

Вовед во науки за податоци (база)

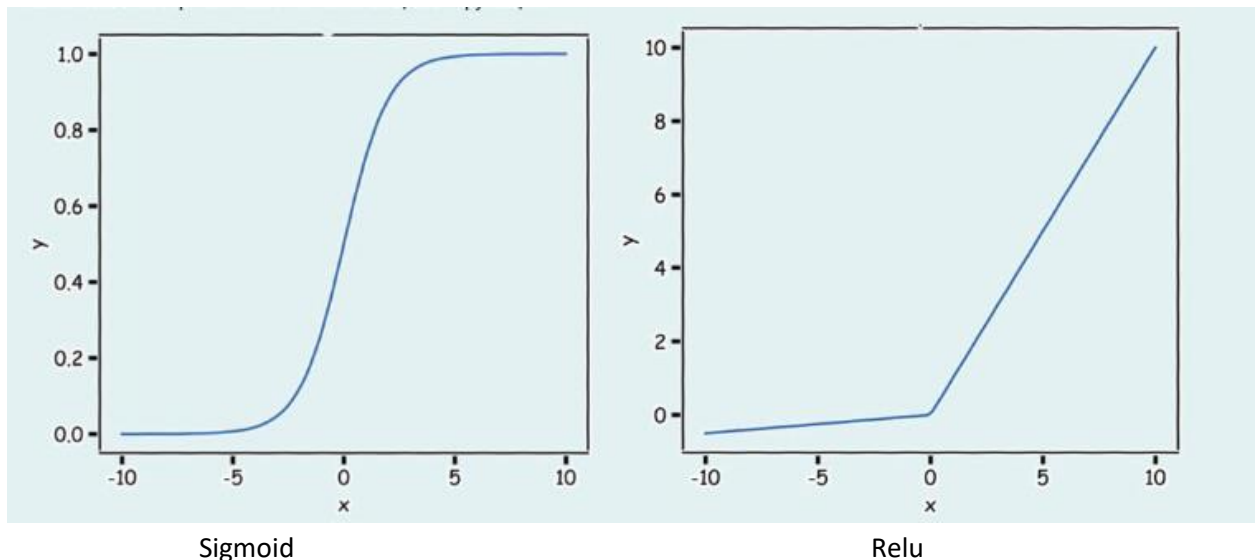
1. За дадената табела во прилог ако се енкодира колоната **Class** со помош на Binary Encoding, како ќе изгледа ново добиената табела?

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

○ a.

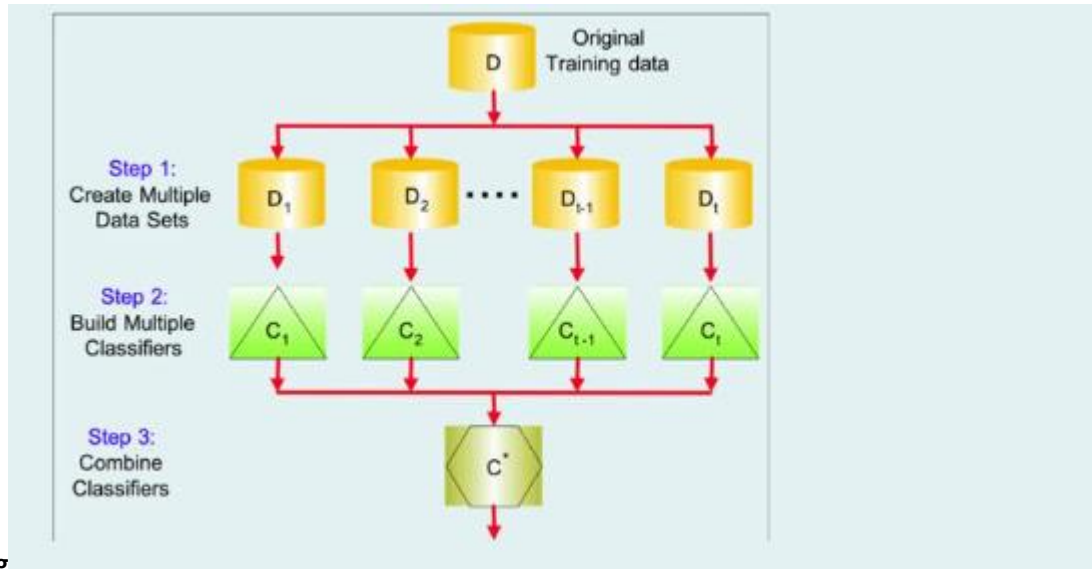
Id	Company_name	Class_1	Class_2	Class_3
1	Pepsi	1	0	0
2	Zara	0	1	0
3	Coca - Cola	1	0	0
4	Apple	1	1	0
5	Mcdonald's	1	1	1

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 ?



