УНИВЕРЗИТЕТ У БЕОГРАДУ
ФАКУЛТЕТ ОРГАНИЗАЦИОНИХ НАУКА

Мастер академске студије

Студијски програм: Пословна аналитика

Предмет: Мултиваријациона анализа

Пројектни рад

Тема: Примена KNN методе и K-means кластеризације

Ментор: Др Милица Маричић

Студент: Алекса Вучетић

Београд,

јун, 2022.

УНИВЕРЗИТЕТ У БЕОГРАДУ
ФАКУЛТЕТ ОРГАНИЗАЦИОНИХ НАУКА

Садржај

Садржај илустрација

1. Увод

Компанија *BlackRock* је један од највећих инвестиционих фондова на свету. Као таква, она не само да се понаша као инвестициони фонд, већ са великим успехом наступа на тржишту и као компанија која пружа финансијске услуге другим пословним ентитетима. Поред тога, и сама развија свој софтвер за управљање алтернативним инвестицијама.

С обзиром на чињеницу да је скуп фондова који је се користи у интерним апликацијама ове компаније огроман, постоји небројено много алата и начина да се њима манипулише. Један вид манипулације овим фондовима јесте и њихово груписање по разним параметрима и индикаторима. Груписање се може чинити веома тривијалним видом манипулације било којим подацима, међутим, оно заправо омогућава да се уоче разне везе између објеката и може бити веома корисно у даљем коришћењу истих. Такође, оно што показује пракса, јесте и да се у разним извештајима који се шаљу компанијама клијентима тражи груписање фондова у различите нивое или скупове.

У овом раду ће бити речи управо о груписању фондова поменуте компаније и то помоћу два алгоритма машинског учења. Наравно, ради заштите података, фондови који се користе у овом раду су тестни примери компаније. Оно што је значајно поменути јесте то да се и сви извештаји у реалном пословању ове фирме креирају прво на тестним подацима па се тек онда примењују и на реалне. Ова реченииица је наведена како би се показала значајност тестних података.

Алгоритми машинског учења који се користе у овом раду су KNN класификација и K-means кластеризација. Истраживања у разним областима показују да је KNN једна од најефикаснијих техника машинског учења која се односи на класификацију, док је K-means један од најефикаснијих алгоритама кластеризације. Такође, истраживања показују да KNN пружа више него завидну прецизност као алгоритам „под надзором“ док то исто чини и K-means као алгоритам „без надзора“.

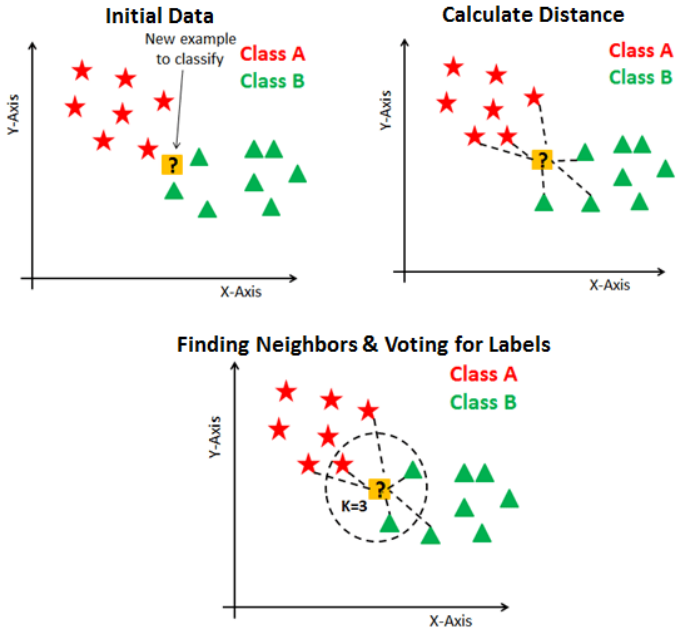
Један од циљева овог рада јесте управо поређење ова два алгоритма и анализа њихових резултата на конкретном примеру.

2. KNN и K-means алгоритми

KNN је анаграм енглеских речи *K Nearest Neighbors* који у преводу значи „*k* најближих суседа“. Овај алгоритам је непараметарски што значи да не прави никакве претпоставке о расподели улазних података. Стога, када је непозната расподела података на основу којих се прави предвиђање, овај алгоритам не би био лош избор. Поред овога, још једна предност овог алгоритма јесте и једноставност примене. Међутим, та једноставност, чини се, нема противутицај на ефикасност овог алгоритма.

