Апартмани за изнајмљивање - САД

Пројекат за курс Истраживање података

Аутор: Петра Игњатовић 63/2020
Асистент: Стефан Капунац
Професор: Ненад Митић

Садржај

Увод

Овај рад се фокусира на анализу базе података Apartments for rent. База садржи поделу јединица за изнајмљивање на територији Сједињених Америчких Држава на пет категорија, оглашених на различитим сајтовима. Боље упознавање са датом базом следи у наредних неколико страна.
У овом раду ће се спровести алгоритми за класификацију (K најближих суседа и Стабла одлучивања) и кластеровање (K средина, хијерархијско и DBSCAN), и на крају је приказана примена Априори алгоритма за проналажење правила придруживања.

Експлоративна анализа података

	id	category	title	body	amenities	bathrooms	bedrooms	currency	fee	has_photo	...	price_display	price_type	square_feet	address
0	5668626895	housing/rent /apartment	Studio apartment 2nd St NE, Uhland Terrace NE,...	This unit is located at second St NE, Uhland T...	NaN	NaN	0.0	USD	No	Thumbnail	...	$790	Monthly	101	NaN
1	5664597177	housing/rent /apartment	Studio apartment 814 Schutte Road	This unit is located at 814 Schutte Road, Evan...	NaN	NaN	1.0	USD	No	Thumbnail	...	$425	Monthly	106	814 Schutte Rd
2	5668626833	housing/rent /apartment	Studio apartment N Scott St, 14th St N, Arling...	This unit is located at N Scott St, 14th St N,...	NaN	1.0	0.0	USD	No	Thumbnail	...	$1,390	Monthly	107	NaN
3	5659918074	housing/rent /apartment	Studio apartment 1717 12th Ave	This unit is located at 1717 12th Ave, Seattle...	NaN	1.0	0.0	USD	No	Thumbnail	...	$925	Monthly	116	1717 12th Avenue
4	5668626759	housing/rent /apartment	Studio apartment Washington Blvd, N Cleveland ...	This unit is located at Washington Blvd, N Cle...	NaN	NaN	0.0	USD	No	Thumbnail	...	$880	Monthly	125	NaN

База се састоји из 10.000 инстанци и 22 атрибута:

- идентификатор огласа
- категорија - апартман, кућа или краткотрајно изнајмљивање
- наслов огласа
- текст огласа
- листа погодности
- број купатила
- број соба
- валута
- провизија агенцији или оглашивачу
- слике
- дозвољени кућни љубимци

- цена
- начин плаћања - месечно или недељно
- адреса
- град
- држава
- квадратура изражена у стопама квадратним
- географска ширина
- географска дужина
- интернет страница огласа

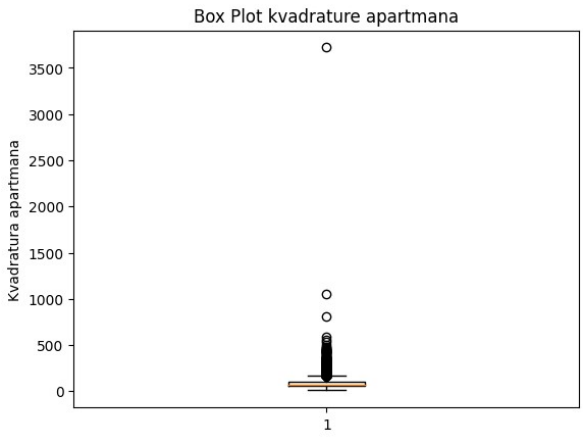
База података као што је ова може бити корисна Агенцијама за некретнине при креирању маркетиншких потеза и стратегије за продају или изнајмљивање јединица циљној групи, WorkAndTravel програму и слично.
При одабиру апартмана, клијента највише интересују цена, квадратура и локација, па ће највећи акценат овог рада бити стављен на те атрибуте.

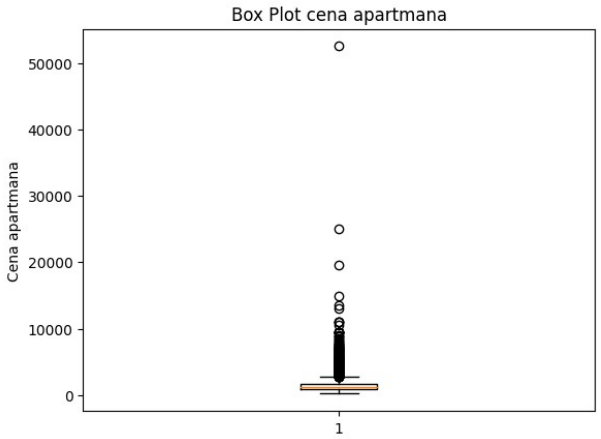
Претпроцесирање

За лакше руковање базом предузете су следеће мере:
1. Категорије су спојене у једну - апартман - из семантичких разлога. Кућа се може схватити као апартман, а тип 'на краћи временски период' подразумева да се јединица издаје на 2-6 месеци.
2. Све валуте су биле амерички долари, па је ова колона избачена.
3. Ниједна вредност у колони 'такса' није била позитивна, па је ова колона избачена.
4. Сви апартмани су имали слике и имали дозвољене кућне љубимце (било то пси, мачке или специјалне сорте истих), па су те колоне избачене.
5. Постојао је један апартман са недељним нивоом плаћања, па је њему ручно постављена потребна цена за месечни ниво.
6. Квадратне стопе су пребачене у квадратне метре, за лакшу интерпретацију резултата.

Аутлајери и екстремне вредности

примењено третирање: ручна корекција, замена средњом вредношћу





Битно је препознати које вредности су аутлајери, а које су екстремне вредности. Кућица (box plot) даје визуелизацију расподеле података и олакшава уочавање елемената који се издвајају.
Квадратура: једна инстанца наводно има преко 3500m^2. Међутим, бољи поглед на ту инстанцу говори да је наслов тог огласа: One BR in New York NY
Овај аутлајер ће бити замењен просечном вредношћу једнособних апартмана у Њујорку, што је око 71m^2.
Нови максимум је око 1000m^2, што понови звучи као грешка, али наслов огласа је овај пут Six BR 9908 Bentcross Drive, тако да вредност може бити легитимна и оставља се у бази.

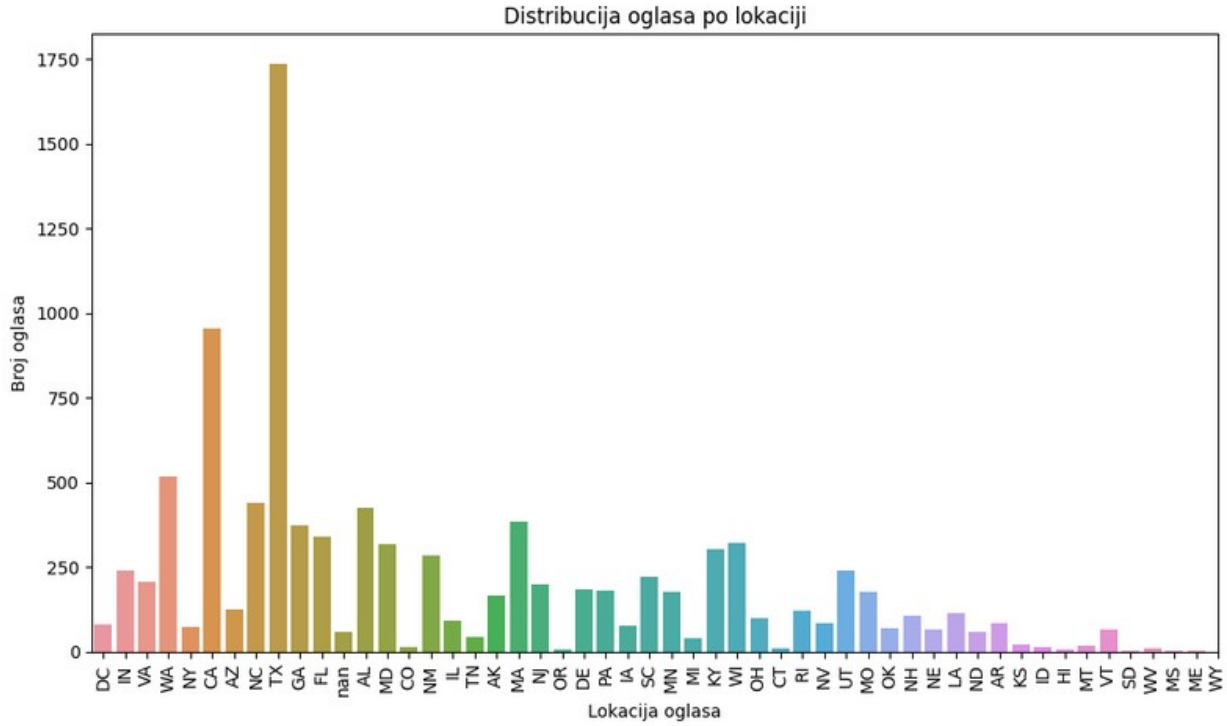
Цена: 50.000 долара је дефинитивно непревазиђена. У тексту огласа стоји следеће: '...$500.00/ month $350.00 deposit...' што говори да је по среди само грешка при уносу и може се ручно исправити.
Следећи максимум је око 25.000 долара месечно. У питању је шестособан апартман у Калифорнији, у елитном крају са вишим животним стандардима: Six BR 256 Las Entradas, тако да се и овај екстремум оставља у бази.

```
count    10000.000000
mean        87.595476
std         48.987726
min          9.393000
25%         60.357000
50%         74.586000
75%        102.300000
max       1052.574000
Name: square_meters, dtype: float64
```

```
count    10000.000000
mean      1481.077500
std        947.983982
min        200.000000
25%        949.000000
50%       1270.000000
75%       1695.000000
max      25000.000000
Name: price, dtype: float64
```

Нема недостајућих вредности код ових атрибута.
Просечна величина апартмана је око 90 квадрата, најмања јединица има 9, а највећа 1052 квадрата.
Просечна цена апартмана је 1500 долара, што има смисла с обзиром на виши стандард у Америци. Најмања вредност је 200 долара и она припада јединици смештеној у вароши Маунт Ери (Maunt Airy) у Северној Каролини, која има око 10 000 становника (по попису из 2021. године).

Локација



Највише смештаја се налази у Тексасу, Калифорнији и Вашингтону, а најмање у Јужној Дакоти, Мисиспију, Мејну и Вајомингу.

