Вовед во науки за податоци (база)

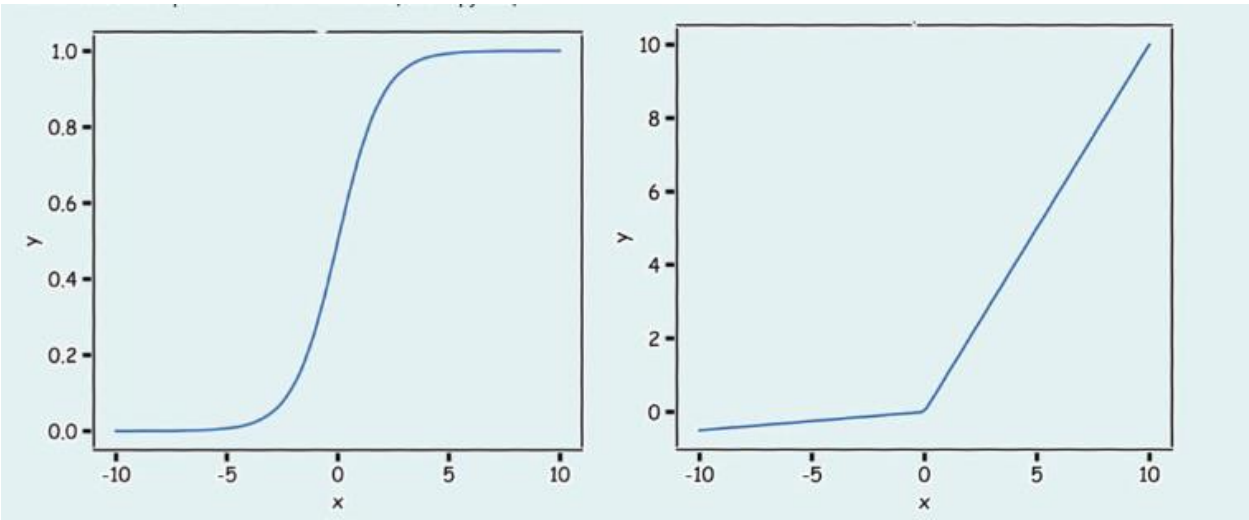
1. За дадената табела во прилог ако се енкодира колоната **Class** со помош на Binary Encoding, како ќе изгледа ново добиената табела?

Id	Company_name	Class
1	Pepsi	Drink
2	Zara	Clothes
3	Coca - Cola	Drink
4	Apple	Technology
5	Mcdonald's	Food

a.

Id	Company_name	Class_1	Class_2	Class_3
1	Pepsi	1	0	0
2	Zara	0	1	0
3	Coca - Cola	1	0	0
4	Apple	1	1	0
5	Mcdonald's	1	1	1

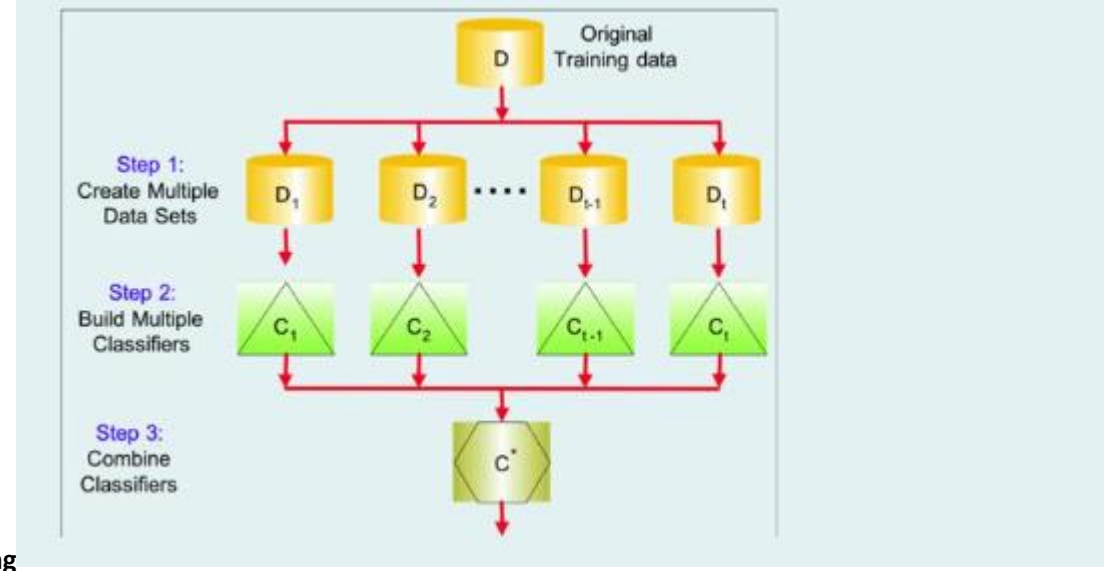
2. Ако податоците се означени кој ред во која класа припаѓа, тогаш станува збор за **Supervised learning**
   Ако податоците не се означени со припадност по класи ,тогаш станува збор за **unsupervised learning.**
   Ако при обуката се добиваат само сигнали дали е нешто добро научено или не, тогаш станува збор за **reinforcement learning.**
3. На сликата се прикажани кои активациски функции ?



Sigmoid Relu

4. Кои од наведените може да се користат како критериум за поделба (Splitting Criterium) на јазлите кај дрвата за одлучување?
   - **Грешка при класификација , Ентропија , Индексот Џини**
5. Кои од следниве репозиториуми/библиотеки се користат за едноставно споделување на претренираните NLP модели.
   - **HuggingFace transformesrs library, PyTorch Hub, TensorFlow-Hub**
6. Кои од следниве се називи на алгоритми за оптимизација кај невронските мрежи?
   - **Adam , Adagrad**

7. Шема за кој вид на учење со ансамбли?



Bagging

8. Кои од наведените дескриптивни статистики е најдобро да се изберат ако податочното множество се состои од континуирани податоци и притоа е потребно да се прикаже варијација на истите?
   - **Ранг , Интер-квартална разлика, Стандардна Варијација**
9. Што претставува Parts of Speech Tagging
   - **Процес на означување на збор во текст што одговара на категоријата зборови (или , поопшто лексички единици) кои имаат слични граматички својства (именки глаголи)**
10. Ако треба да се идентификува кога ќе се случи некоја необична трансакција при плаќањето, за каков вид на машинско учење станува збор?
    – Регресија , Откривање на недостатоци , Класификација, Учење со потикнување (не знам кој е точен)
11. Ако се подели множество на повеќе делови и потоа се остава едно за тестирање, а другите се користат за обука, за која техника станува збор.
    - **Cross Validation**
12. Што печати дадениот код:

Што печати дадениот код:

```
url = "https://sitel.com.mk/"
html = requests.get(url).text
soup = BeautifulSoup(html, 'html.parser')

tags = soup('img')
for tag in tags:
   print(tag.get('src', None))
```

a. Ги принта сите сликите од веб страната "sitel.mk"
b. Ја враќа и прикажува веб страната
c. Ги принта сите извори на сликите 'img' под 'src' тагот на веб страната "sitel.mk"
d. Ја симнува веб страната
e. Ниту едно од наведените

Clear my choice

13. Нека се дадени следниве променливи X и Y

X	Y
1	3
2	3
Null	4
4	3
1	2

Ако се користи k-NN метод со k=2 за пополнување на податоците кои недостасуваат, со која вредност ќе зе замени **Null** за променливата X.

- [x] a. Неможе да се определи
- [ ] b. 3
- [ ] c. 4
- [ ] d. 2

14. Кога дистрибуцијата на податоците е наклонета на десно , што се случува со средната вредност и медијаната кај овие податоци?