У овој методи вредност жељене варијабле се одређује на основу исхода *k* најближих суседа. Шта то значи? Да би запис података *t* био класификован, преузима се његових *k* најближих суседа. Ти суседи формирају околину од *t*. Већинско гласање (да га тако назовемо) међу записима података у окружењу се обично користи за одлучивање о класификацији за *t* са или без разматрања пондерисања заснованог на удаљености. Међутим, да би овај алгоритам био примењен, неопходно је изабрати одговарајућу вредност за *k*, а успех саме класификације умногоме може зависити од избора ове вредности. У извесном смислу, за ову методу се може рећи да је пристрасна са *k*. Постоји много начина за избор најбоље вредности. Најједноставнији начин јесте да се алгоритам покрене много пута и да се изабере она вредност са којом алгоритам пружа најбоље перформансе. Са друге стране, да би KNN био мање зависан од избора *k*, Вангпредлаже да се узме у обзир више скупова најближих суседа, а не само један[1].

[1] Wang, H. (2002). “Nearest Neighbours without k: A Classification Formalism based on Probability”.



Илустрација 1 - KNN класификација[2]

Са друге стране, K-means кластеризација је алгоритам машинског учења „без надзора“. Генерално, алгоритми „без надзора“ доносе закључке користећи само улазне податке без позивања на познате или означене исходе. Циљ ове методе је груписање сличних запажања и откривање основних образаца.

Метода к-means кластеризације има један улазни параметар - *k*, који се односи на број група (кластера, енг. *clusters*) у које ће опсервације бити смештене. Алгоритам заправо проналази центре тих кластера, који се називају центроиди. Затим, додељује сваку опсервацију у кластер заснован на најближем центру кластера. Циљ ове методе је да се минимизира збир квадрата у кластеру, тј. да збир растојања на кватрат сваке тачке од одговарајућег центра кластера буде што мања.

Кораци K-means алгоритма:

1) Почетна селекција центара кластера; центроиди се или генеришу насумично или се бирају из скупа података.
2) Додела кластера; за сваку опсервацију из скупа података, идентификује се најближи центар кластера (обично засновано на еуклидској удаљености) и та опсервација се дедељује групи (кластеру) најближег центроида.
3) Провера вредности центроида; након што се свака опсервација додели одређеном кластеру, рачуна се нови центар тако што се узима средња вредност свих чланова датог кластера.

[2] Слика преузета са: Navlani, A. (avgust 2, 2018) “KNN Classification Tutorial using Scikit-learn”. *datacamp.* https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

Алгоритам понавља кораке 2) и 3) док се не испуни један од следећих услова:

- ниједна опсервација не промени свој кластер;
- збир удаљености је минимизиран;
- достигнут је максимални број итерација.

Резултат алгоритма може бити локални оптимум који није нужно најбољи могући исход. Због тога се овај алгоритам обично покреће неколико пута и то на начин да приликом сваког покретања центроиди буду различити подаци у односу на претходно покретање.

3. Примена одабраних алгоритама статистичког учења у језику *R*

3.1. Опис матрице података

Као што је већ поменуто, матрица података за овај рад преставља тестне податке компаније *BlackRock*. Дакле, фондови који се налазе у њој су непостојећи, али осликавају понашање фондова у реалном пословном окружењу. Сама матрица је настала као производ специјално креираног извештаја за потребе овог рада. Рекавши ово, мисли се да су све варијабле које представљају финансијске индикаторе биране како би испуњавале услове неопходне за примену ове две методе. Оно што је још бинто напоменути, јесте да све ове варијабле чине једну групу индикатора, па се као такве, често и у реалном пословном свету посматрају заједно.

Што се тиче самих финансијксих индикатора, они су објашњени у даљем делу текста. Поред њих објашњене су и остале промењиве коришћене у овом примеру:

- FUND - променљива која представља назив фонда;
- COMMITMENT - променљива која представља висину „обавезе“ коју је сваки инвеститор у фонд дао фонду (пример: особа A улаже у фонд F; он/она даје обавезу да ће у тај фонд уложити одређену своту новца која не мора бити уплаћена одмах у целости, већ у зависности од потреба фонда);
- ADJUSTED_NAV - променљива која представља процењену вредност имовине датог фонда;
- RETURN_OF_CAPITAL - променљива која представља повраћај новца на уложена средства;
- NET_CASH - променљива која представља прилив новца у фонд (али не од инвеститора);
- FUND_STATUS - променљива која предствља статус фонда (фонд у који се улаже, али који не улаже даље, фонд који улаже у друге фондове, итд.);
- CURRENCY - променљива која представља валуту у којој се улаже у фонд;
- REMAINING_COMMITMENT - променљива која представља износ који још није уплаћен у фонд, а обавезан је;
- CASH_CALLED - променљива која представља износ који је уплаћен у фонд;
- CASH_DISTRIBUTED - променљива која представља износ који је фонд исплатио носиоцима акција.

На основу ових финансијских индикатора могуће је утрврдити величина фонда, значајност фонда, активности фонда, успешност фонда, итд. Стога, промењиве које се тичу финансијских показатеља су веома значајне и увек су исказане нумерички (на тај начин се могу квантификовати).