Недостајуће вредности

примењене технике: избацивање колоне, избацивање врсте, попуњавање средњом вредношћу групе, бинаризација атрибута

```
id                0
category          0
title             0
body              0
amenities      3549
bathrooms        34
bedrooms          7
currency          0
fee               0
has_photo         0
pets_allowed   4163
price             0
price_display     0
price_type        0
square_meters     0
address        3327
cityname         77
state            77
latitude         10
longitude        10
source            0
time              0
dtype: int64
```

Недостајуће вредности у колонама које описују географску ширину и дужину представљају 0.1% скупа, па могу бити занемарене.
Редови у којима су непознати градови и/или државе чине око 0.7% скупа, па се могу избацити.
Адреса је податак који неће имати много утицаја у даљем истраживању јер ће већу контрибуцију давати географска ширина и дужина, па се ова колона може у потпуности занемарити. Слично важи и за колону која је показатељ да ли су дозвољени кућни љубимци.
Број купатила и соба је битан податак. Број њихових недостајућих вредности није висок, па се може попунити средњом вредношћу осталих атрибута.
Проблем: колона погодности.
Недостаје готово трећина скупа, па није делотворно избрисати те инстанце, с обзиром да се тиме губи значајна количина података.
Отежавајућа околност је што неки апартман може да има наведено 'фрижидер, патос, микроталасна пећница, кабловска', а други може да има 'кабловска, фрижидер, теретана, базен'. Очекивано је да ће присуство структуре као што је базен индиковати скупљи апартман. Очигледно је да вредности нису уникатне, и могу се појављивати различите варијације скупа погодности.
У већини огласа, тело садржи детаљан опис из кога се могу извући сви неопходни подаци. За тако нешто је, наиме, потребан модел NLP-а.
Уместо тога, једноставније решење је бинаризација. Примењени су следећи кораци:

1. Издвајање скупа већ наведених погодности:
'AC', 'Alarm', 'Basketball', 'Cable or Satellite', 'Clubhouse', 'Dishwasher', 'Doorman', 'Elevator', 'Fireplace', 'Garbage Disposal', 'Gated', 'Golf', 'Gym', 'Hot Tub', 'Internet Access', 'Luxury', 'No listed amenities', 'Parking', 'Patio/Deck', 'Playground', 'Pool', 'Refrigerator', 'Storage', 'TV', 'Tennis', 'View', 'Washer Dryer', 'Wood Floors'

2. Одређивање 'луксузног' подскупа:
Alarm, Basketball, Clubhouse, Doorman, Fireplace, Golf, Gym, Hot Tub, Luxury, Patio/Deck, Playground, Pool, Tennis, View, Wood Floors.

3. Пролазак кроз сваки ред
Идеја је да уместо две колоне - *body* и *amenities* - постоји само једна колона: *luxury*, која ће да буде индикатор да ли је јединица отмена или није. Критеријум за утврђивање ће да буде постојање објеката из луксузног подскупа. Елементи подскупа су одређени субјективном проценом.
За јединице којима вредност у колони '*amenities*' није недостајућа, индикатор ће бити извучен из те вредности, а ако јесте недостајућа, онда ће бити неопходан наиван приступ - пролазак кроз оглас реч по реч и претрага да ли та реч припада луксузном скупу.

Након свих наведених мера, база података је очишћена и спремна за даље коришћење.

Класификација

Задатак класификације је одредити функцију (класификациони модел) који пресликава сваки скуп атрибута X = (x^1, x^2, x^3...) у једну од предефинисаних вредности *y*, при чему је *y* - ознака класе. Овај процес се назива класификацијом уколико је *y* категорички, а регресијом уколико је нумерички атрибут.
Класификација објеката по цени може да буде значајно за анализирање профитабилности класа или предвиђање будућих цена у различитим сегментима тржишта.
Ова секција се бави предвиђањем цене апартмана користећи два алгоритма: K најближих суседа и Стабла одлучивања.

1.K најближих суседа

k nearest neighbors - KNN

KHH алгоритам покушава да предвиди тачну класу за инстанцу из тест скупа на основу израчунате удаљености између те инстанце и свих тачака из тренинг скупа, и онда изабере K тачака које су јој најближе. Она класа која је најдоминантија међу изабраних K ће бити додељена тест-инстанци.
У случају регресије, вредност која ће бити додељена је средња вредност K изабраних тачака.
Циљни атрибут ће бити цена, а с обзиром да је цена нумерички атрибут, за потребе класификације на њега мора бити примењена дискретизација.
Интервали су одређени емпиријски.

интервали:

- cheap/affordable

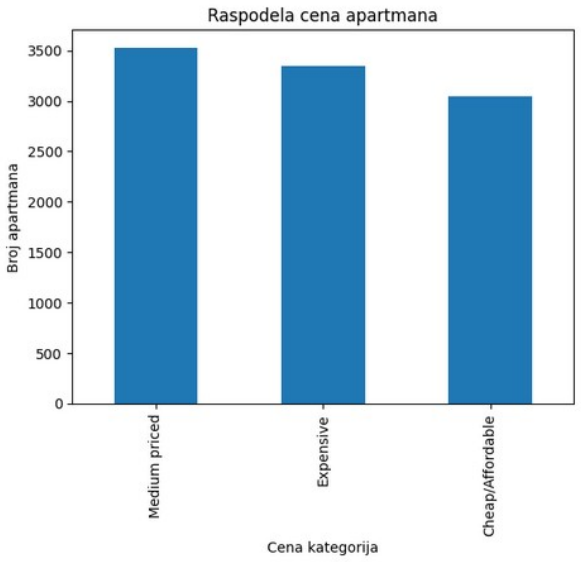
min - 1000

- medium priced

1000 - 1500

- expensive

1500 - max



```
price_category
Medium priced       3528
Expensive           3348
Cheap/Affordable    3046
Name: count, dtype: int64
```

На овај начин класе су приближно балансиране.

Алгоритам се састоји из 3 корака:
1. подела података на тест и тренинг скуп
2. нормализација података
3. тренирање модела.

Неопходно је нормализацију извршити након поделе података на тест и тренинг скуп да се тест скуп не би компромитовао.
У овом примеру је искоришћен *MinMaxScaler*.
Карактеристике које су узете за предвиђање цене су квадратура стана, индикатор лукузности, и географска ширина и дужина. Број соба и купатила нису узети у разматрање да не би X-скуп био превелике димензије, а квадратура је довољна да опише структуру апартмана.

Скуп података је подељен на следећи начин: трећина скупа чине тренинг, остатак чини тест.

```
Classification report for model KNeighborsClassifier on training data
--------------------------------------------------------------------------------
                  precision    recall  f1-score   support

Cheap/Affordable       0.74      0.82      0.78      2437
       Expensive       0.82      0.81      0.81      2678
   Medium priced       0.73      0.67      0.70      2822

        accuracy                           0.76      7937
       macro avg       0.76      0.77      0.76      7937
    weighted avg       0.76      0.76      0.76      7937
```

```
Classification report for model KNeighborsClassifier on test data
--------------------------------------------------------------------------------
                  precision    recall  f1-score   support

Cheap/Affordable       0.65      0.72      0.68       609
       Expensive       0.75      0.71      0.73       670
   Medium priced       0.61      0.59      0.60       706

        accuracy                           0.67      1985
       macro avg       0.67      0.67      0.67      1985
    weighted avg       0.67      0.67      0.67      1985
```

Анализа добијених резултата

Тачност (accuracy) модела је мера која говори колико је модел успешан у предвиђању тачних одговора у односу на укупан број примерака. Тачност се израчунава као однос броја тачних предвиђања које је модел извршио и укупног броја примерака у тестирању.

Тачност модела на тренинг скупу је 76%, док је на тест скупу нешто мања, 67%. Модел је солидан, али има проблем при генерализацији, што захтева експериментисање са параметрима.

1. Прецизност
На тренингу, 74% инстанци које су класификоване као *'Cheap/Affordable*' заиста припадају тој класи, док је на тест скупу тај проценат за 9 мањи (65%). Слично важи и за класу *'Expensive'* - разлика је за 9 процената. Модел је највише проблема имао са класом '*Medium priced',* која обухвата најмањи распон цена (500 долара, у поређењу са првом и трећом класом које имају распон, редом, 800 и 10000)

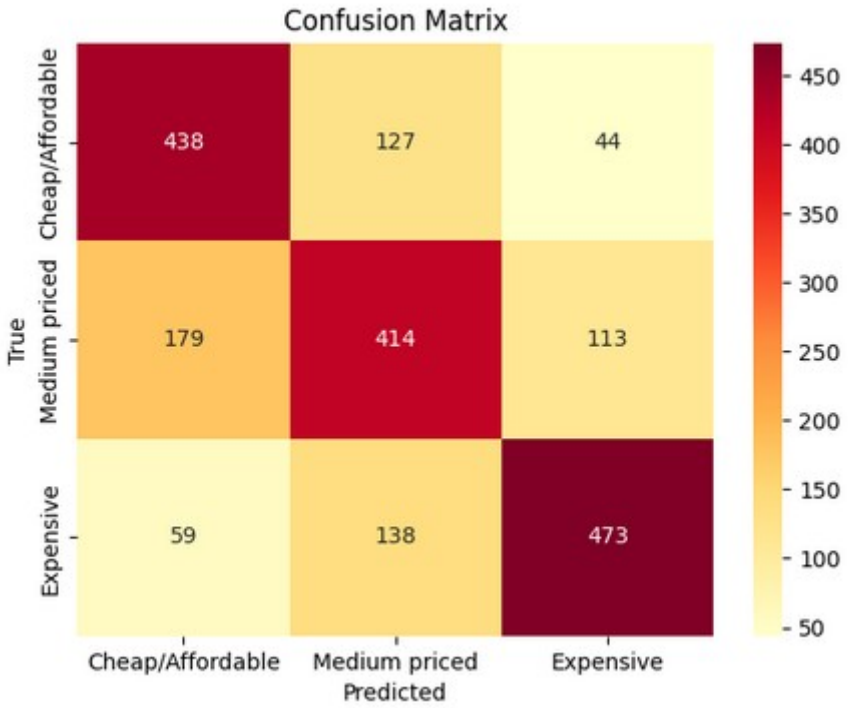
2. Одзив (*Recall*)
Ова вредност показује да је на тренингу модел предвидео 82% *'Cheap/Affordable'* апартмана од њиховог укупног броја. Слично важи и за класе '*Medium priced'* и '*Expencive',* са вредностима 67% и 81%. Разлика у односу на тест скуп је поново око 10 процената, и поново се најлошије показао на класи средњих цена.