-**Медијаната е поголема од средната вредност**

15. За табелата df:

	house_type	price
0	Flat	100000
1	House 1 floor	200000
2	House 2 floors	300000

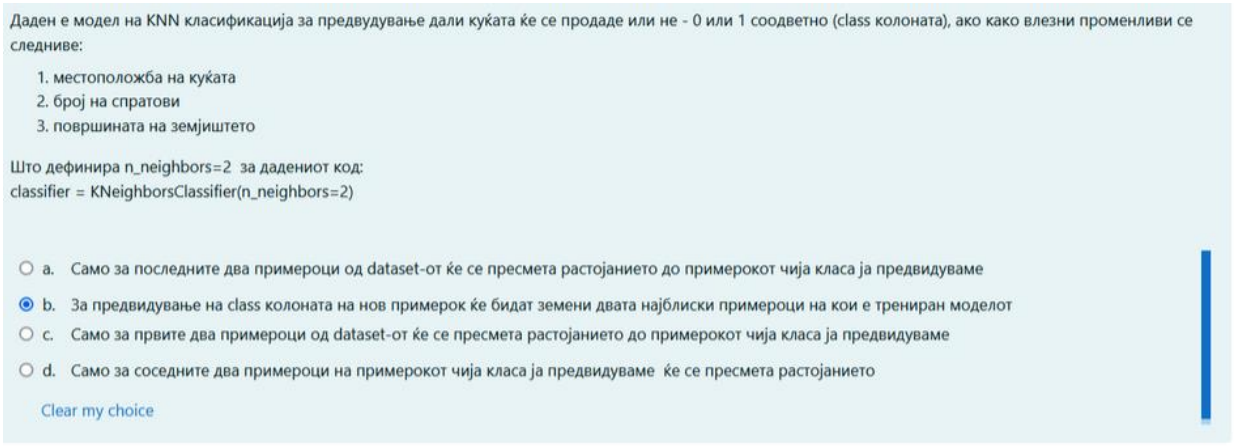
Што резултат ќе изгенерира дадениот код:
df.groupby(['house_type'],as_index=False).mean()

- ○ a. Средната вредност типот на куќи (колоната house_type)
- ○ b. Ниту едно од наведените
- ◉ c. Средната вредност на цената за секој тип на куќа
- ○ d. Средната вредност на цената за сите куќи

Clear my choice

16. За дадениот код: x = {i for i in range(0,15)) кој тип ќе биде променливата x и кои вредности ќе ги содржи
    - **X е од тип set и ги содржи вредностите 0,1,2,3,4,5,6,7,8,9,10,11,12,13,14**
17. Која од наведените кодови ќе ги избрише сите редици кои имаат NAN вредност?

Df.dropna(axis)=0



18.
19. За дадениот модел : model = DecisionTreeClasifier()
    Кој параметар треба дополнително да се додаде како аргумент во зградите да се наведе максимална длабочина на дрвото да е 10?
    -max_depth =10
20. Селектирајте ги особините кои се карактеристични за Data Science а не и за Machine Learning
    -Take action! , Develop/use tools that can handle massive datasets

21. На кои од наведените модели за кластирање потребно е да се наведе бројот на кластери?

    -**K-Means Clustering**

22. На кој начин се добиваат embeddings на зборовите при тренирање на BERT модел?

    Се зима излезот на моделот

23. Кај обработка на NLP се среќаваат следните задачи

    Препознавање на именувани нешта, Категоризација на теми

24. Еден од најдобрите јазични модели GPT-2 се потпира на трансформер архитектурата. Кој

    дел од трансформер архитектурата се користи во GPT-2

    Decoder

25. Кој од дадените алгоритми за кластерирање може да се добијат 6 кластери како што се

    прикажани на сликата :

 **DBSCAN and Hierarchical**

26. Во кој случај би било најдобро да се употреби Softmax како излезно ниво кај невронските мрежи?

    Дасда?????????

27. Што претставува n_estimators = 5 во XGBoost моделот?

    - **5 дрва на одлука кои паралелно ќе се изградат**

28. Кои карактеристики треба да ги има активациската функција кај невронските мрежи ?

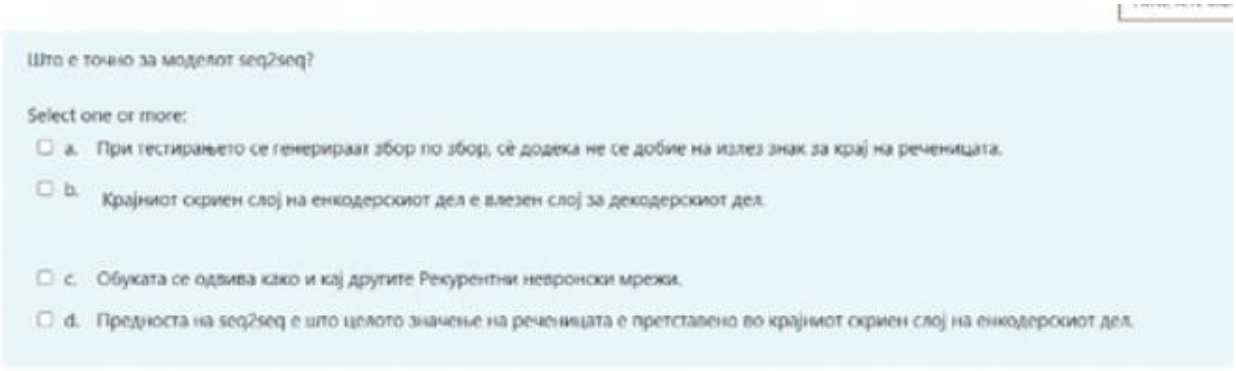
    Да има некаква нелинеарност

29.

1. Кои од следниве карактеристики се новитети кај Трансформер моделите?
   Tokenization, Feedforward Network, Positional embeddings, Self Attention layer
   Одговор: POSITION EMBEDDINGS, SELF ATTENTION LAYER

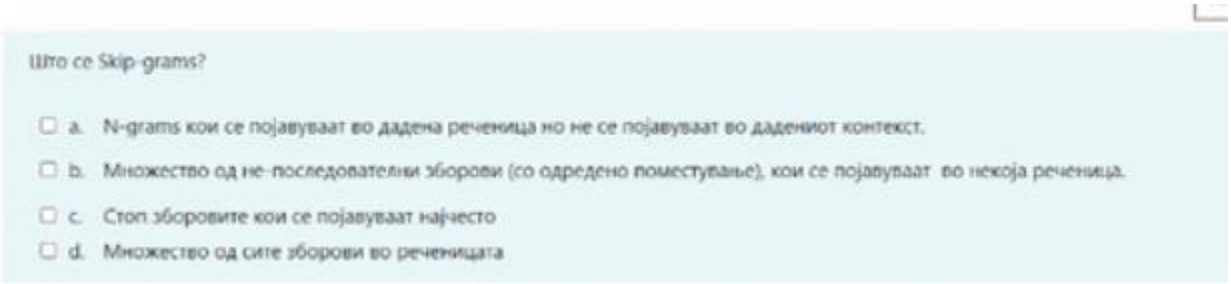
2. Каква димензионалност треба да е влезно тренирачкото множество кај LSTM невронската мрежа?
   3D, 1D, 2D-матрица
   Одговор: 2d-матрица

3. Што е точно за моделот seq2seq



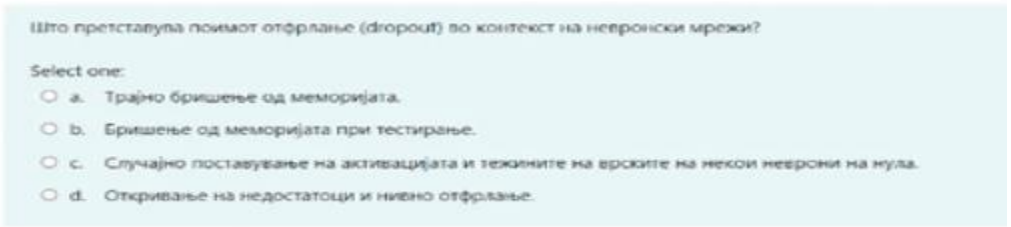
   Одговор:

4. Што е skip grams?