Након уноса матрице података у *R-studio* окружење и након спровођења команде за одређивање типа података сваке променљиве, установљено је да нису све променљиве које би требало да буду нумеричког типа, заправо нумеричког типа. Разлог томе је углавном тај што поље за унос тих података у апликацији компаније користи запете као начин раздвајања хиљада. Самим тим, та променљива се учитава као променљива типа *char*. Стога, неопходно је проверити које су све променљиве учитане на тај начин и променити њихов тип.

Такође, иницијална матрица броји 214 инстанци, тј. 214 различитих фондова са својим финансијским индикаторима. Оно што је битно поменути јесте да су одређена поља ове матрице заправо празна и са њима се мора пажљиво руковати.

Пре него што се пажња посвети посебно обема методама, биће представљено сређивање података које је урађено на исти начин за обе методе.

3.2. Сређивање података

Прво што је урађено јесте то да су све варијабле које би требало да су типа *integer* или *numeric*, а нису, претворене у променљиве типа *numeric* помоћу функције:

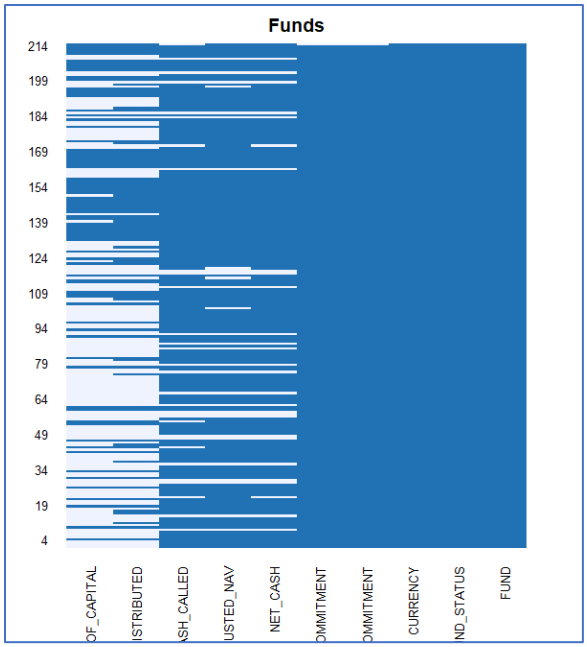
```
fund_indicators$COMMITMENT <- as.numeric(fund_indicators$COMMITMENT);
```

Променљиве FUND_STATUS и CURRENCY су претворене у варијабле факторског типа:

```
fund_indicators$FUND_STATUS <- as.factor(fund_indicators$FUND_STATUS);
```

Након овога, свим вредностима 0, је уместо нуле додељена NA вредност како би подаци били манипулативнији, али како би опет број инстанци био задовољавајући за даљу анализу.

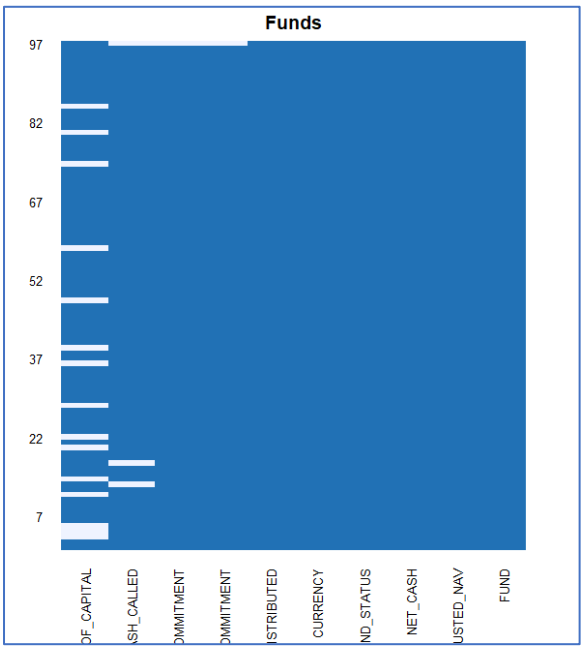
Када је ово завршено, пажња се обратила на недостајуће вредности. Стање пре даљег сређивања ја изгледало овако:



Илустрација 2 - Приказ недостајућих вредности пре сређивања података

Уочава се велики број недостајућих вредности за варијабле RETURN_OF_CAPITAL и CASH_DISTRIBUTED.

Начин који је изабран за сређивање недостајућих вредности јесте да се избаце све инстанце које имају недостајуће вредности променљиве CASH_DISTRIBUTED. Након овог корака, број инстанци који је остао у сету података јесте 97. Стање након овога изгледа овако:



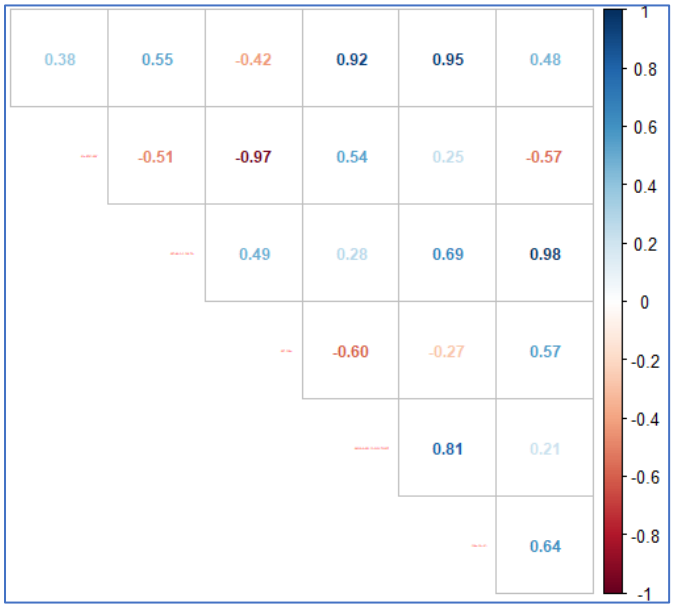
Илустрација 3 - Приказ недостајућих вредности након делимичног сређивања података

Сада се види да постоји само неколико NA вредности. Те вредности ће бити замењене средњом вредношћу или медијаном у зависности од расподеле коју има свака промнљива посебно. Да би се утврдила расподела, користи се Шапиров тест нормалности:

```
shapiro.test(fund_indicators$COMMITMENT);
```

Како све тестиране варијабле имају п-вредност мању од 2,2e-16 закључује се да ниједна варијабла са недотајућим вредностима (COMMITMENT, ADJUSTED_NAV, RETURN_OF_CAPITAL, NET_CASH, CASH_DISTRIBUTED) нема нормалну расподелу. Стога недостајуће вредности се мењају њиховим медијанама.

Када су сређене недостајуће вредности проверава се међузависност сваке варијабле како би се утврдило да ли нека има утицај на другу.



Илустрација 4 - Приказ међузависности варијабли

На претходној слици види се које варијабле утичу једна на другу и те варијабле се избацују из даљег посматрања. Променљиве које су из овог конкретног рада избачене су REMAINING_COMMITMENT и CASH_CALLED.

Након овога, сет података је спреман за примену алгоритама.

3.3. KNN алгоритам

Циљана варијабла овог алгоритма је COMMITMENT и на основу ње ствара се нова која представља да-не варијаблу факторског типа која приказује да ли је за дати фонд COMMITMENT већи од 3. квартила читаве променљиве или није.

Да би могао да се примени овај алгоритам неопходно је да се стандардизују вредности. Опет треба узети у обзир нормалност променљивих. Како је већ показано да варијабле у овом

примеру немају нормалну расподелу, оне ће бити стандардизоване помоћу медијане и интерквартилног размака. То је урађено на следећи начин:

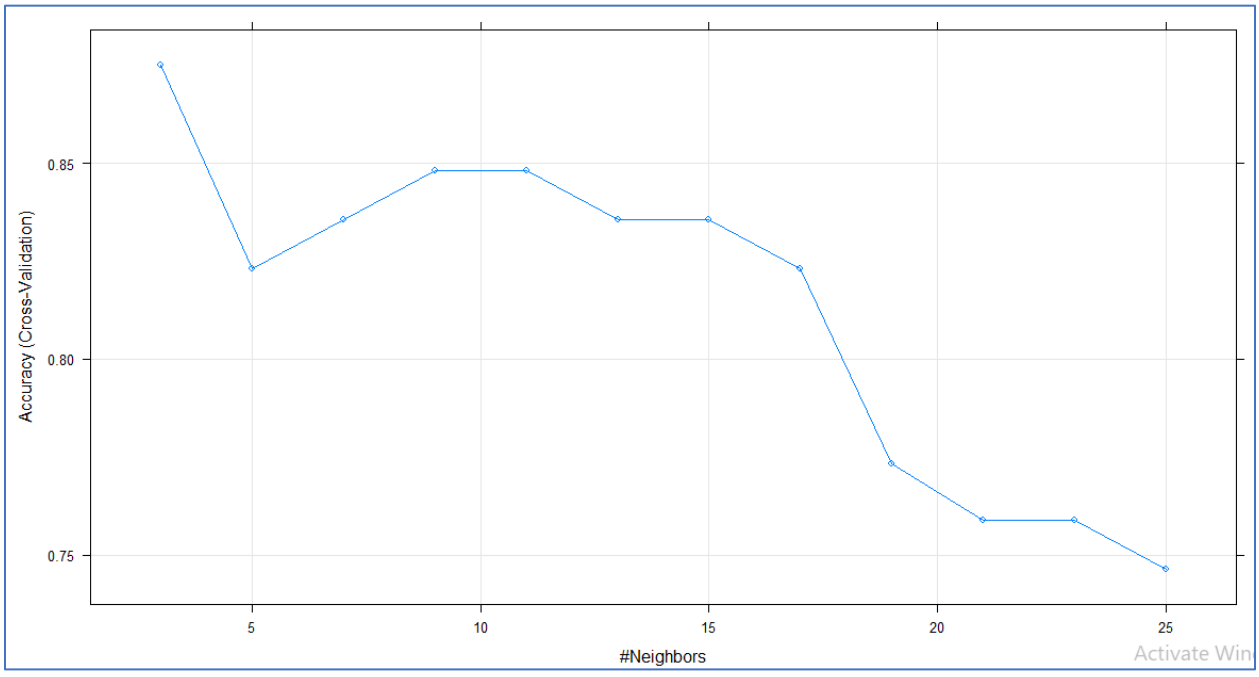
```
fund_indicators.standardized <- apply(X = fund_indicators[,numeric.vars],
                                      MARGIN = 2,
                                      FUN = function(x) scale(x, center = median(x), scale = IQR(x)));
```