3. F1-*score*
F1-*score* је мера која комбинује прецизност и одзив модела и пружа информације о равнотежи између тачно позитивних и лажно позитивних предикција, као и тачно позитивних и лажно негативних предикција. У свим класама је негде око 70%.

Матрица конфузије

На у-оси налазе се праве вредности класа, а на х-оси се налазе предвиђене класе. Пожељно је да вредности на дијагонали буду што више - да јефтини буду предвиђени као јефтини, средњи као средњи и скупи као скупи. Ова матрица је веома задовољавајућа. [1]

[1] интерпретација је преузета из jupyter свеске овог пројекта



Interpretacija rezultata:

Cheap/Affordable:

```
438 uzoraka su tacno predvidjeni kao Cheap/Affordable
127 uzoraka su bili predvidjeni kao Medium priced, a zapravo su Cheap/Affordable
44 uzoraka su bili predvidjeni kao Expensive, a zapravo su Cheap/Affordable
```

Medium priced:

```
414 uzoraka su tacno predvidjeni kao Medium priced
59 uzoraka su bili predvidjeni kao Cheap/Affordable, a zapravo su Medium priced
113 uzoraka su bili predvidjeni kao Expensive, a zapravo su Medium priced
```

Expensive:

```
473 uzoraka su tacno predvidjeni kao Expensive
138 uzoraka su bili predvidjeni kao Medium priced, a zapravo su Expensive
59 uzoraka su bili predvidjeni kao Cheap/Affordable, a zapravo su Expensive
```

Мере оптимизације

➔ **GridSearchCV**

Хипер-параметри су параметри који се не уче током тренирања и прослеђују се као аргументи конструктору модела.

GridSearchCV је техника за тражење оптималних параметара за креирање модела применом крос-валидације свих наведених опција за параметре. Уобичајен приступ њене примене код KHH алгоритма је избор између *gini* и *entropy* критеријума, и бирање максималне дубине стабла како би се постигао најбољи перформанс модела.

```
{'n_neighbors': 10, 'p': 1, 'weights': 'distance'}
```

- n_neighbors = 10: оптималан број суседа за процену је 10
- p = 1: Овај параметар дефинише рачунање растојања између података. Када је једнако јединици, користи се Менхетн растојање, а кад је једнако двојци, користи се Еуклидско растојање.
- weights = distance: Овај параметар дефинише како ће се узети у обзир удаљеност од суседа при процени. Ако је постављено на uniform, сви суседи имају исту важност. Пошто је постављено на distance, суседи који су ближи имају већи утицај при доношењу одлуке.

Classification report for model KNeighborsClassifier on training data

	precision	recall	f1-score	support
Cheap/Affordable	0.97	0.99	0.98	2437
Expensive	0.98	0.99	0.99	2678
Medium priced	0.99	0.96	0.98	2822
accuracy			0.98	7937
macro avg	0.98	0.98	0.98	7937
weighted avg	0.98	0.98	0.98	7937

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
Cheap/Affordable	0.74	0.73	0.73	609
Expensive	0.78	0.79	0.78	670
Medium priced	0.63	0.63	0.63	706
accuracy			0.71	1985
macro avg	0.71	0.71	0.71	1985
weighted avg	0.71	0.71	0.71	1985

Велика тачност на тренинг скупу а мала на тест скупу говори о томе да се модел преприлагодио и да има проблем при генерализацији података. Овај проблем се може решити другачијом поделом базе - на тест, тренинг и трећи, валидациони скуп. Валидациони скуп је скуп података који се користи за процену перформанси модела током процеса подешавања параметара.

Classification report for model KNeighborsClassifier on validation data

	precision	recall	f1-score	support
Cheap/Affordable	0.76	0.68	0.72	507
Expensive	0.73	0.76	0.74	543
Medium priced	0.57	0.60	0.58	538
accuracy			0.68	1588
macro avg	0.69	0.68	0.68	1588
weighted avg	0.68	0.68	0.68	1588

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
Cheap/Affordable	0.74	0.68	0.71	614
Expensive	0.73	0.78	0.75	663
Medium priced	0.60	0.60	0.60	708
accuracy			0.69	1985
macro avg	0.69	0.69	0.69	1985
weighted avg	0.69	0.69	0.68	1985

Модел још увек не показује напредније перформансе у односу на раније добијене резултате.

➔ Ансамбли

Ансамбли су један од најпопуларнијих метода за оптимизацију у машинском учењу. Заснивају се на комбиновању више модела како би они заједнички добили оптималне резултате.

Classification report for model RandomForestClassifier on test data

	precision	recall	f1-score	support
Cheap/Affordable	0.74	0.74	0.74	614
Expensive	0.76	0.79	0.78	663
Medium priced	0.63	0.61	0.62	708
accuracy			0.71	1985
macro avg	0.71	0.71	0.71	1985
weighted avg	0.71	0.71	0.71	1985

■ KHH у случају регресије

Пошто је цена апартмана нумеричка вредност, природније је користити KHH као регресиони модел. Кораци су веома слични као у случају класификације, са разликом метрике код одабира вредности за тест-инстанцу (не узима се најбројнија вредност међу K суседа већ аритметичка средина)
Битан фактор је изабрано K јер успешност модела у изразитој мери зависи од њега. Одабир K зависи од средње апсолутне грешке (MAE) и средње квадратне грешке (MCE) модела.

Mean Squared Error (MSE - Средња квадратна грешка):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- n - број примерака
- Y_i - стварна вредност за i-ти примерак

- $\hat{Y}_i$ - предвиђена вредност за i-ти примерак

Величина MCE може бити велика ако постоји велики размак између стварних и предвиђених вредности.

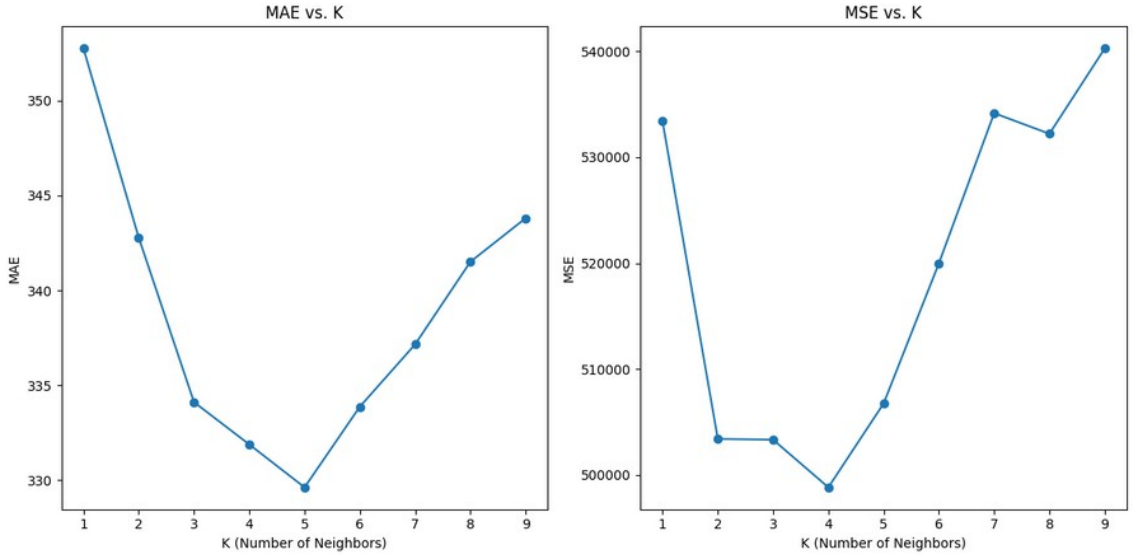
Mean Absolute Error (MAE - Средња апсолутна грешка):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

И овде важи исто:

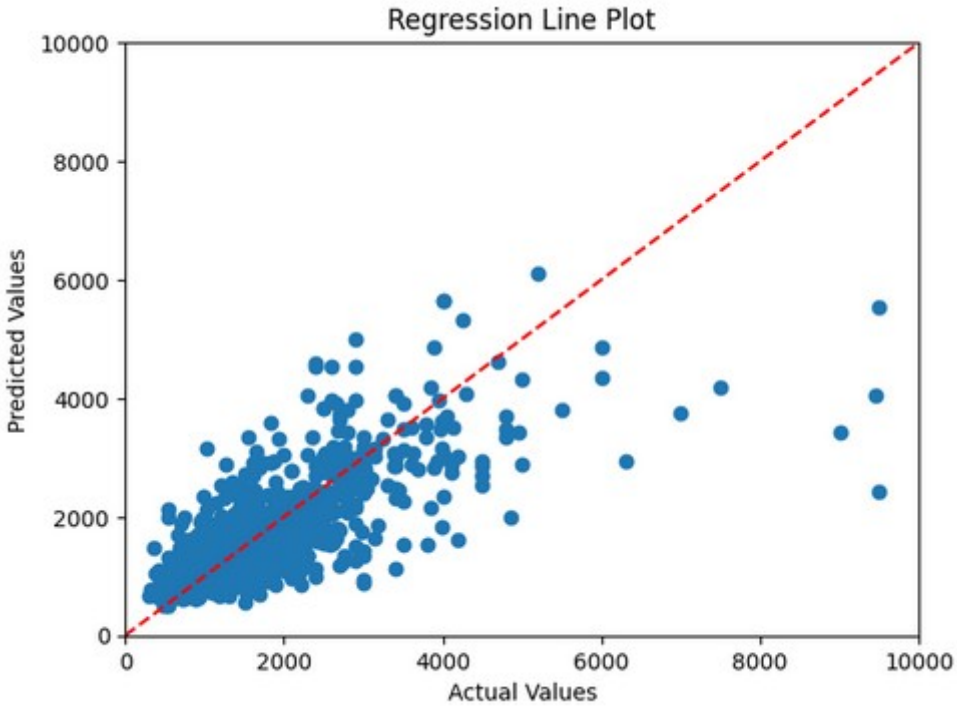
- n - број примерака
- Y_i - стварна вредност за i-ти примерак
- $\hat{Y}_i$ - предвиђена вредност за i-ти примерак

MAE је груб према великим аутлајерима и може бити бољи избор ако подаци имају значајне варијације



Приоритет при избору добиће оно K за које је MAE најмање. Разлог за овај корак је велика варијација цена.