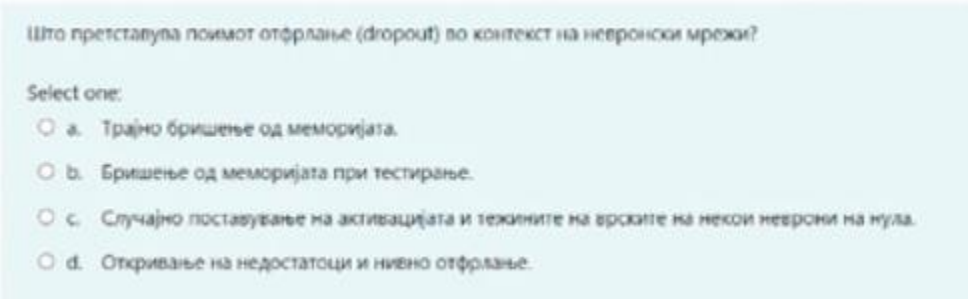
   Одговор: Множество од не-последователни зборови кои се појавуваат во некоја реченица

5. Што претставува поимот отфрлање во контекст на невронски мрежи?



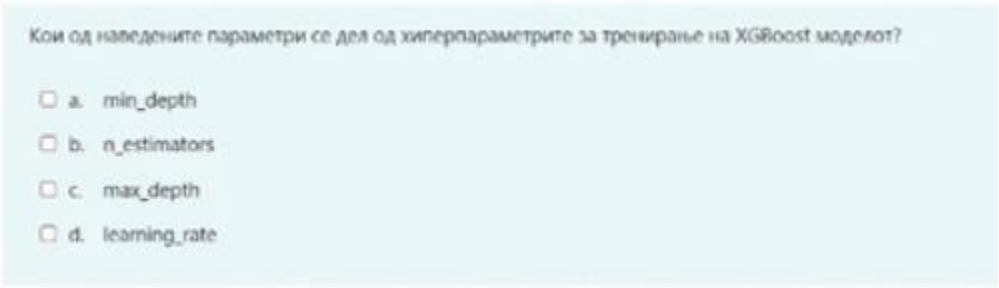
   Одговор: случајно претставување на активација и техниките на врските на некои невронски на нула

6. Кои се предностите на Long short term memory LSTM мрежите?



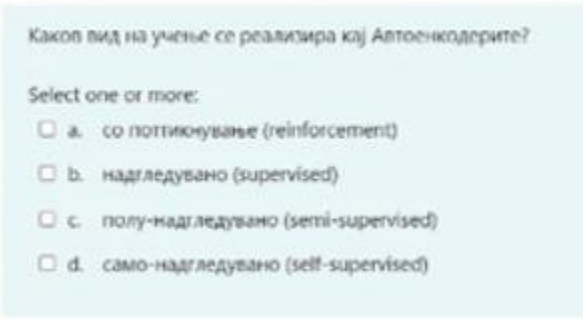
Одговор: Можност за учење на долги низи

7. Кои од наведените параметри се дел од хиперпараметритте за тренирање на XGBOOST моделот?



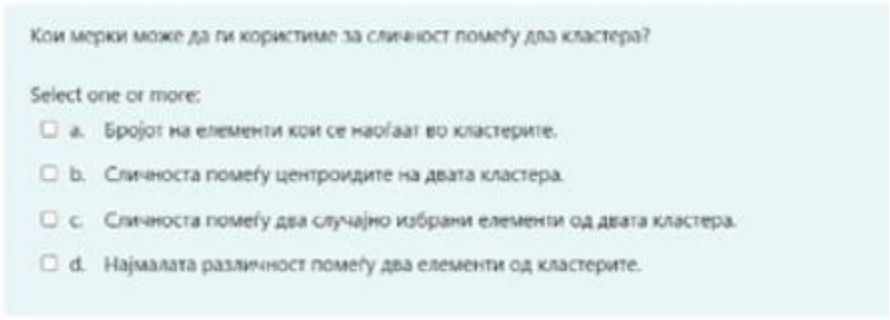
Одговор: n_estimators, learning_Rate, max_Depth

8. Каков вид на учење се реализира кај Автоенкодерите?



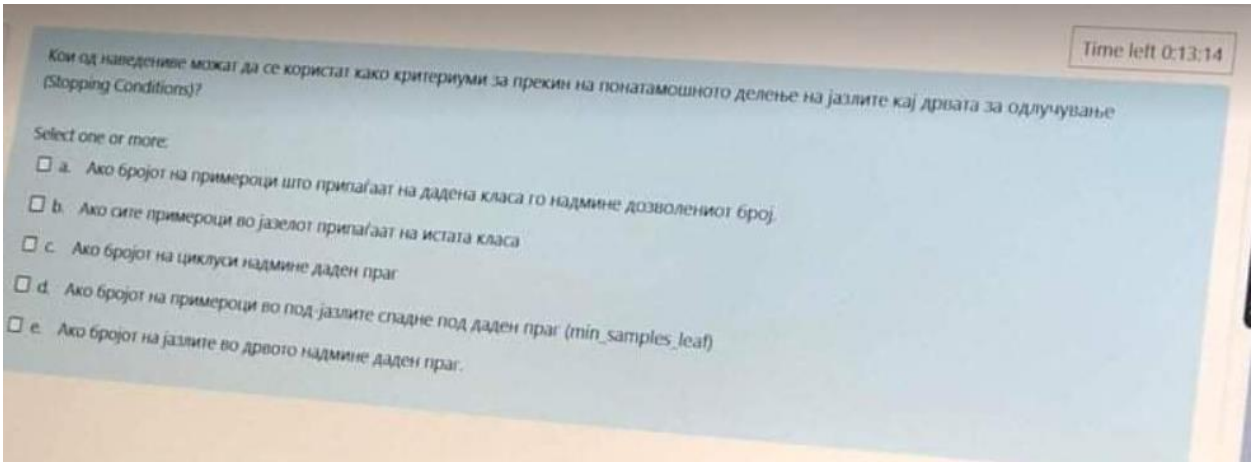
Одговор: полу-надлегувано, само-надлегувано

9. Кои мерки може да ги користиме за сличност помегу два кластера?



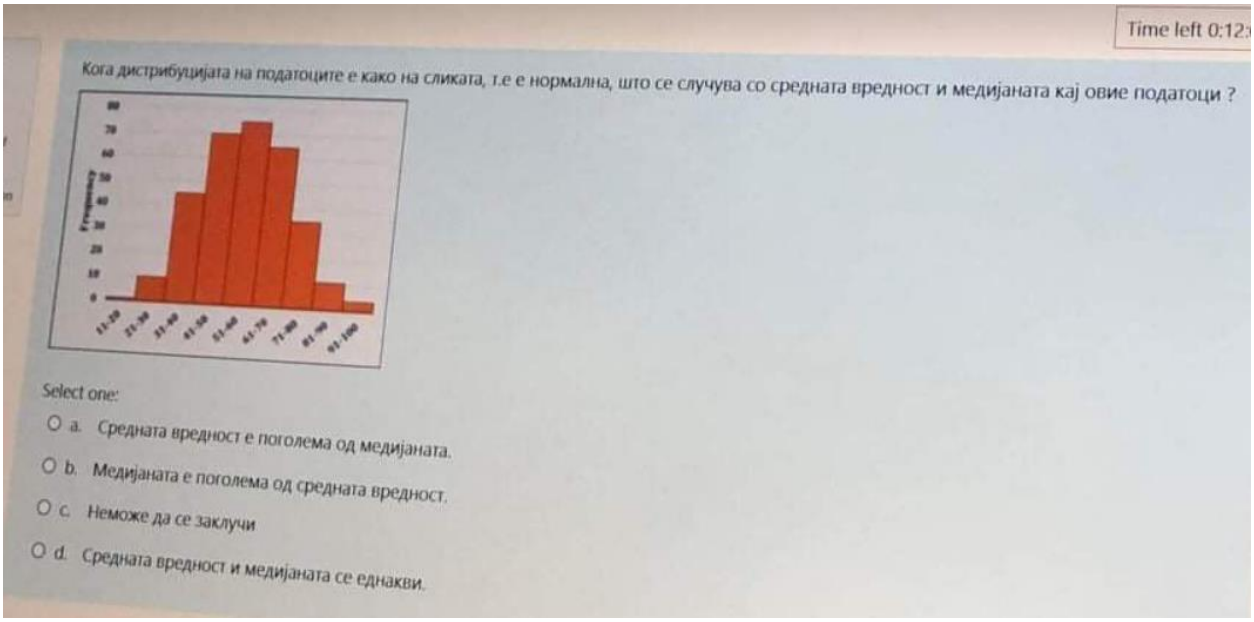
Одговор: Најмала различност помегу два елементи од кластерите, сличност помегу центроидите на двата кластера

10. Кои од наведените може да продолжат да се користат како критериум за понатамошно делење на јазлите кај дрвата за откачување?