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 transformers 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 избераат 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 поттикнување (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 се замени **Null** за променливата 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

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 (колонтата 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](#)

15.

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

Даден е модел на KNN класификација за предвидување дали куќата ќе се продаде или не - 0 или 1 соодветно (class колоната), ако како влезни променливи се следниве:

1. местоположба на куќата
2. број на спратови
3. површината на земјиштето

Што дефинира `n_neighbors=2` за дадениот код:
`classifier = KNeighborsClassifier(n_neighbors=2)`

- ☐ a. Само за последните два примероци од dataset-от ќе се пресмета растојанието до примерокот чија класа ја предвидуваме
- ☒ b. За предвидување на class колоната на нов примерок ќе бидат земени двата најблиски примероци на кои е трениран моделот
- ☐ c. Само за првите два примероци од dataset-от ќе се пресмета растојанието до примерокот чија класа ја предвидуваме
- ☐ d. Само за соседните два примероци на примерокот чија класа ја предвидуваме ќе се пресмета растојанието

[Clear my choice](#)

18.

19. За дадениот модел : `model = DecisionTreeClassifier()`

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 особените 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

Што е точно за моделот seq2seq?

Select one or more:

- ☐ a. При тестирањето се генерираат збор по збор, сè додека не се добие на излез знак за крај на реченицата.
- ☐ b. Крајниот скриен слој на енкодерскиот дел е влезен слој за декодерскиот дел.
- ☐ c. Обуката се одвива како и кај другите Рекурентни невронски мрежи.
- ☐ d. Предноста на seq2seq е што целото значење на реченицата е претставено во крајниот скриен слој на енкодерскиот дел.

Одговор:

4. Што е skip grams?

Што се Skip-grams?

- ☐ a. N-grams кои се појавуваат во дадена реченица но не се појавуваат во дадениот контекст.
- ☐ b. Множество од не-последователни зборови (со одредено поместување), кои се појавуваат во некоја реченица.
- ☐ c. Стоп-зборовите кои се појавуваат најчесто
- ☐ d. Множество од сите зборови во реченицата

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

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

Што претставува поимот отфрлање (dropout) во контекст на невронски мрежи?

Select one:

- ☐ a. Трајно бришење од меморијата.
- ☐ b. Бришење од меморијата при тестирање.
- ☐ c. Случајно поставување на активацијата и тежините на врските на некои неврони на нула.
- ☐ d. Откривање на недостатоци и нивно отфрлање.

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?

Што претставува поимот отфрлање (dropout) во контекст на невронски мрежи?

Select one:

- ☐ a. Трајно бришење од меморијата.
- ☐ b. Бришење од меморијата при тестирање.
- ☐ c. Случајно поставување на активацијата и тежините на врските на некои неврони на нула.
- ☐ d. Откривање на недостатоци и нивно отфрлање.

Одговор: Можност за учење на долги низи

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

Кои од наведените параметри се дел од хиперпараметрите за тренирање на XGBoost моделот?

- ☐ a. min_depth
- ☐ b. n_estimators
- ☐ c. max_depth
- ☐ d. learning_rate

Одговор: n_estimators, learning_Rate, max_Depth

8. Каков вид на учење се реализира кај Автоенкодерите?

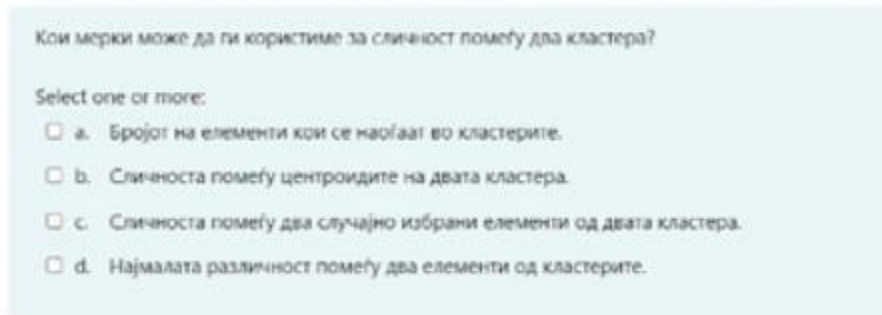
Каков вид на учење се реализира кај Автоенкодерите?