Након тренирања модела, добија се следећа визуелизација перформансе:



Пожељно је да подаци што мање одударају од регресионе линије представљене црвеном бојом.

```
MAE: 329.6302267002519
MSE: 506751.8651889168
RMSE: 711.8650610817451
```

Модел се показује релативно добро у процени цене.

2.Стабла одлучивања

DecisionTrees

Стабло одлучивања је интуитиван алгоритам за класификацију.
Састоји се из следећих принципа:
- стабло се састоји из чворова и грана
- сваки чвор представља тест на одређеном атрибуту
- гране из чворова воде до других чворова или до листова који садрже коначну класу
- почетак стабла одлучивања је корен.

Стабло се гради рекурентно тако да се на сваком кораку бира најбољи атрибут за поделе. Критеријуми као што су Гинијев индекс или ентропија се користе за мерење чистоће података у сваком чвору. Циљ је смањити нечистоћу или несигурност у сваком чвору, а избор мере нечистоће нема велики утицај на перформансе.

Стабла одлучивања су применљива на све типове података и могу да представе сваку функцију дискретних атрибута.

Као и у претходно приказаном примеру, у фази претпроцесирања базе атрибут цене је дискретизован у 3 категорије: *Cheap/Affordable, Medium priced, Expensive.*
Класе су прилично балансиране па то неће представљати проблем при примени алгоритма.
Карактеристике ће и овај пут бити исте: квадратура, луксузност, географска ширина и географска дужина.

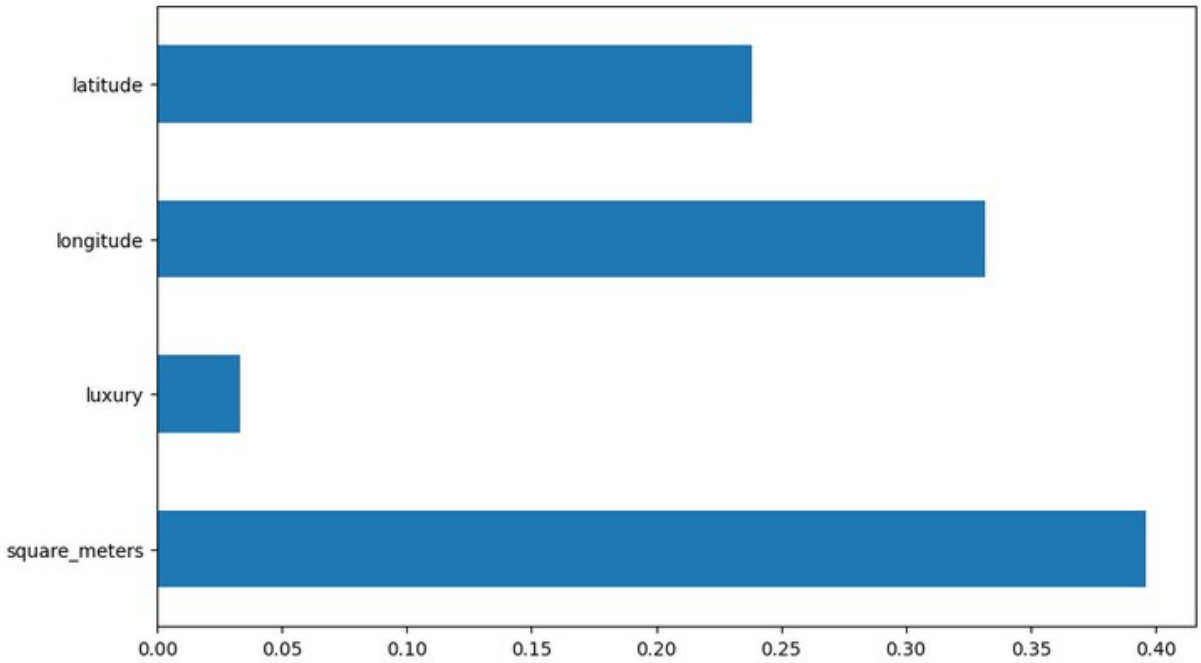
Classification report for model DecisionTreeClassifier on training data

	precision	recall	f1-score	support
Cheap/Affordable	0.97	0.99	0.98	2437
Expensive	0.98	0.99	0.99	2678
Medium priced	0.99	0.96	0.98	2822
accuracy			0.98	7937
macro avg	0.98	0.98	0.98	7937
weighted avg	0.98	0.98	0.98	7937

Classification report for model DecisionTreeClassifier on test data

	precision	recall	f1-score	support
Cheap/Affordable	0.71	0.73	0.72	609
Expensive	0.77	0.77	0.77	670
Medium priced	0.61	0.59	0.60	706
accuracy			0.70	1985
macro avg	0.70	0.70	0.70	1985
weighted avg	0.70	0.70	0.70	1985

И поново се јавља преприлагођеност. Могућ разлог за то је да је дрво сложеније него што је потребно. Корисно је имати увид у значајност карактеристика које се испитују.



Супротно полазној претпоставци, индикатор да ли је апартман луксузан или не и није толико пресудан при класификацији.
Да ли модел даје боље резултате уколико се димензија улазних података смањи? Уколико се из улазних карактеристика избаци '*luxury*' добијају се следећи резултати.

```
report(dtc, X_train, Y_train)
```

```
Classification report for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------
                  precision    recall  f1-score   support

Cheap/Affordable       0.89      0.98      0.93      2437
       Expensive       0.94      0.94      0.94      2678
   Medium priced       0.99      0.90      0.94      2822

        accuracy                           0.94      7937
       macro avg       0.94      0.94      0.94      7937
    weighted avg       0.94      0.94      0.94      7937

--------------------------------------------------------------------------
```

```
report(dtc, X_test, Y_test, "test")
```

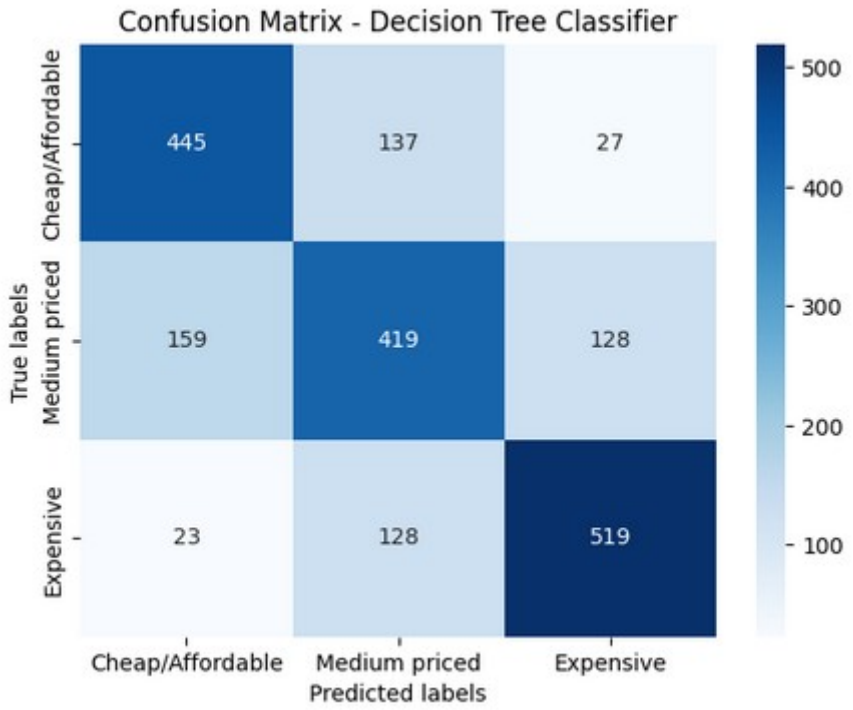
```
Classification report for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------
                  precision    recall  f1-score   support

Cheap/Affordable       0.30      0.31      0.31       609
       Expensive       0.32      0.33      0.33       670
   Medium priced       0.35      0.33      0.34       706

        accuracy                           0.32      1985
       macro avg       0.32      0.32      0.32      1985
    weighted avg       0.32      0.32      0.32      1985
```

Резултати су евидентно лошији, па ово није одговарајућ приступ. У даљим разматрањима и оптимизацијама радиће се искључиво на првобитном моделу који укључује '*luxury*' атрибут.

Матрица конфузије

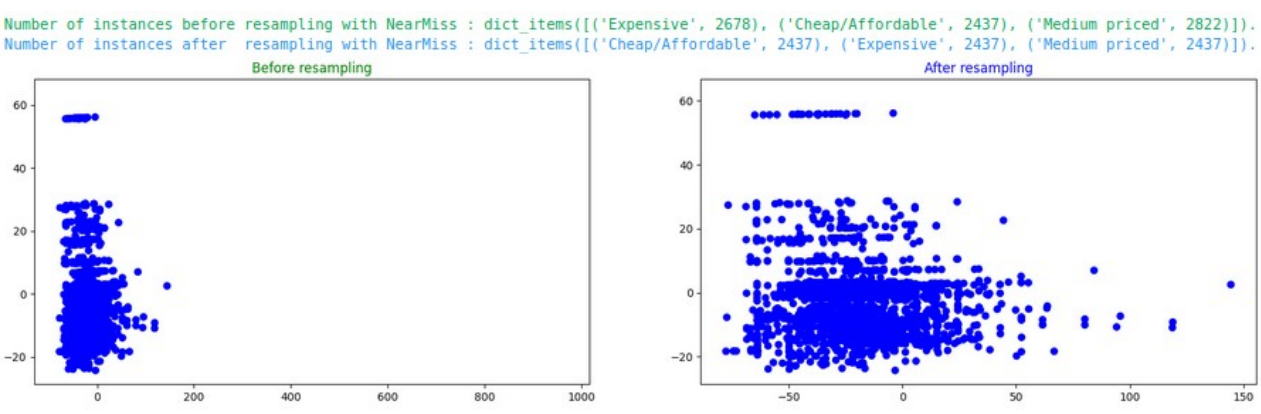


И овај пут матрица конфузије приказује одличне резултате у предвиђању класа - на дијагонали су убедљиво највише вредности.

Мере оптимизације

Приступ GridSearchCV и приступ базиран на случајним шумама је већ описан у претходном подеоку, стога у овом неће бити пуно речи о њему, већ о другачијем приступу - NearMiss.