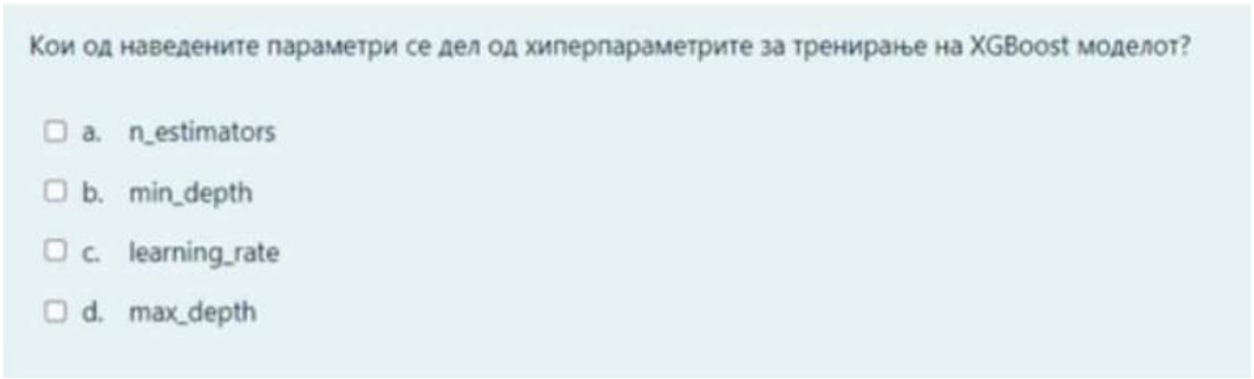
Одговор:

11. Кога дистрибуцијата на податоци е како на сликата, т.е. нормална што се случува со средната вредност и медијаната на овие податоци?



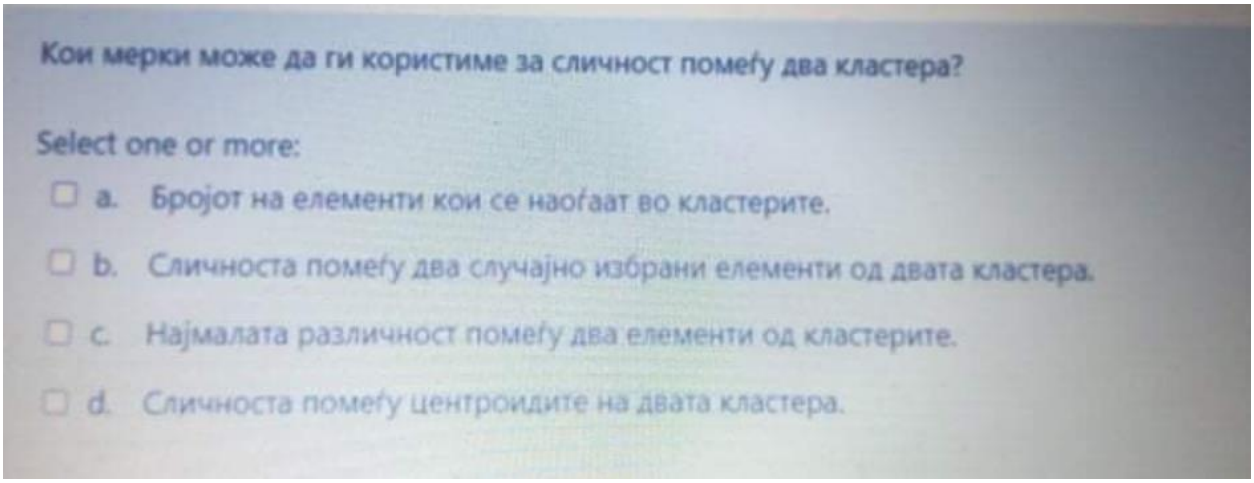
Одговор: Медијаната е поголема од средната вредност

12. Кои од наведените параметри се дел од хиперпараметрите за тренирање К моделот?



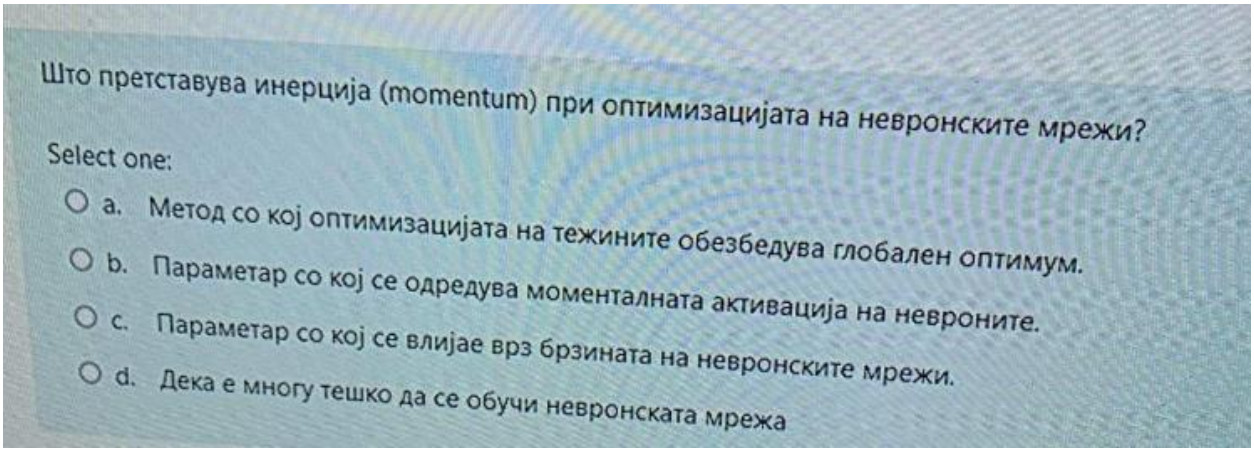
Одговор:

13. Кои мерки можеме да ги користиме за сличност мегу два кластера?



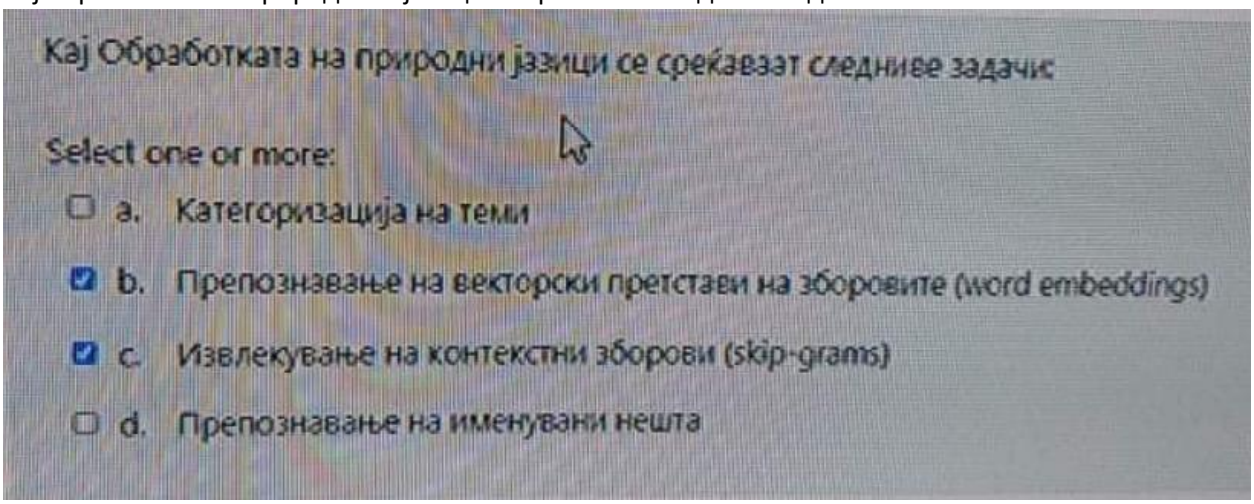
Одговор:

14. Што претставува инерција при оптимизација на невронски мрежи?