Након стандардизације, сет је подељен на тест сет и тренинг сет где тренинг сет чини 80% инстанци док је преосталих двадесет резервисано за тест сет.

Да би се одабрао оптималан број к, тј. оптималан број суседа користи се крос валидација

```
knn.cv <- train(x = train.podaci[,-7],
                y = train.podaci$HIGH_COMMITMENT,
                method = "knn",
                trControl = numFolds,
                tuneGrid = cpGrid),
```

где су numFolds и cpGrid параметри крос валидације. У теоријском делу рада биће приказана визуелизација крос валидације.



Илустрација 5 - Оптималан број суседа

Са слике се види да је оптималан број суседа три. Када је одређен оптималан број суседа сада се може применити и сам модел:

```
knn.pred <- knn(train = train.podaci[,-7],
                test = test.podaci[,-7],
                cl = train.podaci$HIGH_COMMITMENT,
                k = 3);
```

Да би добијене вредности биле приказане на прави начин, креирана је матрица конфузије која изгледа овако, а за њом је креирана и матрица евалуационих метрика која ће бити анализирана у даљем тексту. Као што се може видети у *R* фајловима који иду уз овај рад, све четири евалуационе метрике имају вредности близу 1 што значи да модел ради готово савршено.

3.4. K-means кластеризација

Код ове методе, након сређивања података неопходно је обратити пажњу на екстремне вредности јер оне негативно утичу на саму методу. Након урађеног теста на *outlier*-е, установљено је да:

- COMMITMENT има 14 екстремних вредности;
- ADJUSTED_NAV има 19 екстремних вредности;
- RETURN_OF_CAPITAL има 5 екстремних вредности;
- NET_CASH има 28 екстремних вредности;
- CASH_DISTRIBUTED има 13 есктремних вредности.

За сређивање ових вредности коришћена је Виндзорска техника која прво проверава да ли променљива има доње, горње или и једне и друге екстремне вредности. Она их елиминише на начин да уколико има горње екстремне вредности узима све вредности које су веће од 3. квартила (укључујући и *outlier*-е) и мења их самом вредношћу 3. квартила (3. квартила је узет као пример; не мора нужно да то буде та вредност). Иста ситуација је и за доње.

Након елиминисања екстремних вредности, све нумеричке варијабле су нормалнизоване помоћу формуле:

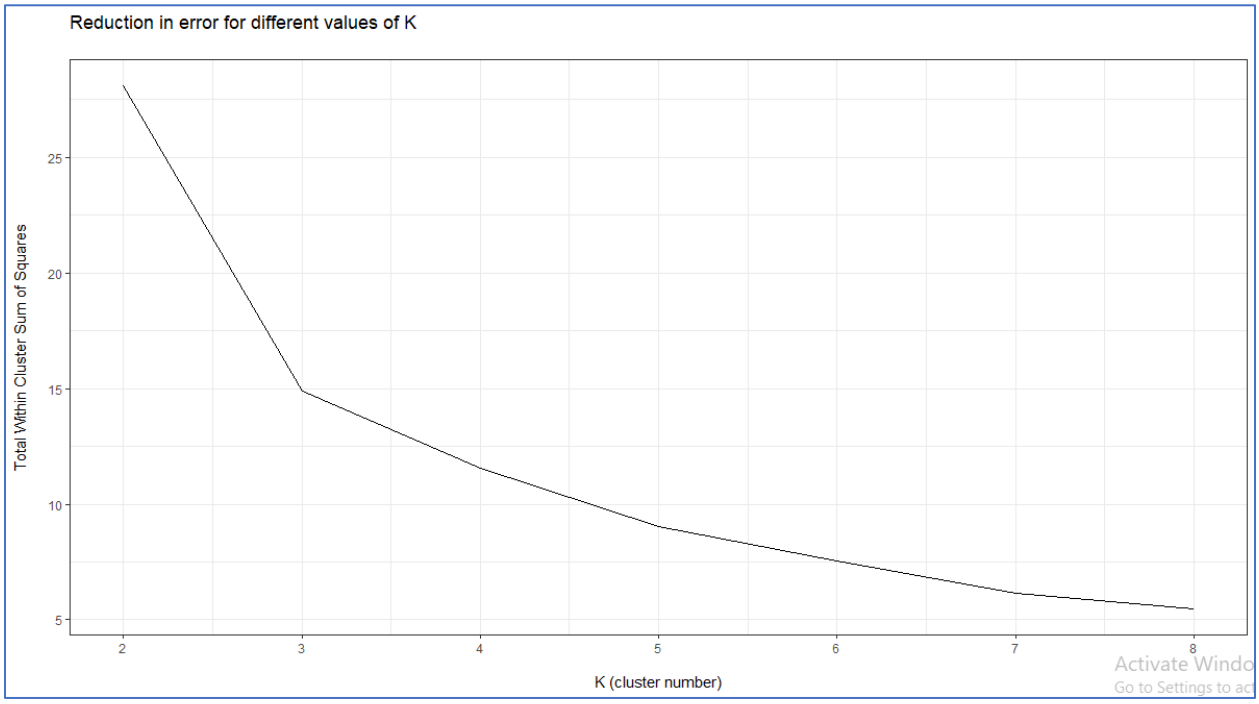
$$\frac{x-\min(x)}{\max(x)-\min(x)}.$$

Када је све сређено, неопходно је одредити оптималан број кластера. То се ради коришћењем *Elbow* метода:

```
for (k in 2:8) {
  set.seed(111)
  km.res <- kmeans(x = fund_indicators.norm,
```

```
                centers = k,
                iter.max = 20,
                nstart = 1000)
        evaluacione.metrike <- rbind(evaluacione.metrike,
                        c(k, km.res$tot.withinss, km.res$betweenss/km.res$totss))
};
```

Како би се лакше одредио оптималан број кластера, урађена је и визуелизација:

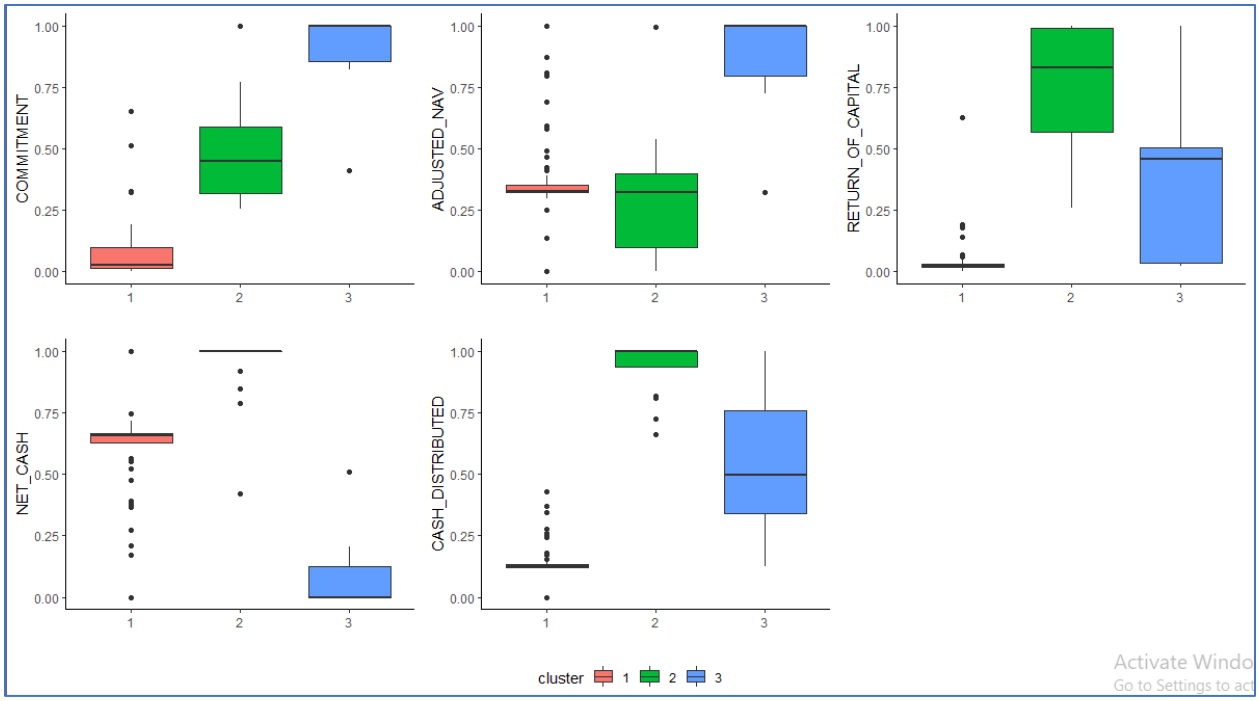


Илустрација 6 - Elbow метода

Са слике се одлично види да је оптималан број кластера три.