Select one or more:

- ☐ a. со поттикнување (reinforcement)
- ☐ b. надгледувано (supervised)
- ☐ c. полу-надгледувано (semi-supervised)
- ☐ d. само-надгледувано (self-supervised)

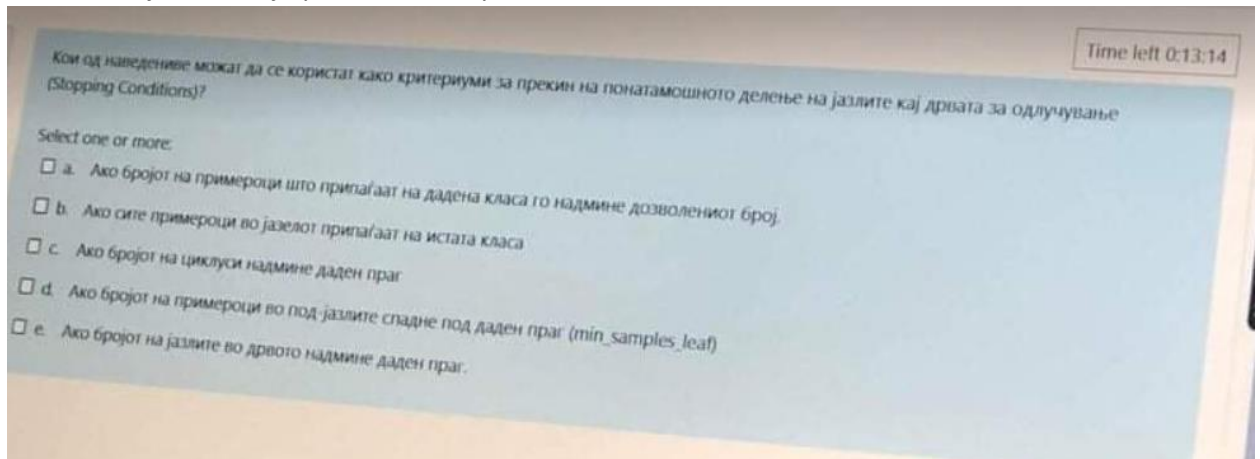
Одговор: полу-надгледувано, само-надгледувано

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



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

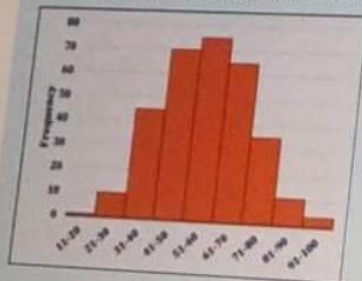
10. Кои од наведените може да продолжат да се користат како критериум за понатамошно делење на јазлите кај дрвата за откачување?



Одговор:

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

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



Select one:

- ☐ a. Средната вредност е поголема од медијаната.
- ☐ b. Медијаната е поголема од средната вредност.
- ☐ c. Неможе да се заклучи
- ☐ d. Средната вредност и медијаната се еднакви.

Одговор: Медијаната е поголема од средната вредност

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

Кои од наведените параметри се дел од хиперпараметрите за тренирање на XGBoost моделот?

- ☐ a. `n_estimators`
- ☐ b. `min_depth`
- ☐ c. `learning_rate`
- ☐ d. `max_depth`

Одговор:

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

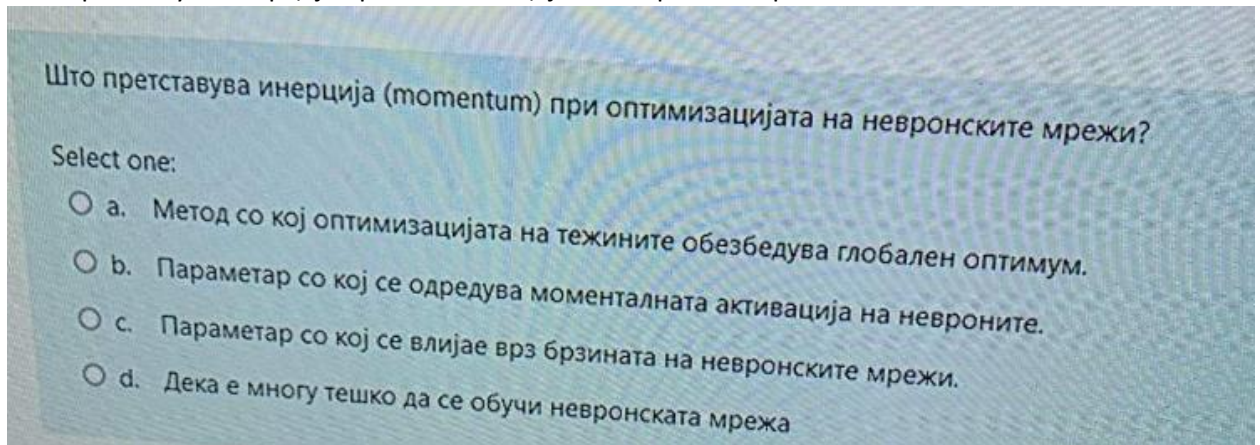
Кои мерки може да ги користиме за сличност помеѓу два кластера?

Select one or more:

- ☐ a. Бројот на елементи кои се наоѓаат во кластерите.
- ☐ b. Сличноста помеѓу два случајно избрани елементи од двата кластера.
- ☐ c. Најмалата различност помеѓу два елементи од кластерите.
- ☐ d. Сличноста помеѓу центроидите на двата кластера.

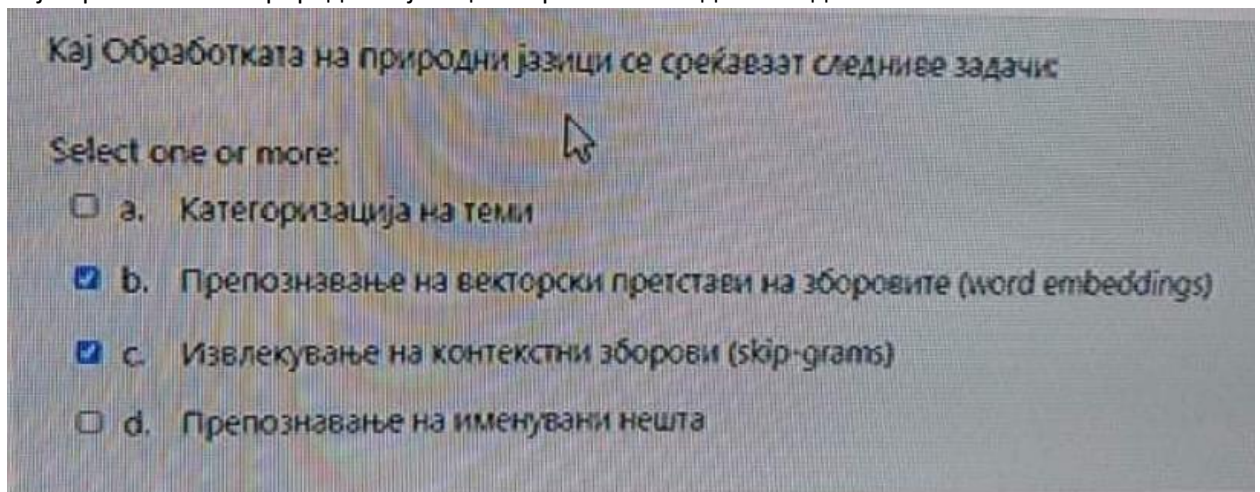
Одговор:

14. Што претставува инерција при оптимизација на невронски мрежи?



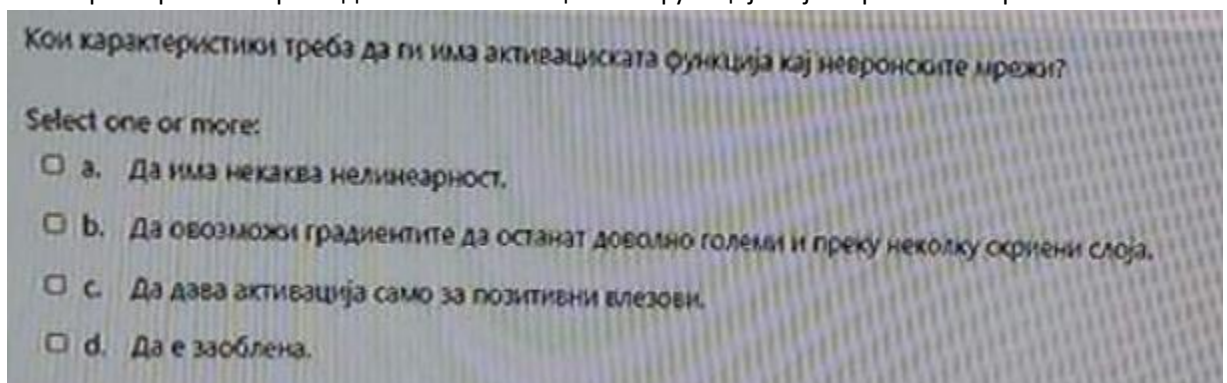
Одговор: Метод со кој оптимизацијата на тежините обезбедува глобален оптимум