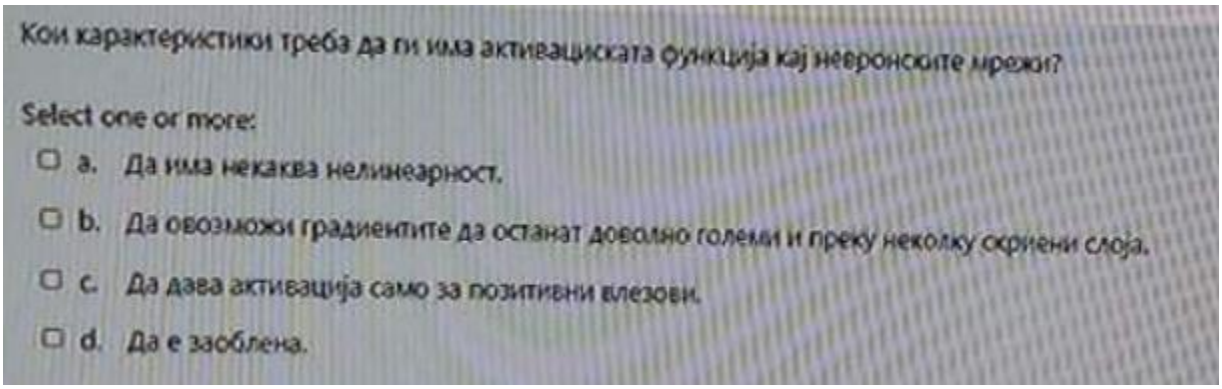
Одговор: Метод со кој оптимизацијата на тежините обезбедува глобален оптимум

15. Кај обработката на природните јазици се среќаваат следниве задачи:



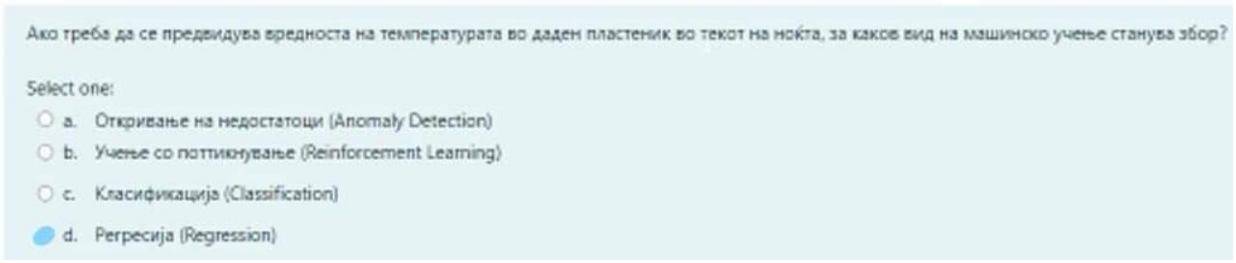
Одговор: двете дадени, плус препознавање на именувани нешта

16. Кои карактеристики треба да ги има активациската функција кај невронските мрежи?

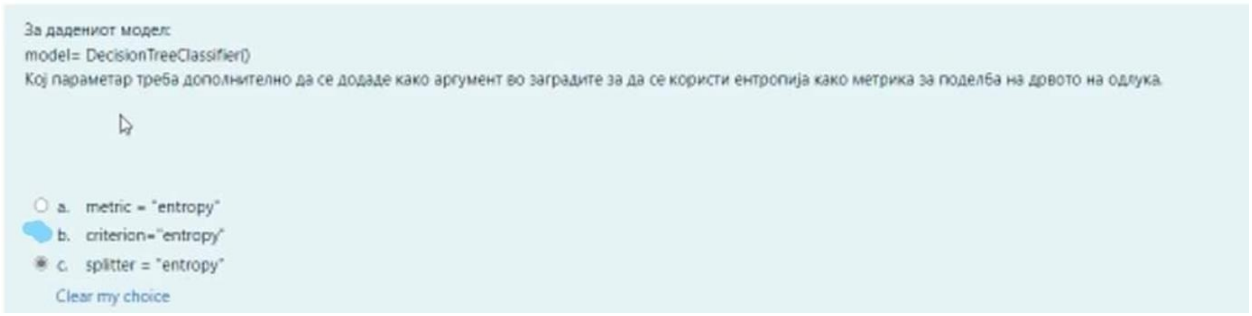


Одговор: Да има некаква нелинеарност, да овозможи градиентите да останат доволно големи и преку неколку скриени слоја

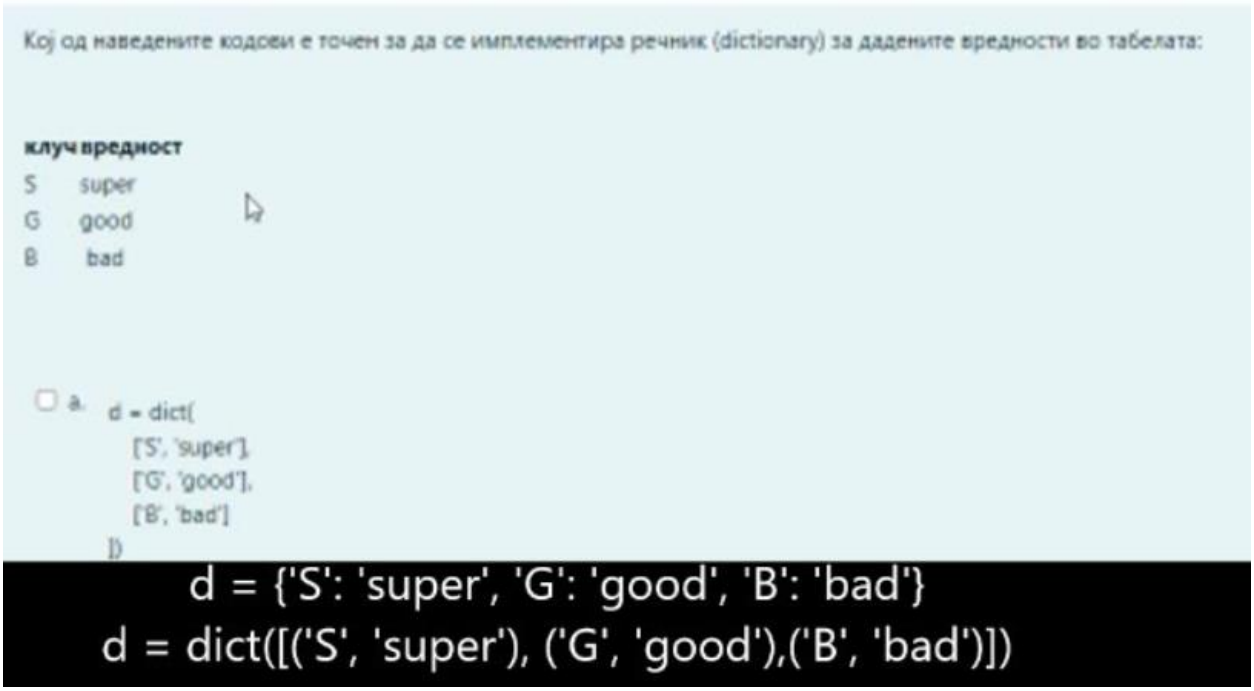
17. Ако треба да се предвидува вредноста на температурата во даден пластеник во текот на ноќта, за каков вид на машинско учење станува збор?
    Одговор: регресија