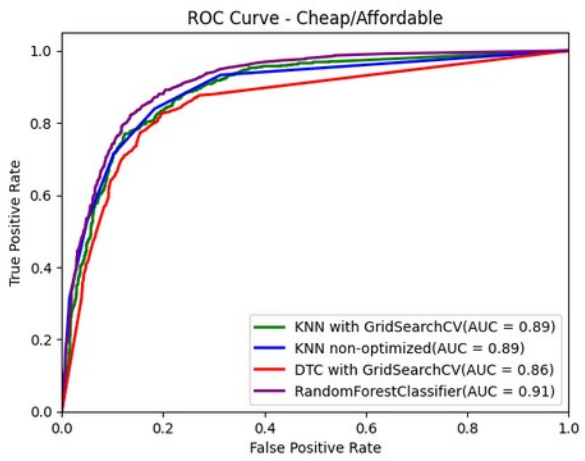
NearMiss бира инстанце већинске класе које имају најмање средње растојање од K најближих инстанци мањинске класе.

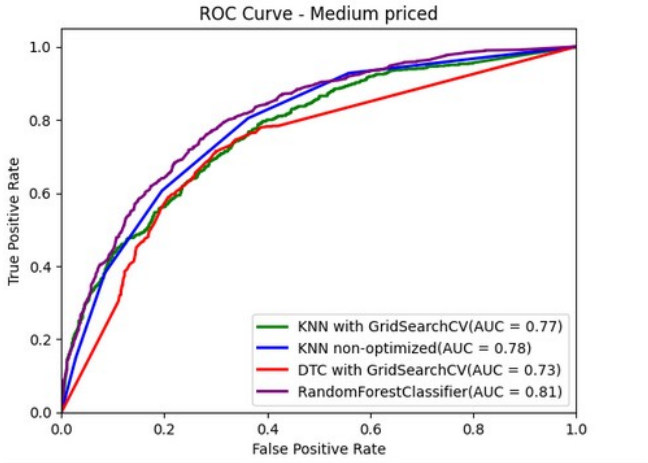


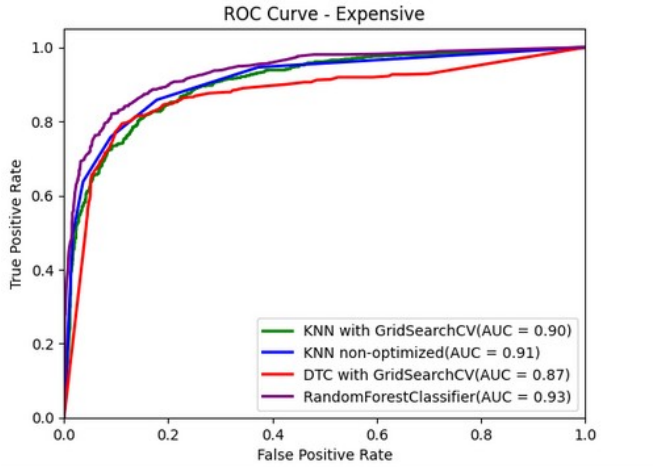
■ Поређење свих евалуираних модела

За визуелизацију добијених резултата и помоћ у одабиру најбољег модела користи се ROC крива. Крива анализира способност модела да раздвоји две класе (у случају бинарног класификатора), или у процени One Versus Rest у случају више класа.
Што је већа површина испод криве (AUC - area under curve), модел је бољи. Ова крива помаже у процени способности модела да раздвоји позитивне и негативне класе. Процена перформанси модела је вредност која је најмање удаљена од тачке (0,1).
На графику испод се може видети да се најбоље показао стандардни KHH модел.

За бољи увид у ефикасност свих наведених модела, корисно је погледати њихов учинак према свакој од издвојених класа. У сваком случају, као најбољи се издвојио RandomForestClassifer.







Кластеровање

Кластеровање је проналажење група објеката таквих да су објекти у групи међусобно слични (или повезани) и да су објекти у различитим групама међусобно различити (или неповезани). Карактеристика добро раздвојених кластера је да им припадају елементи такви да су ближе било ком другом елементу у кластеру него осталим елементима који нису у кластеру.
Могу бити корисни приликом анализе тржишта, постављања маркетиншких циљева и персонализовања понуда које фирма пружа.
На конкретном примеру биће приказано ексклузивно, комплетно, нехомогено кластеровање користећи алгоритам K-средина, хијерархијско кластеровање и DBSCAN алгоритам.[2]

1. K-средина

K-means

Свака тачка кластеру је ближа прототипу (центроиду) кластера у односу на прототипове осталих кластера. Центар кластера је центроид - просек свих тачака у групи.

[2] ексклузивно - појединачни елемент се може истовремено налазити само у једном кластеру
комплетно - комплетан скуп учествује у кластеровању
нехомогено - кластери су различите величине

Уместо да интерно одреди оптималан број кластера, овај алгоритам захтева да се тај број наведе унапред. Центроиде се бирају насумично. Иницијално, свака тачка ће припадати оном кластеру чијем је центроиду најближа. Итеративно се израчунавају нове центроиде и поново израчунавају припадности сваке тачке појединим кластеру. Алгоритам се зауставља када број тачака које промене кластер падне испод унапред задатог прага.

Напомена: За добијање оперативних резултата, неким државама су додељени посебни 'статуси' који говоре о гласу на ком се та држава налази. У причи се налазе 3 статуса: Nature-lover, Family-friendly, Turbulent. Тако су на пример државе као што су Њујорк и Вашингтон добиле статус Turbulent што значи да су жижа дешавања, одликује их доста пословних прилика и привлаче млађу публику . Државе као што су Јута и Калифорнија су добиле статус Nature-lover због прелепих предела. Државе као што су Масачусетс и Минесота су добиле статус Family-friendly јер су по одређеним критеријумима погодне за подизање породице. Извор за сваки статус су биле статистике нађене на интернету.

Избор броја кластера K је важан за испитивање да ли су добијени резултати коректни. Можда се за неки други број K добије бољи резултат. Мере које се користе за процену квалитета резултата укључују:

- SSE - sum of squared erros
- $SSE = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)^2$ po c_i, gde je ***d*** euklidsko rastojanje (ali je moguće koristiti i neko drugo).

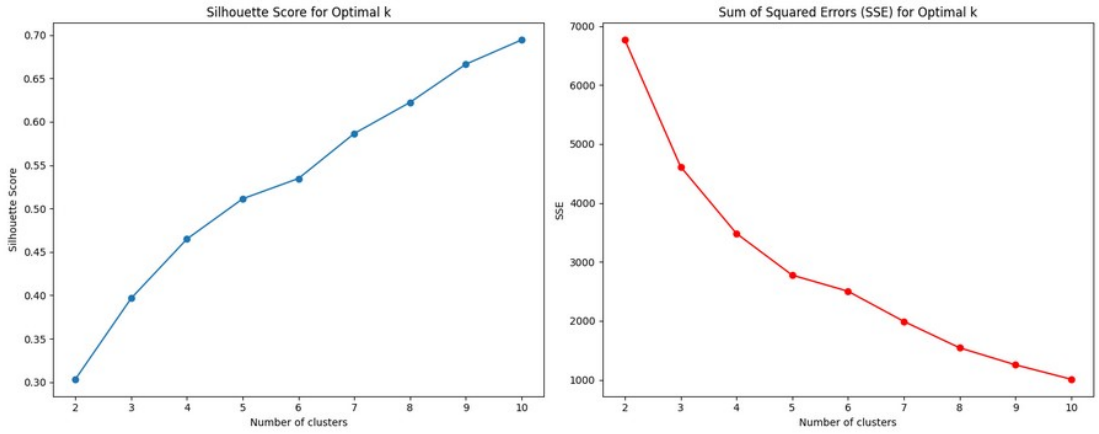
кохезија - мери степен сличности елемената у кластеру са центроидом

- сепарација - мери колико су различити објекти из различитих кластера
- коефицијент сенке - комбинација сепарације и кохезије

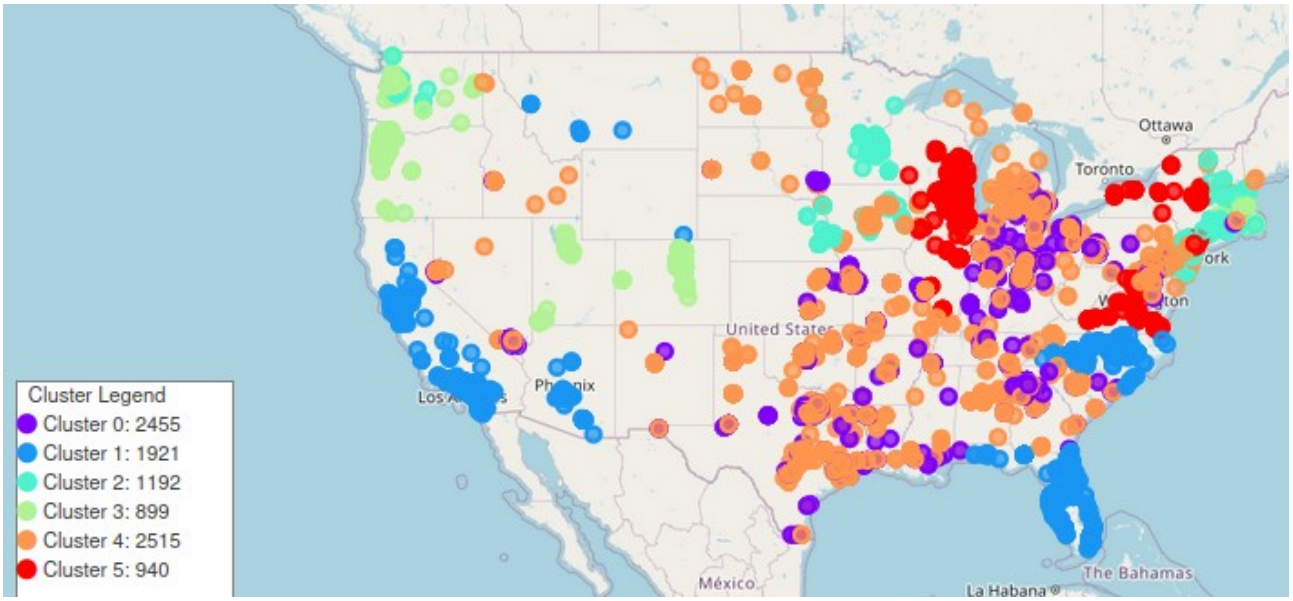
Koeficijent senke se računa za svaku instancu pojedinačno, na sledeći način:

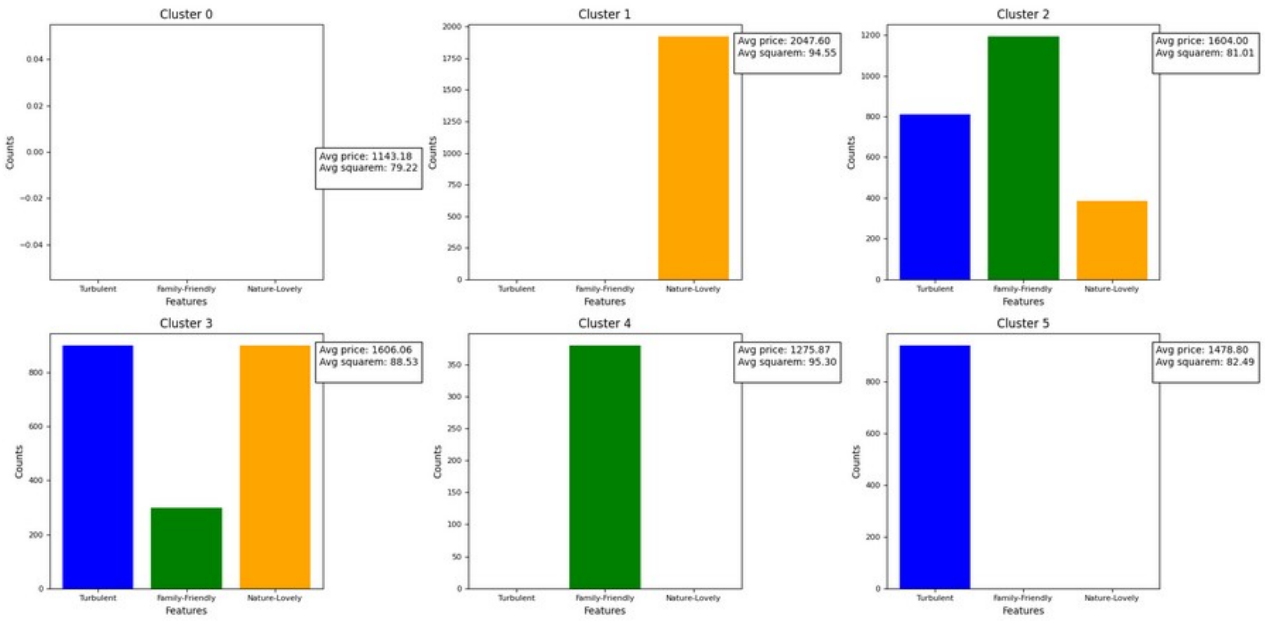
1. Za i-tu instancu računamo usrednjeno rastojanje od svih instanci iz istog klastera (Umesto rastojanja može da se koriste i druge mere različitosti). Obeležimo izračunatu vrednost sa a_i.
2. Za i-tu instancu i sve klastere koji ne sadrže i-tu instancu računamo usrednjeno rastojanje instance od svih elemenata iz svakog klastera. Pronalazimo minimalno rastojanje i obeležimo ga sa b_i.
3. silhouette coefficient za i-tu instancu računamo : $s_i = \frac{b_i - a_i}{max(a_i, b_i)}$

Једна од најпознатијих метода је правило лакта. Након визуелизације промене SSE бира се број кластера на лакту, тј. у тачки где SSE најбрже опадне. Проблем може да настане када се веома блиски кластери спајају у један, јер њиховим раздвајањем SSE неће много опасти. Решење овог проблема је коришћење коефицијента сенки. Велике вредности коефицијента сигнализирају добру кохезију унутар кластера и сепарацију између кластера.



Што је већи број кластера, већи је коефицијент сенки, али већи број кластера доводи до тешко интерпретабилних резултата. У наставку ће бити приказани резултати добијени поделом на 6 кластера, што се емпиријски показало као најделотворније.





Сваки кластер одговара одређеном профилу клијента. Ни један кластер се не састоји искључиво од станова чија је локација у неким од држава 'на гласу', већ приказује њихову доминантност у одређеној регији.
Нулти кластер садржи по просеку најјефтиније и најмање станове, и на мапи се налазе раштркани по источном делу Америке и он је погодан за оне који немају велике материјалне ресурсе. У првом кластеру доминирају апартмани из Nature-lover категорије и уједно је и најскупљи, погодан је за пензионере, авантуристе и оне који имају већу финансијску стабилност. [3]

Алгоритам K-средина је једноставан и флексибилан. Важно је напоменути да може бити више подједнако добрих подела (више глобалних минимума). Мана је што је осетљив на аутлајере због квадрирања Еуклидског растојања и они могу да доведу до појединачних или празних кластера.
Временска сложеност: O(n * K * I * d)
Просторна сложеност: O((n + K) * d),
где је n број тачака, K број кластера, I број итерација и d број атрибута.

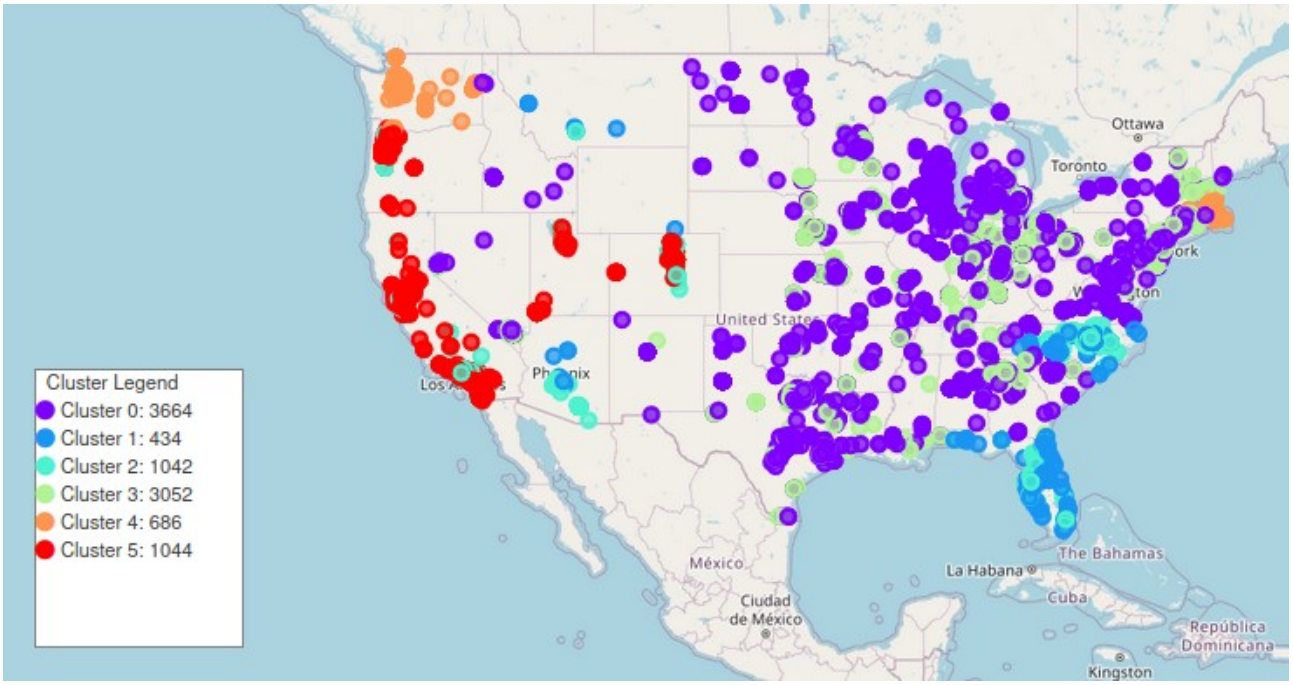
2. Хијерархијско спајајуће кластеровање

Agglomerative clustering

Основна идеја ове групе алгоритама је формирање скупа угњежђених кластера који су организовани у облику хијерархије по нивоима. Резултати се најчешће визуализују у облику дендограма или дијаграма са угнежденим кластерима на којима се виде однос и редослед формирања кластера.
Код спајајућег кластеровања, хијерархија се формира одоздо-навише. У почетку се свака тачка посматра као посебан кластер, и у сваком од наредних корака се врши спајање два најближа.

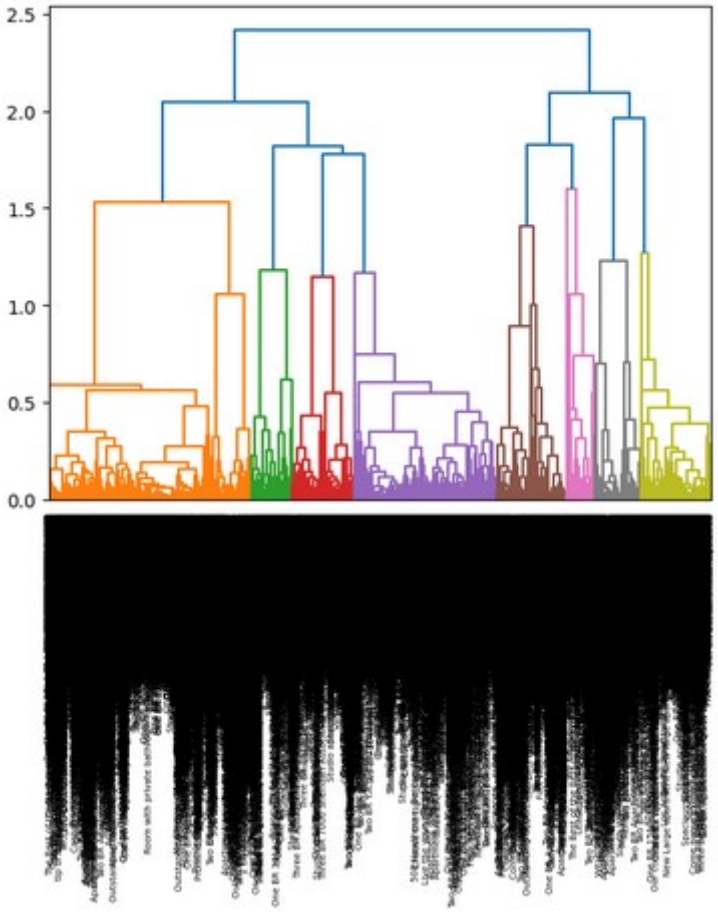
[3] Детаљнија анализа кластера се може наћи у jupyter свескама намењеним за овај пројекат

Детаљнија анализа сваког кластера ће овог пута бити изостављена јер је сврха овог рада демонстрација различитих алгоритама за истраживање података.