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

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

Select one:

- ☐ a. Откривање на недостатоци (Anomaly Detection)
- ☐ b. Учење со поттикнување (Reinforcement Learning)
- ☐ c. Класификација (Classification)
- ☒ d. Регресија (Regression)

18. За дадениот модел

За дадениот модел:
model= DecisionTreeClassifier()
Кој параметар треба дополнително да се додаде како аргумент во заградите за да се користи ентропија како метрика за поделба на дрвото на одлука.

☐ a. metric = "entropy"

☒ b. criterion="entropy"

☐ c. splitter = "entropy"

[Clear my choice](#)

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

Кој од наведените кодови е точен за да се имплементира речник (dictionary) за дадените вредности во табелата:

ключ	вредност
S	super
G	good
B	bad

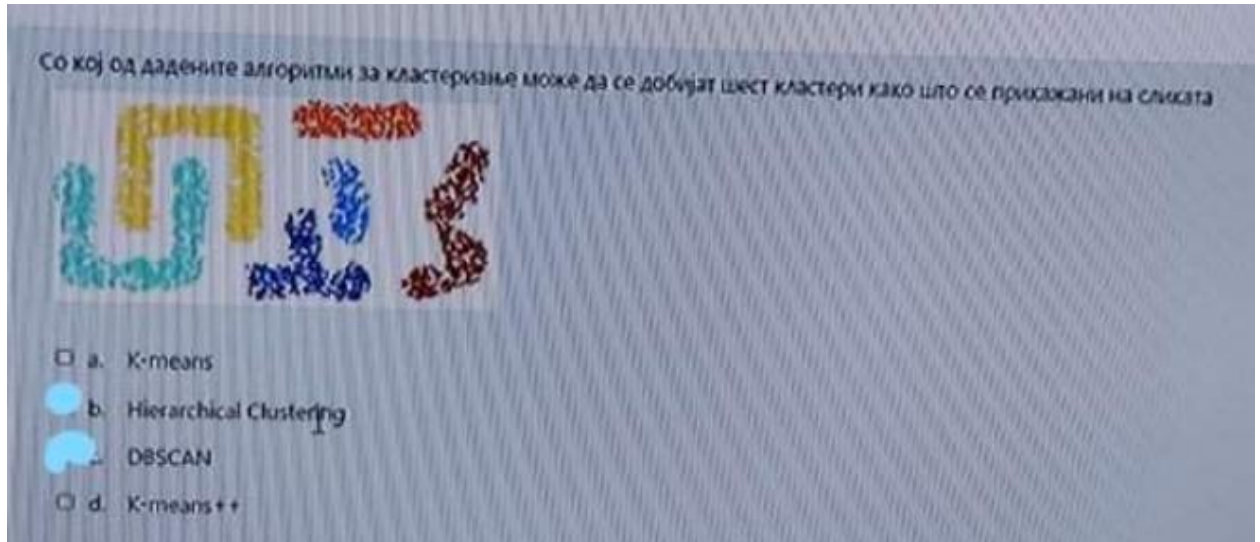
☐ a.

```
d = dict([  
    ['S', 'super'],  
    ['G', 'good'],  
    ['B', 'bad']  
])
```