18. За дадениот модел

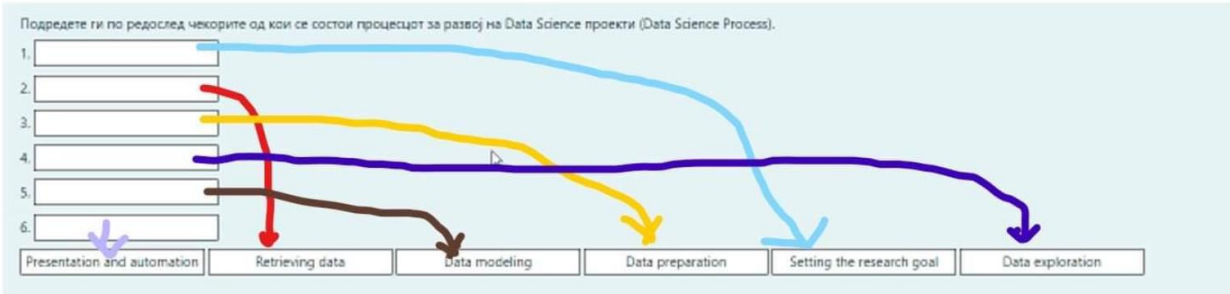


19. Кој од наведените кодови е точен за да се имплементира речник за дадените вредности во табелата

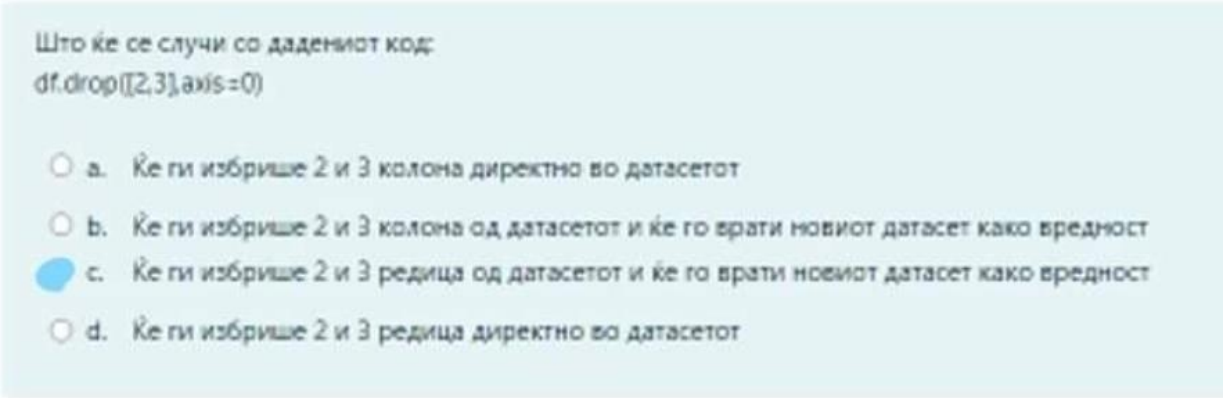


20. Со кој од дадените алгоритми за кластирање може да се добијат шест кластери како што се прикажани на сликата

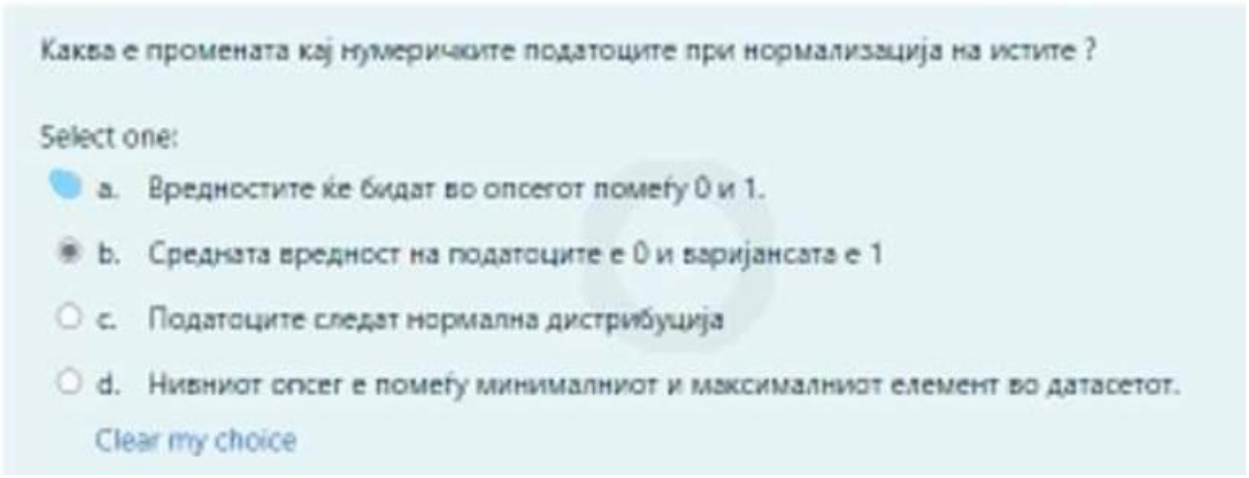
21. Подредете ги по ред чекорите по редослед од кои се состои процесот за развој на data science проектите data science process



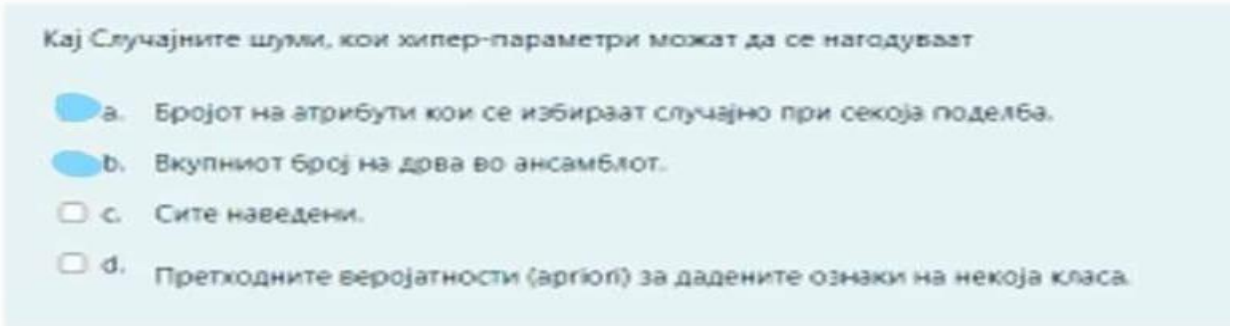
22. Што ќе се случи со дадениот код



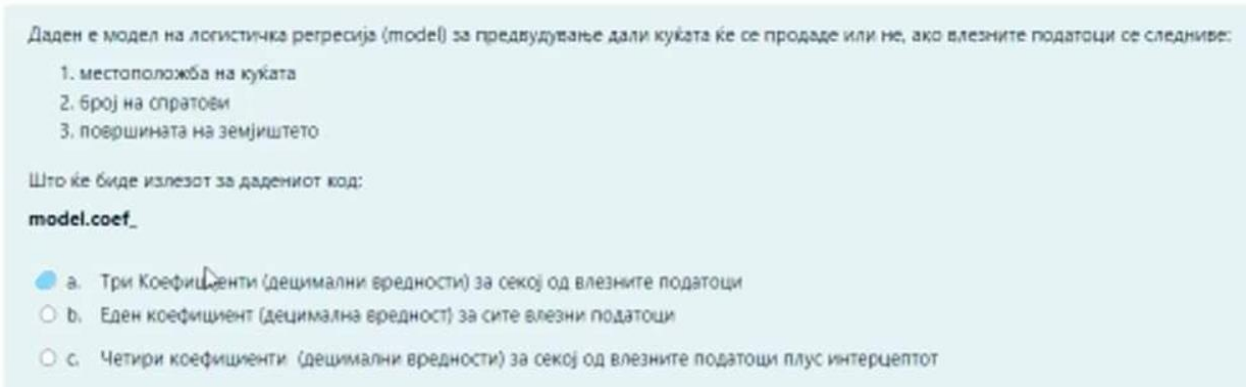
23. Каква е промената кај нумеричките податоци при нормализација на истите?