Када је одређен оптималан број кластера, коначно се спроводи кластеризација:

```
fund_ind.clusters <- kmeans(x = fund_indicators.norm,
                            centers = 3,
                            iter.max = 20,
                            nstart = 1000);
```



Илустрација 7 - Кластери

Након визуелизације решења, може се закључити да први кластер (црвени) броји најмање чланова и има највише вредности које заправо могу да припадају и неком другом клстеру. Кластер два (зелени) окупља највише вредности променљивих NET_CASH, CASH_DISTRIBUTED и RETURN_OF_CAPITAL док такође окупља средње вредности преостале две варијабле. За разлику од њега и трећи кластер (плави) има две варијабле по којима окупља највише вредности док такође има и две у којима окупља оне инстанце са најнижим вредностима.

4. Закључак и анализа резултата

Након спроведених метода, анализиране су перформансе модела као и сами резултати.

Када се говори о KNN методологији, већ је речено да је то метода „са надзором“. Тако је и изабрана променљива COMMITMENT као циљана променљива. Дакле, оно што је био циљ овог модела јесте да на основу осталих променљивих класификује све инстанце у две групе. Прву групу чине фондови, односно опсервације, чија је вредност променљиве COMMITMENT већа од трећег квартила исте проемнљиве, док другу групу чине остале инстанце. Како је 80% опсервација било тестног карактера, модел је имао за циљ да осталих 20% расподели у две групе на основу онога што је „научио“.

	predicted	
true	No	Yes
No	13	1
Yes	2	2

Илустрација 8 - Матрица конфузије

Оно што говори ова матрица јесте да је од 13 инстанци које заправо немају вредност задате променљиве већу од 3. квартила дате променљиве, заиста предвиђено да немају дату вредност док је једна инстанца која је заправо мања, предвиђена као већа. Са друге стране, од четири инстанце које су веће од задате вредности две су предвиђене на прави начин, док две нису. На онсову матрице конфузије може се видети да је од осамнаест примена, модел промашио само 3, што је прилично висок проценат.

Ти резултати су анализирани у матрици евалуационих метрика.

Прва метрика која је коришћена у овом раду је тачност која се добија тако што се број погођених вредности подели са бројем укупних вредности и добија се да је прецизност 15/18, тј. 0.8333 што је прилично висок проценат. Друга метрика која је коришћена је прецизност, тј. однос броја тачно предвиђених позитивних вредности и збира тачно и лажно предвиђених вредности. Тај број износи 0.8667 што је такође висока вредност. Претпоследња метрика је такозвана *recall* метрика која приказује однос између тачно предвиђених позитивних вредности и укупних предвиђених позитивних вредности и тај број износи 0.9286. Последња метрика је заправо однос претходних метриак и добија се по формули:

$$\frac{2*\text{прецизност}*recall}{\text{прецизност}+recall}.$$

Њена вредност је 0.8966. Као што се види, модел даје прилично задовољавајуће резултате што значи да је прилично ефикасан.

Када се говори о K-means моделу, већ је приказана визуелизација кластера и остављен је коментар на добијена решења. Оно што ће бити предмет дискусије тог решења у овој секцији јесте COMMITMENT варијабла.

С обзиром да је ова променљива постављена као циљана променљива у претходном алгоритму овде ће бити речи само о њој како би се направила поређења између два алгоритма.

Као што се види са слике 7 (Илустрација 7), кластер 3 чине инстанце чије се вредности варијабле COMMITMENT налазе изнад вредности трећег квартила дате варијабле. Остала два кластера чине остале инстанце. Наравно, извесна одступања постоје.

Узевши у обзир чињеницу да су у KNN методи предвиђени резултати у само две групе док је у другој методи то био случај са три кластера, резултати K-means алгоритма се могу апроксимовати да такође и ти резултати буду приказани у два кластера. На тај начин, долази се до првог предлога за даље истраживање. Наиме, ова три кластера могу и сами да се провуку кроз овај алгоритам тако да се ова три кластера групишу у два нова кластера који ће имати нове центроиде који ће по свему судећи груписати та три кластера у две групе где ће једну групу чинити два садашња кластера, и то кластери 1 и 2, тј. они кластери који се налазе испод вредности 3. квартила. Другу гурпу ће чинити садашњи кластер 3 и на тај начин добијене су две групе. Ако се упореде резултати, вероватно ће решења бити таква да ће готово подједнак проценат инстанци у обе методе отићи у одговорајћуе кластере.

Тренутно је ситуација таква да у KNN методи је око 20% предвиђених резултата добило исход *Yes* што значи да су њене вредности веће од 3. квартила. Може се претпоставити да ће се добити сличан проценат и самом кластеризацијом. Међутим, то остаје предмет могућег даљег истраживања.

Са друге стране, оно што се може урадити са KNN алгоритмом јесте да се циљана вараијабла може начинити да буде груписана у три групе (уместо досадашњих *Yes* и *No* резултата) које би осликавале вредности мање од првог квартила, веће од трећег квартила и све остале оне између. На тај начин добиле би се три групе које би могле да се пореде са три кластера из друге методе.

Међутим, оно што се чини на основу тренутног спроведеног истраживања јесте да се резултати обе методе у великој мери поклапају.