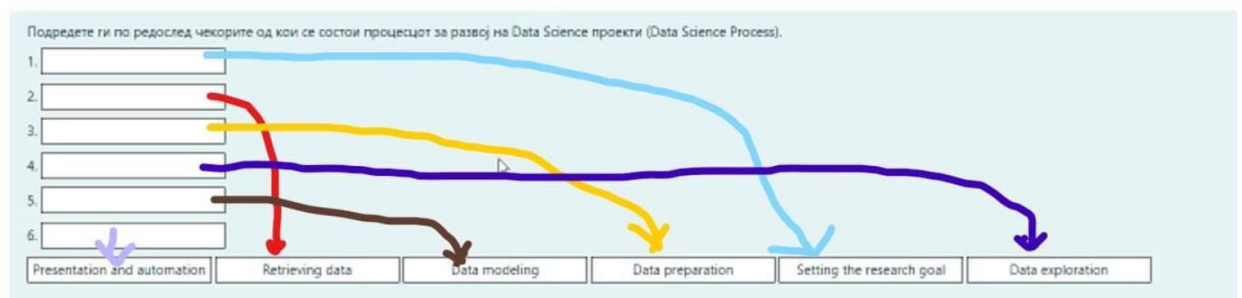
**```
d = {'S': 'super', 'G': 'good', 'B': 'bad'}
d = dict([('S', 'super'), ('G', 'good'), ('B', 'bad')])
```**



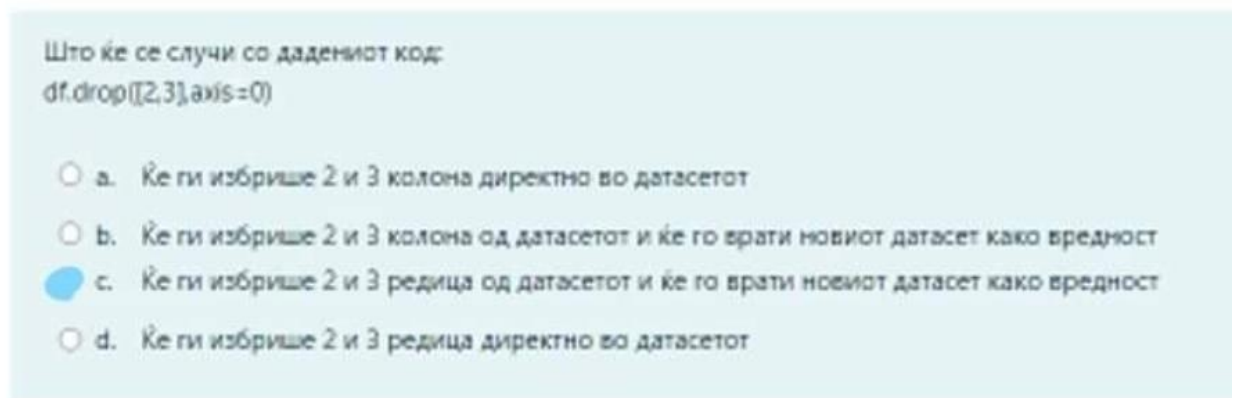
20. Со кој од дадените алгоритми за кластирање може да се добијат шест кластери како што се прикажани на сликата



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?

Каква е промената кај нумеричките податоците при нормализација на истите ?

Select one:

- ☒ a. Вредностите ќе бидат во опсегот помеѓу 0 и 1.
- ☐ b. Средната вредност на податоците е 0 и варијансата е 1
- ☐ c. Податоците следат нормална дистрибуција
- ☐ d. Нивниот опсег е помеѓу минималниот и максималниот елемент во датасетот.

[Clear my choice](#)

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

Кај Случајните шуми, кои хипер-параметри можат да се нагудуваат

- ☒ a. Бројот на атрибути кои се избираат случајно при секоја поделба.
- ☒ b. Вкупниот број на дрва во ансамблот.
- ☐ c. Сите наведени.
- ☐ d. Претходните веројатности (argprior) за дадените ознаки на некоја класа.

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

Даден е модел на логистичка регресија (model) за предвидување дали куќата ќе се продаде или не, ако влезните податоци се следниве:

1. местоположба на куќата
2. број на спратови
3. површината на земјиштето

Што ќе биде излезот за дадениот код:

`model.coef_`

- ☒ a. Три Коефициенти (децимални вредности) за секој од влезните податоци
- ☐ b. Еден коефициент (децимална вредност) за сите влезни податоци
- ☐ c. Четири коефициенти (децимални вредности) за секој од влезните податоци плус интерцептот

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
- ☒ c. Histogram
- ☒ d. Bar charts

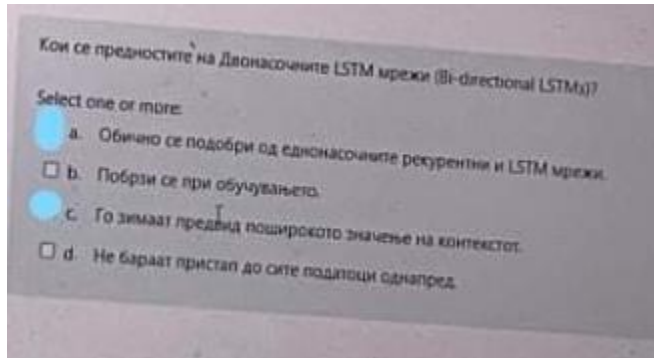
28. Што резултат враќа дадениот код