24. Кај случајните шуми кои хипер параметри може да се нагодуваат



25. Даден е модел на логистичка регресија за предвидување дали куќата ке се продаде или не, ако влезните податоци се следниве



26. За дадениот код која визуелизација ќе се прикаже

За дадениот код која визуелизација ќе се прикаже?
df.hist(bins = 5)

- температура (децимални вредности)
- влажност на воздухот (децимални вредности)
- дали врнело во текот на денот(категориска вредност -> Yes / No)

a. Хистограм за секоја од колоните
b. Хистограм на целиот датасет
c. Хистограм за секоја од нумеричките колони
d. Ниту едно од наведените

27. Кои од визуелизациите е најдобро да се изберат кога станува збор за датасет од категориски податоци?

Кои од визуелизациите е најдобро да се изберат кога станува збор за датасет од категориски податоци ?

Select one or more:

- [ ] a. Dot plot
- [ ] b. Scatter plot
- [x] c. Histogram
- [x] d. Bar charts

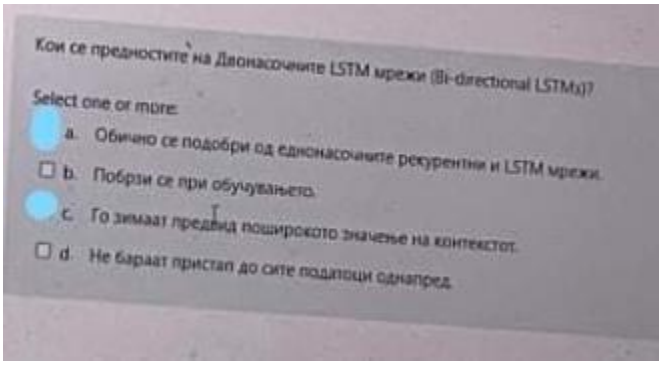
28. Што резултат враќа дадениот код

Што резултат враќа дадениот код:

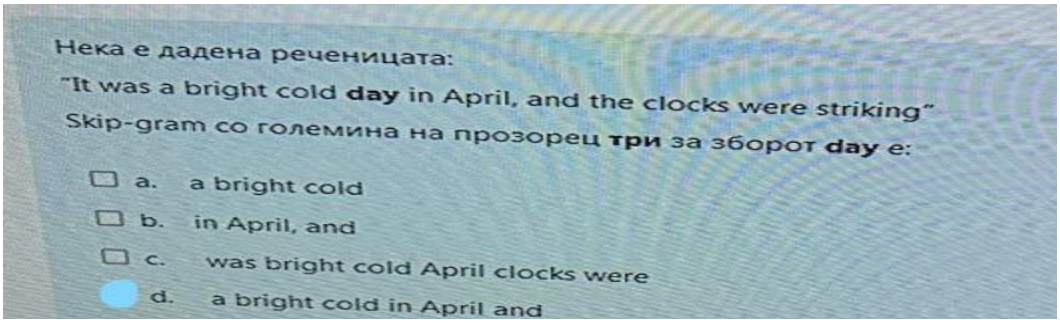
df.isnull()

a. Целата табела (df) со True/False вредности во зависност дали на дадената позиција има/нема NAN вредност
b. Целата табела (df) само со позициите каде има NAN вредност
c. Целата табела (df) само со позициите каде нема NAN вредност
d. Број на NAN вредности по колона

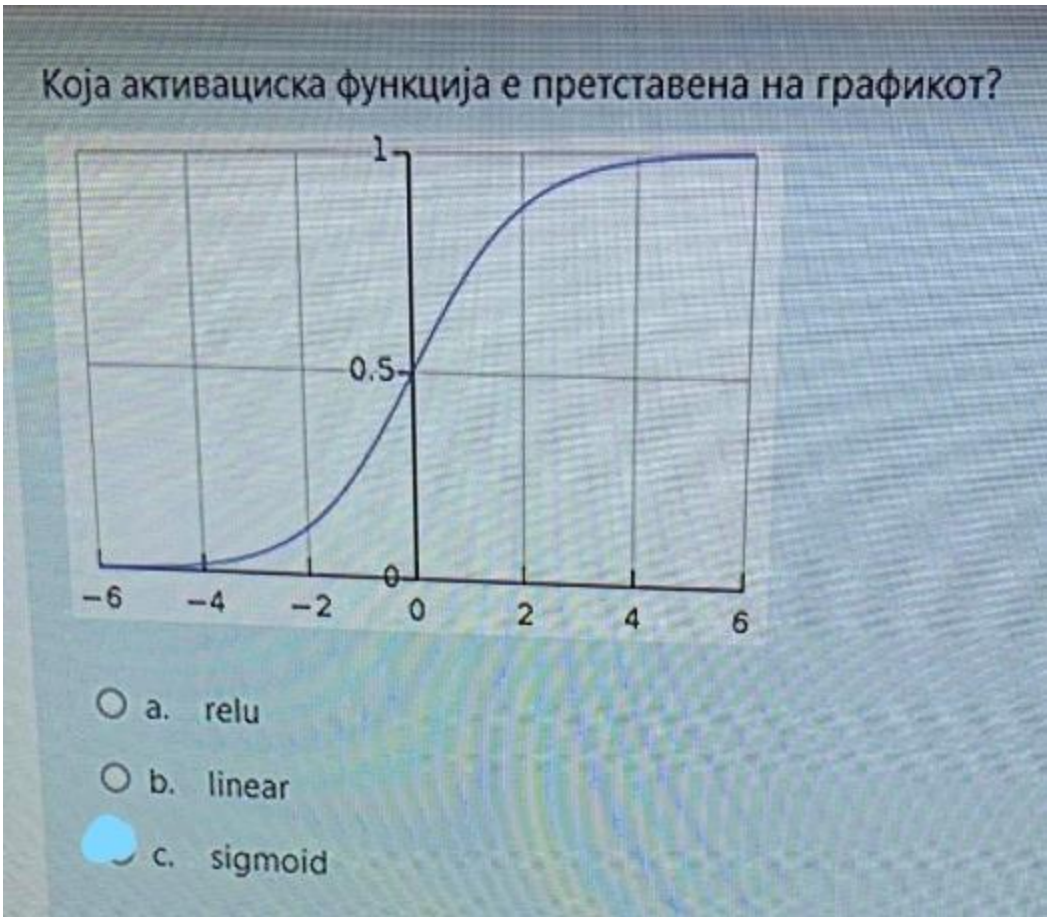
29. Кои се предностите на двонасочните lstm мрежи



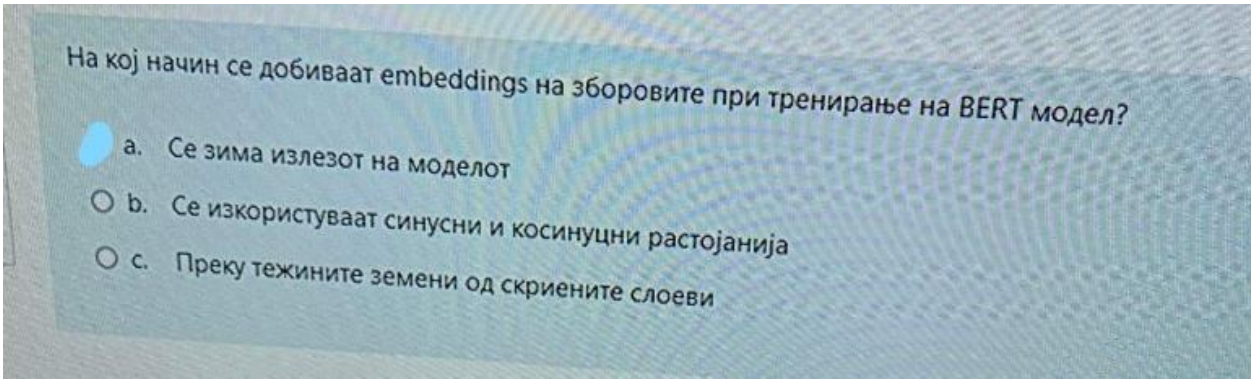
30. Нека е дадена реченицата:



