне регресије како би се адекватно попуниле вредности у колони .

|  |  |
| --- | --- |
|  | (4.1) |

|  |  |
| --- | --- |
|  | (4.2) |

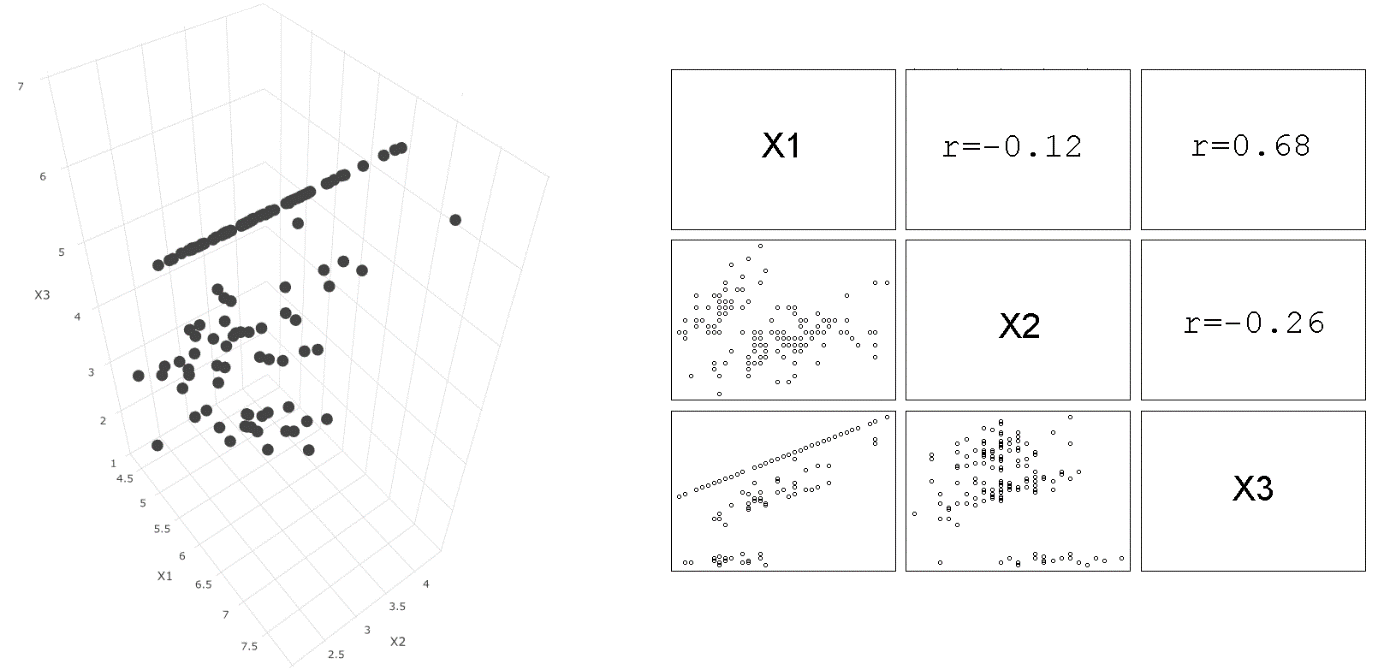
|  |  |
| --- | --- |
|  | (4.3) |

Пример се састоји од само три променљиве и потребно је креирати чак 9 једначина (по 3 једначине за сваку променљиву) како би се адекватно извршила импутација. Поред јасне комплексности проблема, овом техником се повећава корелација између променљивих.

Замислимо да сваки пут када недостаје променљива , такође недостаје и проенљива . Другим речима, подаци недостају по MAR механизму. У том случају, за одређивање вредности користила би се једначина (4.3) и то специјалан облике те једначине:

|  |  |
| --- | --- |
|  | (4.4) |

Овакав начин уметања значајно би повећао вредност корелације између и . На слици (3) је приказана та корелација као и просторни приказ једног скупа података.