Оно што су потенцијални недостаци овог истраживања су пре свега недостајуће вредности саме матрице као и *outlier*-и. Ови недостаци се свакако везују за матрицу података и на методе одређене за решавање ових проблема. Што се тиче самих метода, не постоје неки велики недостаци, али оно што се може употребити као побољшање јесу неке софистицираније методе одређивања *k* у оба случаја, као и већи број спроведених предвиђања. Ипак, у случају тренутне матрице података, повећан број спроведених предвиђања можда не би имао неки утицај, али у случају неупоредиво веће матрице података, ово би могло да буде веома значајно.

5. Слична истраживања

Постоји велики број истраживања на тему поређења KNN и K-means алгоритама. Једно од таквих истраживања јесте и студија коју су спорвели Кавита Митал, Гаурав Агарвал и Прерна Махаџан са Института компјутерске примене и менаџмента у Њу Делхију у Индији (2018). Они су правили поређење између две методе, али у индустрији здравства. Оно што се може извући као препорука из овог истраживања јесте већ наведена идеја о примени KNN метологије за више инстанци *k* вредности, а затим и поређење са резултатима добијених методом K-means кластеризације.

Такође, студија спроведена од стране Мохудина Ахмеда са Едит Коан универзитета у Аустралији, Раихана Сераџа са МекГили универзитета у Монтреалу у Канади и Сајед Мухамед Шамсул Ислама са Универзитета Западне Аустралије у Аустралији (2020) указује на недостатке K-means алгоритма наводећи да, иако је овај алгоритам један од најмоћнијих и најпопуларнијих, он има одређена ограничења, укључујући проблеме са насумичним налажењем центроида кластера што према њиховим речима доводи до неочекиване конвергенције.

Још једно истраживање се бави предикцијама у свету финансија користећи KNN алгоритам. Истраживање о класификацији компанија зарад предвиђања кретања цена њихових акција се може успешно применити и на класификацију фондова. Ово истраживање су спровели Калид Алктаиб, Исмаил Хмеиди, Мохамед К. Али Шатнави и Хасан Нацадат (2013). Прва тројица су са Универзитета науке и технологије у Ирбиду у Јордану, док је последњи научник са Таиф Универзитета у Саудијској Арабији.

Једна од већих књига коју такође треба узети у обзир горови о примени неколико метода комбинованих у једну. У таквом скупу метода главно место заузима метода „без надзора“ K-means кластеризација. Сами детаљи ове примене превазилазе опсег овог предмета, али свакако се могу употребити у даљем истраживању. Књигу „Машинско учење у области финансија - од теорије до праксе“ написали су Метју Ф. Диксон, Игор Халперин и Пол Билокон (2020).

Биће наведено још једно истраживање у области финансија и примене KNN алгоритма. То истраживање се бави класификацијом инсанци акција на индијском тржишту. Међутим, на исти начин се може применити алгоритам и у случају овог рада. Ова студија је значајна из разлога што наводи груписање инстанци било чега као веома битну ствар зарад будуће манипулације и одређивања одређених циљева за одређене групе. Студију су спровели Франческо Рундо, Франческа Трента и Себастијано Батиато са Универзитета Катанија у Катанији у Италији и Агатино Луиђи ди Стаљо са Департмана примењен математике у Рагузи у Италији.

Наведена истраживања се могу применити са практичне стране делимично и у овом раду, док са теоријске стране, ова истраживања указују и на неке од недостатака овог рада.

6. Референце

- Ahmed, M. & Seraj, R. & Islam, S. M. S. (2020). "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation", str: 1-3.
- Mittal, K. & Aggarwal, G. & Mahajan, P. (2018). "Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy", str: 2.
- Guo, G. & Wang, H. & Bell, D. & Bi, Y. & Greer, K. (2003) "KNN Model-Based Approach in Classification", str: 3.
- Wang, H. (2002). "Nearest Neighbours without k: A Classification Formalism based on Probability".
- Dixon, M. F. & Halperin, I. & Bilokon, P. (2020). "Machine Learning in Finance".
- Alkhatib, K. & Najadat, H. & Hmeidi, I. & Shatnawi, M. K. A. (2013). "Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm", str: 2.
- Rundo, F. & Trenta, F. & Stallo, A. L. D. & Battiato, S. (2019). "Machine Learning for Quantitative Finance Applications: A Survey", str: 5 - 18.
- Nayak, R. K. & Mishra, D. & Rath, A. K. (2015). "A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices", str: 670 - 680.
- Katedra za veštačku inteligneciju, Fakultet organizacionih nauka. (2022). „K Nearest Neighbours (KNN)", str: 1 – 13.
- Katedra za veštačku inteligneciju, Fakultet organizacionih nauka. (2022). „K-means Clustering", str: 1 – 27.
- Laboratorija za statistiku, Fakultet organizacionih nauka. (2022). "Data clustering with R", str: 1 – 36.