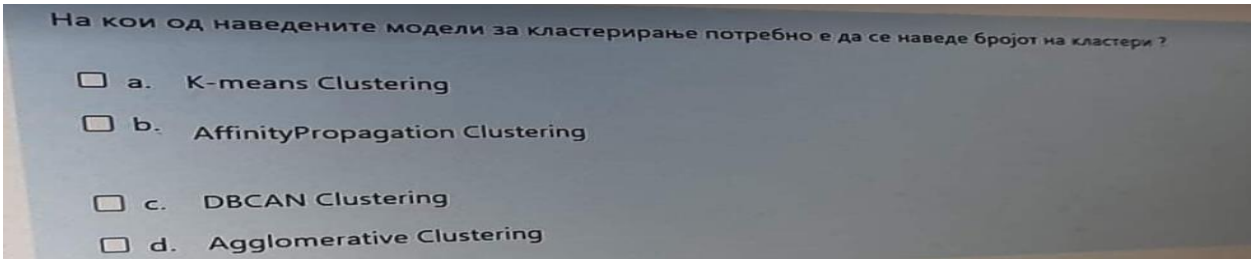
31. Која активациска функција е претставена на графикот



32. На кој начин се добиваат embbedings на зборовите при тренирање на bert модел

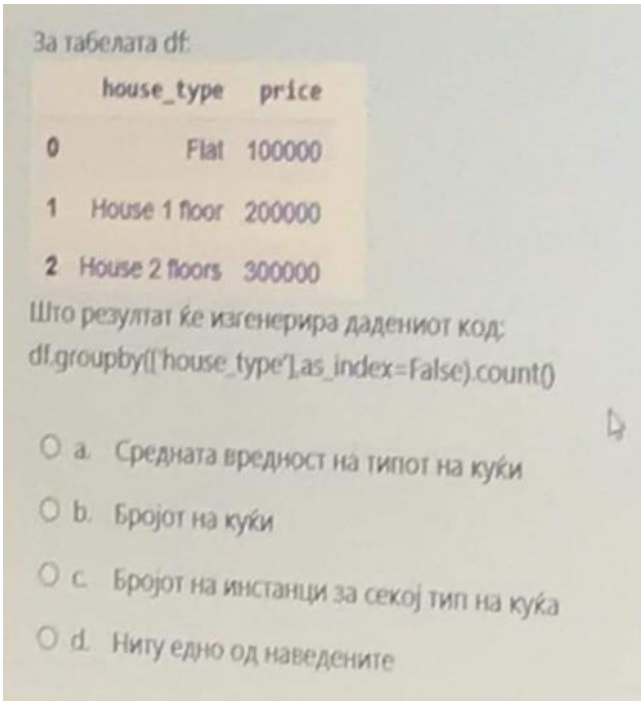


33. На кои од наведените модели за кластрирање потребно е да се наведе бројот на кластери?



Одговор: k-means claustering, agglomerative clausstering

34. За табелата


За табелата df:

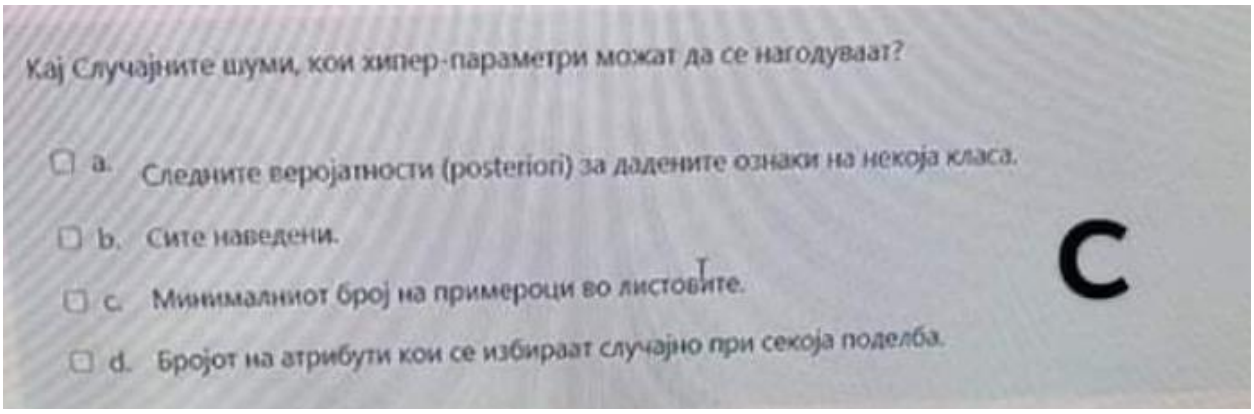
	house_type	price
0	Flat	100000
1	House 1 floor	200000
2	House 2 floors	300000

Што резултат ќе изгенерира дадениот код:
df.groupby(['house_type'],as_index=False).count()

- a. Средната вредност на типот на куќи
- b. Бројот на куќи
- c. Бројот на инстанци за секој тип на куќа
- d. Ниту едно од наведените

Одговор: Бројот на инстанци за секој тип на кука

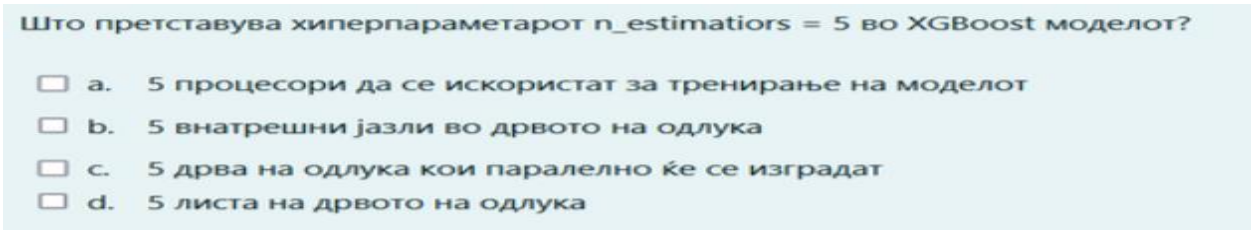
35. Кај случајните шуми, кои хипер-параметри може да се нагодуваат?


Кај Случајните шуми, кои хипер-параметри можат да се нагодуваат?

- a. Следните веројатности (posteriori) за дадените ознаки на некоја класа.
- b. Сите наведени.
- c. Минималниот број на примероци во листовите.
- d. Бројот на атрибути кои се избираат случајно при секоја поделба.

C

36. Што претставува хиперпараметарот n_estimatiors = 5 во xgboost моделот?


Што претставува хиперпараметарот n_estimatiors = 5 во XGBoost моделот?

- a. 5 процесори да се искористат за тренирање на моделот
- b. 5 внатрешни јазли во дрвото на одлука
- c. 5 дрва на одлука кои паралелно ќе се изградат
- d. 5 листа на дрвото на одлука

Одговор: 5 дрва на одлука кои пararелно ке се изградат

37. Во кој случај би било најдобро да се употреби softmax како излезно ниво кај невронските мрежи?

Во кој случај би било најдобро да се употреби Softmax како излезно ниво кај невронските мрежи

- [x] a. Кога имаме класификација во повеќе од две класи
- [ ] b. Кога сакаме да добиеме по брзо процесирање на резултатите на GPU
- [ ] c. Кога како мрека за пресметка на загуба во мрежата се користи MSE (Mean Squared Error)
- [x] d. Кога бројот на влезови е поголем од бројот на излези во невронската мрежа
- [ ] e. Кога имаме длабока невронска мрежа

38. Еден од најдобрите јазични модели gpt-2 се потпира на трансформер архитектура.

Еден од најдобрите јазични модели GPT-2 се потпира на трансформер архитектурата. Кој дел од трансформер архитектурата се користи во GPT-2?

- [ ] a. Decoder+Encoder
- [x] b. Првите 9 нивоа од Decoder делот
- [x] c. Decoder
- [ ] d. Encoder

Одговор:

39.