Слика Корелација након уметања података линеарном регресијом

Уметање линеарном регресијом захтева да подаци недостају по MCAR механизму, што ће управо бити случај у експериментима у овом раду. Због тога резултујућа корелација између променљивих неће имати вердности као на слици 3. Иако се подаци за експеримент састоје од укупно 11 променљивих () што може да проузрокује велики број једначина линеарне регресије, ова техника ће се користити у даљем раду.

### 4.2.4 Уметање података стохастичком регресијом

У претходном примеру су приказани недостаци уметања линеарном регресијом. Као покушај превзилажења тих недостатака, понекад се користи модификована верзија уметања података која се назива и уметање података стохастичком регресијом. [9].

Идеја је поприлично интуитивна; у једначину линеарне регресије додати још један параметар који ће на случајан начин да промени резултујућу вредност. Самим тим једначина (2.5) постаје:

|  |  |
| --- | --- |
|  | (4.5) |

Како би се једноставније објаснила случајна променљива уместо векторског записа, стохастичку регресију ћемо представити:

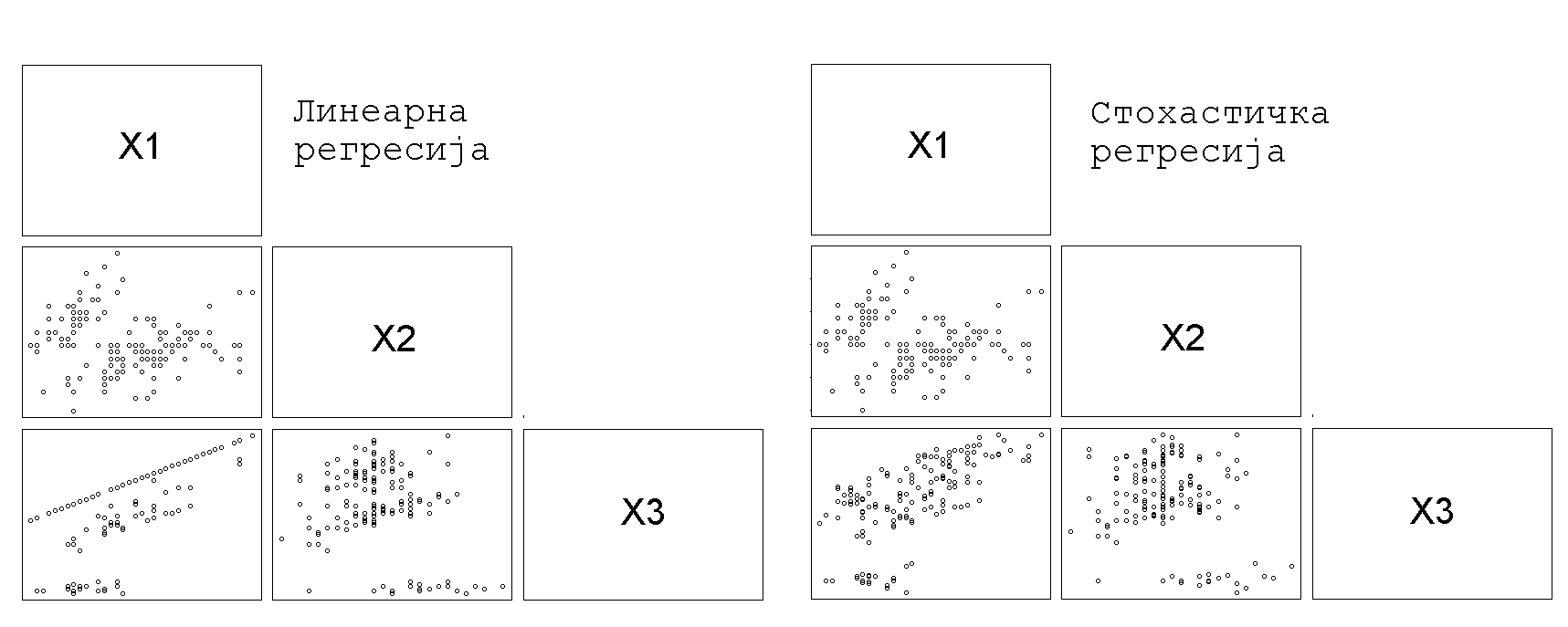
|  |  |
| --- | --- |
|  | (4.6) |

У једначини (4.6), узима вредности од 1 до где је број обсервација у скупу података. Битно је напоменути да је вредност другачија за сваку инстанцу (ред) у скупу. Због те особине није могуће драстично повећати корелацију код попуњеног скупа.

Случајна променљива може да подлеже било којој расподели. Најчешће се користи нормална расподела са очекиваном вредношћу једнаком нули, и стандардном девијацијом једнаком грешци варијанси приликом регресије. [10]

|  |  |
| --- | --- |
|  | (4.7) |

На слици (4) визуелно су приказане обсервације линеарне и стохастичке регресије. Јасно се види да се јака линеарна зависност код линеарне регресије изгубила код стохастичке регресије, мада и даље постоји висок ниво корелације.



Слика Упоредна анлиза импутације линеарном и стохастичком регресијом

Као и техника уметања линеарнорм регресијом, и ова техника ће се користити даље у раду.

## 4.2.5 Уметање коришћењем стабала (шума) одлучивања

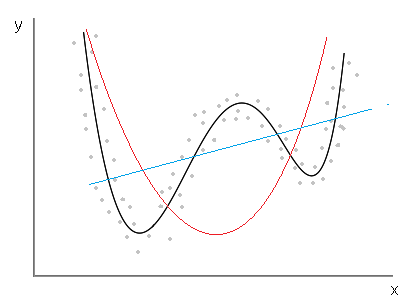