Наизглед, додуше, изгледа као да је ситуација слична малопређашњој, са разликом у бојама.

Као средство визуелизације за хијерархијско кластеровање углавном се узима дендограм. Нажалост због количине инстанци у бази података, резултати нису читљиви.
Што су веће удаљености између две инстанце, њихове карактеристике се више разликују.
У овој имплементацији није искоришћено одсецање уколико се дође до одређеног нивоа хијерархије.



Недостаци агломеративног кластеровања су:

- после комбиновања кластери не могу да се раздвоје
- не постоји глобална функција која директно минимизује
- могу да се јаве осетљивост на шум и елементе ван граница, тешкоће у обради кластера различитих величина, тенденција ка разбијању великих кластера

За алтернацију и покушај побољшања добијених резултата, у пракси се користе различите везе (*linkage*) или методе за рачунање сличности, па тако се уместо овде искоришћене везе *complete*, где се удаљеност између кластера мери као највећа удаљеност између тачака у различитим кластерима, могу користити везе *linkage*[4] или *average*[5]. За методе рачунања сличности може се употребити нпр. Ward-ов метод у коме се уместо варијансе користи збир квадрата грешака. Нови кластер се формира од два кластера чијим спајањем се добија минимално повећање збира квадрата грешака унутар новог кластера.

3. DBSCAN

Density-Based Spatial Clustering of Applications with Noise

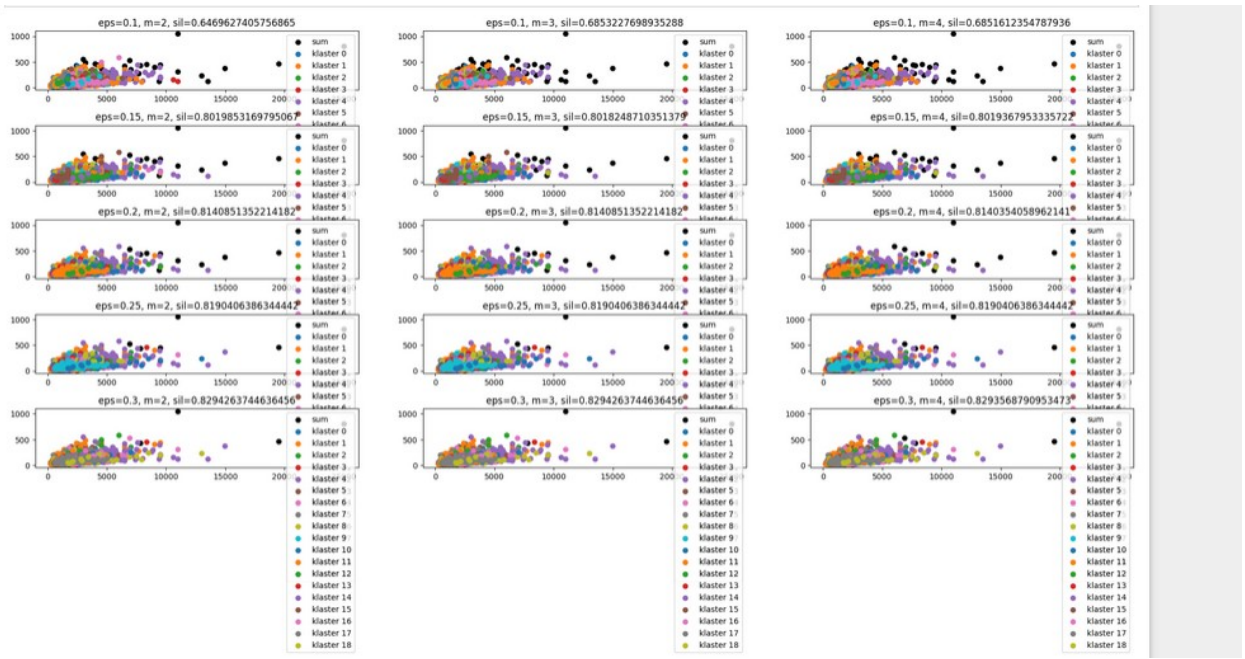
DBSCAN се истиче по својој способности да открије скривене обрасце и структуре у подацима на основу густине распореда тачака. Флексибилан је и није неопходно дефинисање броја кластера унапред. Уместо тога, овај алгоритам динамички идентификује кластере на основу густоће тачака у подацима. Кластери се формирају око централних тачака високе густине, а раздвајају регионима ниске густине, чиме омогућава откривање кластера различитих облика и величина. Отпоран је на шумове и може их прецизно идентификовати.

Алгоритам је аутоматски нашао око 20 кластера и коефицијент сенки му је око 0.83

Epsilon и *min_samples* су два кључна параметра, где *epsilon* представља радијус око сваке тачке, а *min_samples* одређује минималан број суседа који морају бити унутар радијуса за тачку да буде класификована као део кластера.

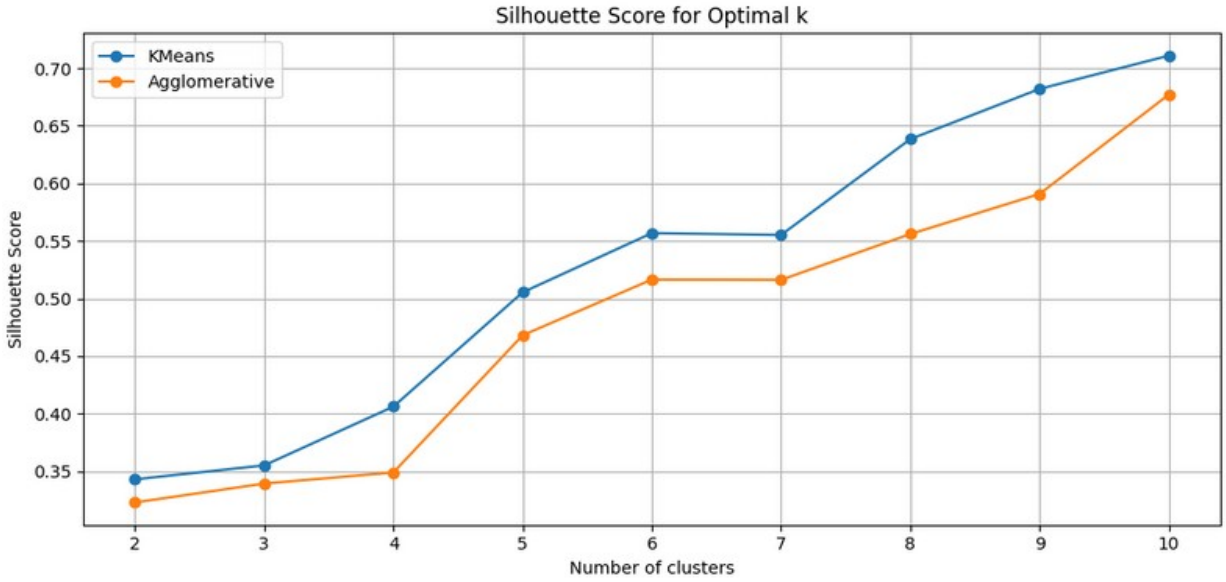
[4] Удаљеност између кластера се мери као најмања удаљеност између тачака у различитим кластерима. Ово значи да ће се кластери спојити ако постоји бар једна тачка у сваком од кластера која је блиска једна другој.

[5] Овај тип везе користи средњу вредност удаљености између свих тачака у различитим кластерима.



Поређење свих добијених модела

на основу коефицијента сенки



Са графика се може закључити да је алгоритам K-средина бољи за број кластера K = 6. Може се експериментисати са различитим бројем K и различитим параметрима. Међутим, човека теорија може довести само до одређене границе, тако да је пракса најбољи показатељ мере квалитета модела.

Правила придруживања

Процес у ком се проналазе корелације између појављивања изоловане ставке (објекта) и њене присутности у скупу ставки из трансакција (слогова). Често се користи да послодавци открију скривене или не тако очигледне везе које ће им помоћи у унапређивању свог бизниса.
Класичан пример за илустрацију метода за одређивање правила придруживања користи се потрошачка корпа. Свака потрошачка корпа је скуп артикала које је купио појединачан купац и она представља једну трансакцију.

Основни појмови:

- подршка правила (*support*) мери колико се одређено правило налази у трансакцијама. Издваја правила која нуде више могућности за добијање корисних информација.
- поузданост правила (confidence) A => B мери колико је вероватно да се B појави уколико се појави A. Високо поверење указује на правило које је од интереса.
- лифт је једна од мера која узима у обзир подршку десног дела правила A => B. Лифт вредности веће од 1 означавају да је последична страна правила много чешћа у трансакцијама које садрже леву страну правила него у трансакцијама које је не садрже, док вредност мања од 1 означавају правила чија је поузданост мања од очекиване.

Обично се минимални ниво поверења поставља високо (око 80%), а минимални ниво подршке око 5-10%, због велике разноликости у случају јако великих база података.
Најпростији начин за одређивање правила придруживања је пука примена грубе силе, што је изузетно временски и рачунски захтевно. Начини за оптимизацију укључују смањење броја кандидатских стваки, смањење броја трансакција или смањење броја поређења. Априори алгоритам, обрађен у наставку, фокусира се на први начин.