Што резултат враќа дадениот код:

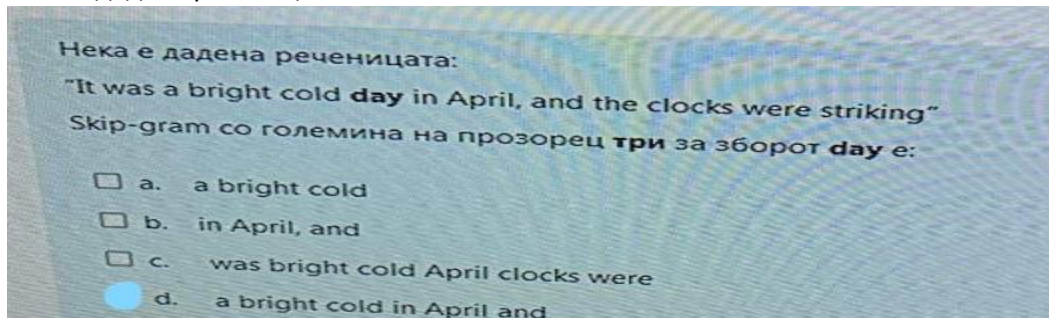
```
df.isnull()
```

- ☒ 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



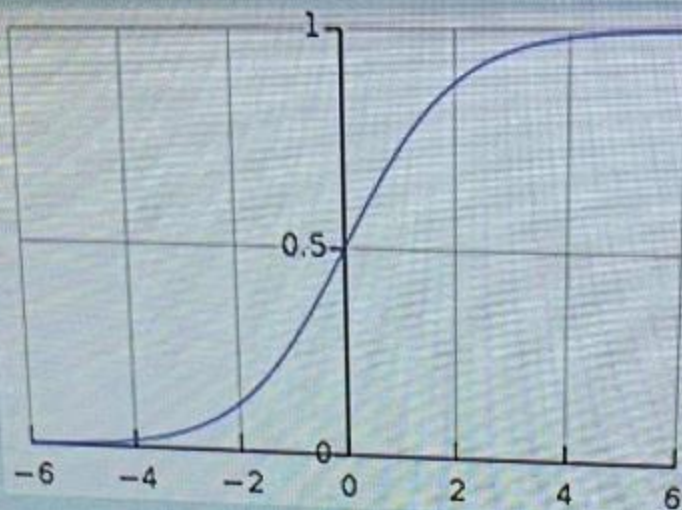
30. Нека е дадена реченицата:



31. Која активациска функција е претставена на графикот



Која активациска функција е претставена на графикот?



- ☐ a. relu
- ☐ b. linear
- ☒ c. sigmoid

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

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

- ☒ a. Се зима излезот на моделот
- ☐ b. Се изкористуваат синусни и косинусни растојанија
- ☐ c. Преку тежините земени од скриените слоеви

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

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

- ☐ a. K-means Clustering
- ☐ b. AffinityPropagation Clustering
- ☐ c. DBSCAN Clustering
- ☐ d. Agglomerative Clustering

Одговор: k-means clustering, agglomerative clustering



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 куќа

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

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

- ☐ a. Следните веројатности (posterior) 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.

36. Што претставува хиперпараметарот `n_estimators = 5` во `xgboost` моделот?

Што претставува хиперпараметарот `n_estimators = 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 паралелно ќе 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

- ☒ a. Кога имаме класификација во повеќе од две класи
- ☐ b. Кога сакаме да добиеме по брзо процесирање на резултатите на GPU
- ☐ c. Кога како мрежа за пресметка на загуба во мрежата се користи MSE (Mean Squared Error)
- ☒ 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
- ☒ b. Првите 9 нивоа од Decoder делот
- ☒ c. Decoder
- ☐ d. Encoder

Одговор:

39.