Проблеми са којима се свакодневно сусрећемо често нису линеарни, па коришћење линеарне регресије не даје увек најбоље резултате. Технике 4.2.3 и 4.2.4. се могу унапредити уколико линеарну регресију заменимо са нелинеарним моделом.

Креирање нелинеарне функције није увек тривијалан задатак. Понекад се зависна променљива једино може описати веома комплексном функцијом независне променљиве. Слика 5 показује упоредну анализу различитих регресионих функција на датом скупу података. Подаци се састоје од две променљиве , и где је зависна променљива а независна. Плавом линијом је представљена линеарна функција (4.8), црвеном линијом квадратна функција (4.9), а црном линијом функција вишег реда (4.10).

|  |  |
| --- | --- |
|  | (4.8) |

|  |  |
| --- | --- |
|  | (4.9) |

|  |  |
| --- | --- |
|  | (4.10) |



Слика Упоредни приказ регресионих функција

Ста је уметанје података, мало историје о томе кад је поцело

Које су технике погодне ѕа мцар

О свакој техници по мало

Фокус рада на мцар, како це се мцар користити мало псеудо кода, датасет који имам

[1] Applied Missing Data Analysis, 3

[2] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York

[3] Prevention And Treatment Of Item Nonresponse, 155

[4] Regression Analysis By Example, 1

[5] Least Angle Regression

[6] Data Analysis Using Regression and Multilevel/Hierarchical Models, 531

[7] Principal component analysis with missing values, 659

[8] The Application of Last Observation Carried Forward in the Persistent Binary Case

[9] An introduction to modern missing data analyses, 13

[10] <https://www.ma.utexas.edu/users/davis/375/popecol/lec4/stoch.html>

1. Заправо уколико је апроксимациона грешка једнака нули (или веома блиска нули), може се доћи до претренираности алгоритма што је непожељан ефекат. Другим речима, алгоритам би одлично радио са тренинг подацима али показао би веома лоше резултате над тестном скупу података. Претренираност је честа појава код нелинеарних апроксимација, што у овом раду није случај, па је апроксимирајућа грешка блиска нули пожељна. [↑](#footnote-ref-1)