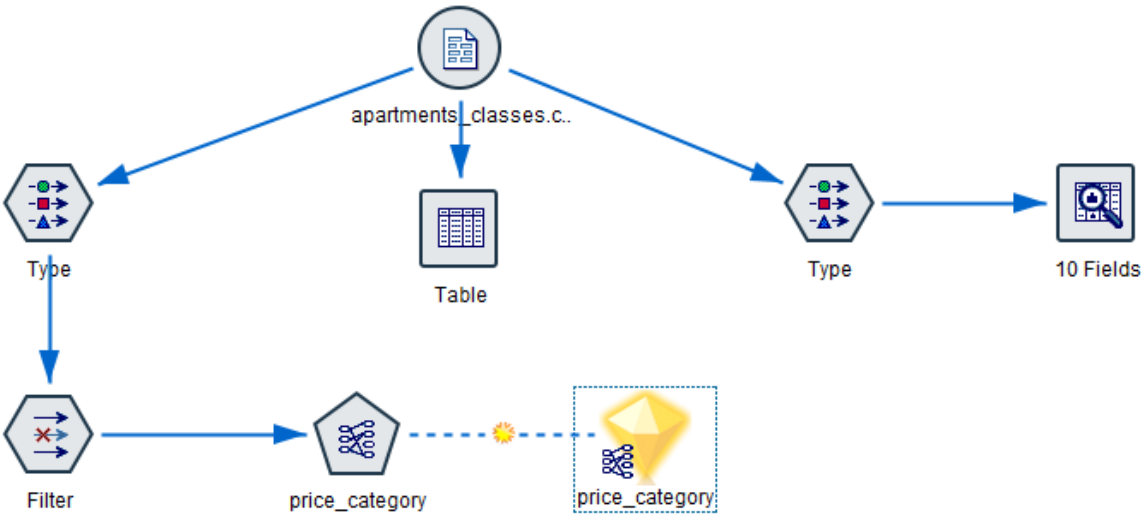
Априори алгоритам

Алгоритам грубе силе конструише решетку са свим могућим подскуповима и правилима, израчунава подршку и поузданост за сваки од њих и врши селекцију оних који задовољавају кориснички дефинисан праг. Априори алгоритам смањује количину разматраних скупова тако што примењује превремено одсецање подскупова на основу учесталости њиховог надскупа (downward closure). Просто речено, ако је неки скуп ставки чест, тада су и сви његови

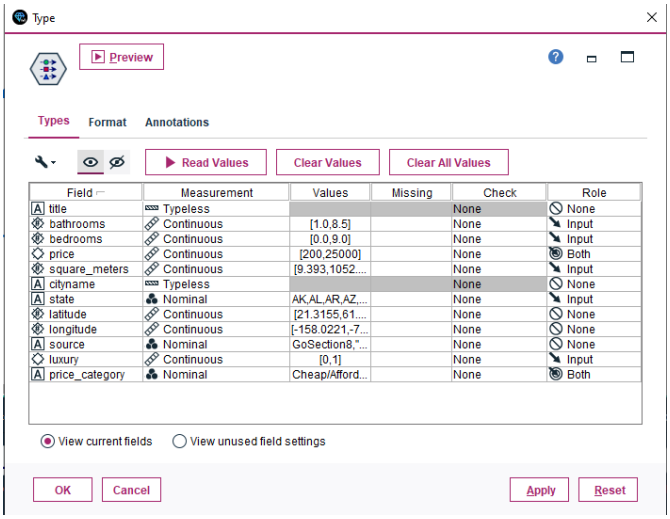
подскупови такође чести, и обрнуто, ако неки скуп артикала није чест, онда ни његови подскупови нису чести, па нема смисла испитивати их и одбацују се.

Смањење броја поређења:
Алгоритам смешта ставке у решетку у ширину и групише их у корпе по дужини скупова. Корпе су представљене хеш структуром са фиксним бројем грана у чвору. На сваком нивоу решетке примењује се хеш функција на одговарајућу ставку скупа. Потом се увећава број ставки које дођу у листовима хеш структуре. Поређење ставки из трансакција врши се са садржајем корпе уместо целим скупом кандидата. Овим се смањује број поређења и обрада, што помаже у одређивању броја појављивања скупа ставки (подршке).

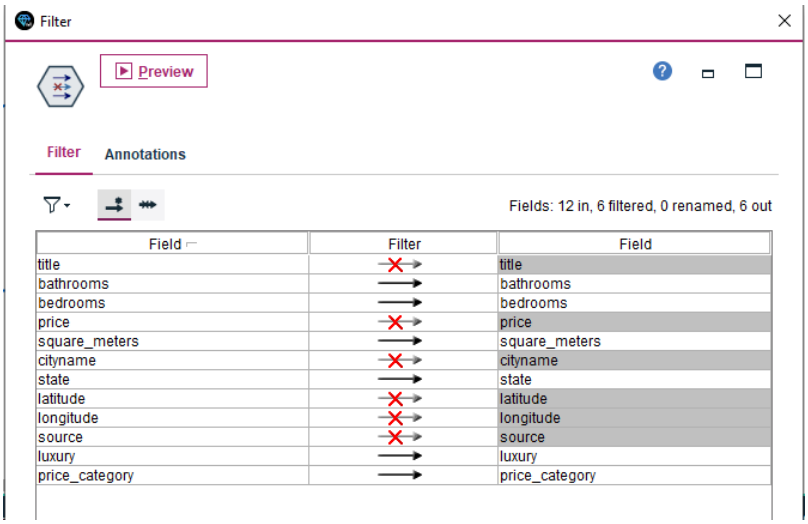
У пракси се користи SPSS Modeler за имплементацију Априори алгоритма. У конкретном случају базе података са апартманима за изнајмљивање, једног купца представљаће један апартман, а садржај корпе представљаће његови атрибути.



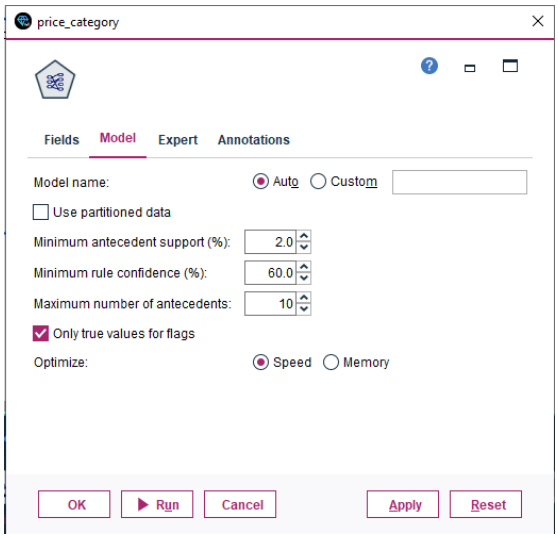
- Корен графа је улазни фајл који садржи листу апартмана са атрибутима и дискретизацијом поља цене као што је то урађено код класификације.
- Чвор *type* одређује функцију атрибута, да ли су улазни, излазни, или оба



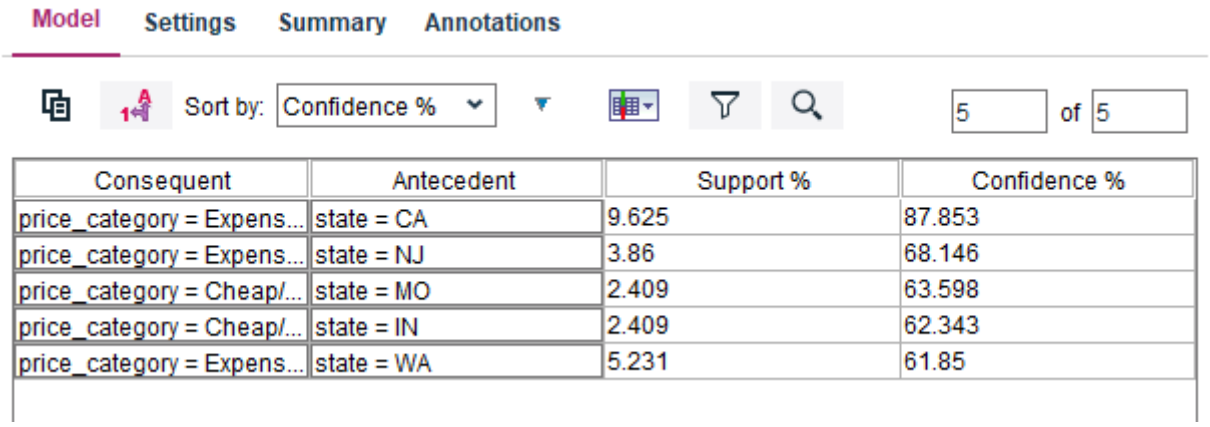
- Чвор *filter* уклања некорисне атрибуте из рачунице



- Чвор *Apriori* омогућава кориснику да унесе жељени праг подршке и поузданости. У овом примеру, захтеви су прилично слаби због разноликости података, а практично нема смисла ићи испод 60% за поузданост.



Добијени резултати:



Consequent	Antecedent	Support %	Confidence %
price_category = Expens...	state = CA	9.625	87.853
price_category = Expens...	state = NJ	3.86	68.146
price_category = Cheap/...	state = MO	2.409	63.598
price_category = Cheap/...	state = IN	2.409	62.343
price_category = Expens...	state = WA	5.231	61.85

Најјаче правило гласи да се, ако се апартман налази у Калифорнији, са 87% сигурности може тврдити да припада скупој категорији, и то правило подржава 9.6% скупа.
Иако су услови постављени прилично неригорозни, број правила добијен алгоритмом није велики, што може бити до кардиналности саме базе (можда је потребан већи скуп) или подаци нису довољно колерисани међу собом.

Априори алгоритам има широку примену и лак је за имплементацију чиме је заслужио звање најпопуларнијег алгоритма за проналажење правила придруживања. После њега долазе FP-growth и Eclat алгоритам.

Закључак

Истраживање података нуди моћна оружја за разлагање сирових информација које се касније могу трансформисати у нешто невидљиво голим оком. На статистичару је да прикупи те информације, на програмеру да употреби оружје, а на добром дата-аналитичару је да резултате преведе на језик који сви разумеју.

"Data will talk to you if you're willing to listen." - Jim Bergeson

Ресурси

1. http://poincare.matf.bg.ac.rs/~nenad.mitic/ip1.html
2. https://github.com/MATF-istrazivanje-podataka-1/materijali_2022-2023
3. https://www.linkedin.com/pulse/clustering-affordable-rental-housing-developments-martinez-quintero?trk=public_profile_article_view
4. https://medium.com/mlearning-ai/designing-a-optimal-knn-regression-model-for-predicting-house-price-with-boston-housing-dataset-faef377536e3
5. https://www.cnbc.com/2022/05/25/these-are-the-best-and-worst-states-to-live-in-for-millennials-in-2022.html
6. https://www.cnbc.com/2023/01/18/best-states-to-raise-a-family-in-the-united-states.html
7. https://www.travelandleisure.com/naturally-beautiful-states-in-the-country-6543572
8. https://icdm.zhonghuapu.com/algorithms/10Algorithms-08.pdf