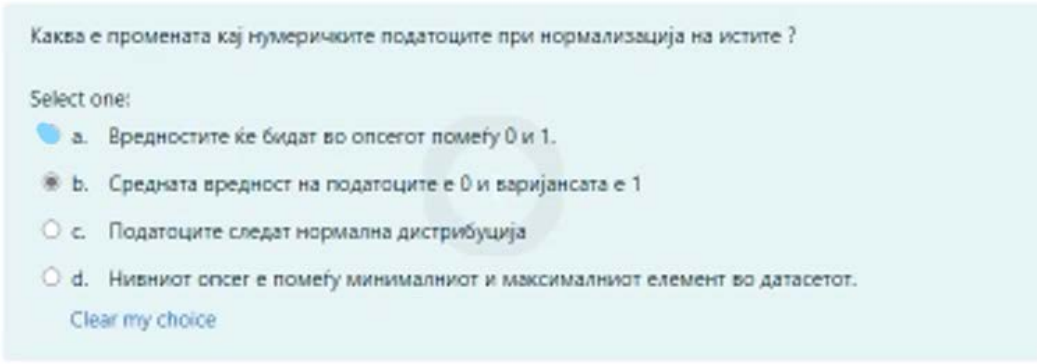
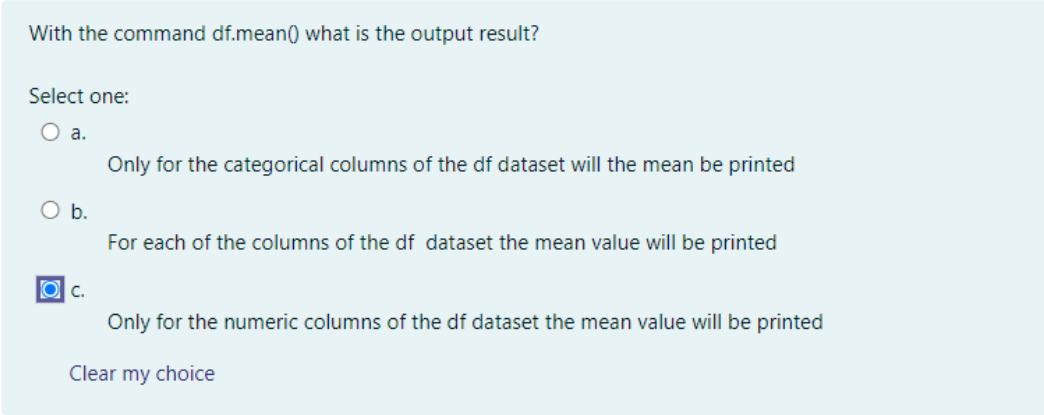
Вовед во науката за податоци

1. Каква е промената кај нумеричките податоци при нормализација на истите?
   - Вредностите ќе бидат во опсегот помеѓу 0 и 1.
   - Средната вредност на податоците е 0 и варијансата е 1
   - Податоците следат нормална дистрибуција
   - Нивниот опсег е помеѓу минималниот и максималниот елемент во датасетот.
   - Нормализација ги праве вредностите меѓу 0 и 1 а стандардизација, со средна вредност 0 а варијанса 1



2. With the command df.mean() what is the output result?
   - Only for the categorical columns of the df dataset will the mean be printed
   - For each of the columns od the df dataset the mean value will be printed.
   - Only for the numeric columns of the df dataset the mean value will be printed.



3. Which of the following descriptive statistics is best to choose if the dataset contains continuous data?
   - Frequency
   - Median value
   - Percent (row, column or total)

- Mean value

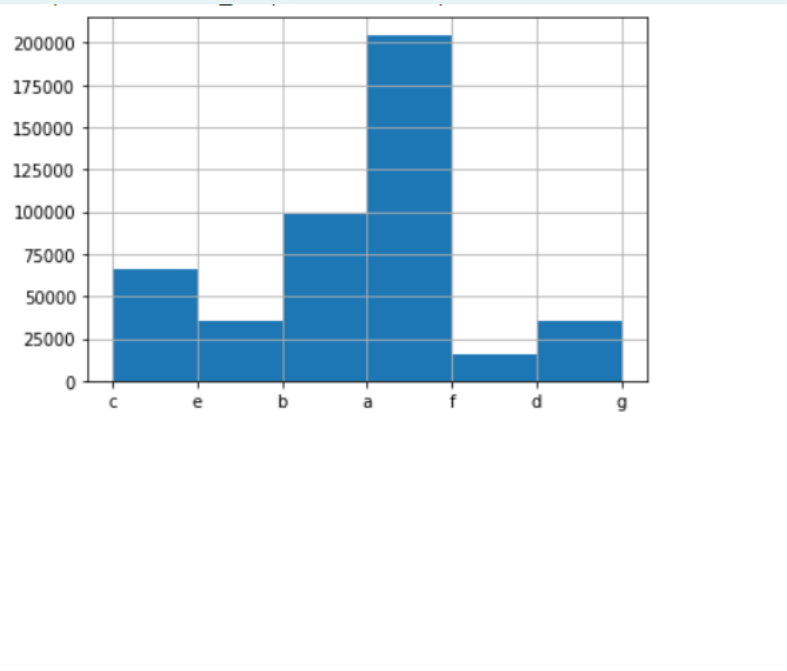
Which of the following descriptive statistics is best to choose if the dataset contains continuous data?

Select one or more:

- [ ] a. Frequency
- [x] b. Median value
- [ ] c. Percent (row, column or total)
- [x] d. Mean value

4. Which of the following commands is appropriate for the given visualization?
   - seaborn.displot(df[‘Bed Grade’], bins=3, kde=True, rug=True)
   - df[‘Hospital_type_code’].hist(bins = 3)
   - seaborn.distplot(df[‘Bed Grade’], bins=6, kde=True, rug=True)
   - df[‘Hospital_type_code’].hist(bins = 6)

Which of the following commands is appropriate for the given visualization?

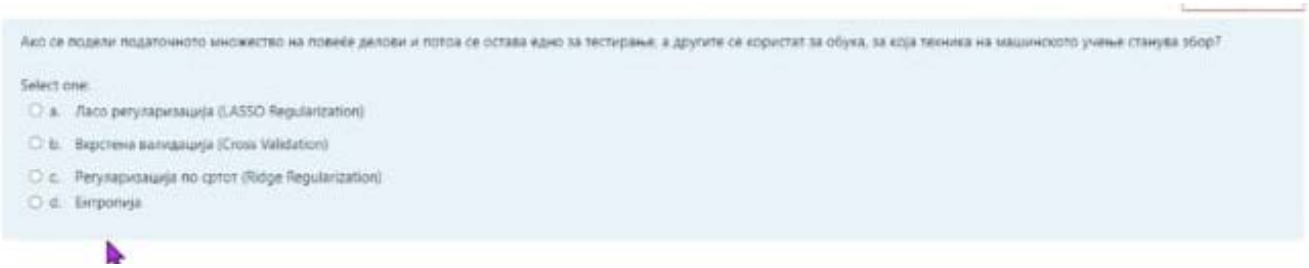


Select one:

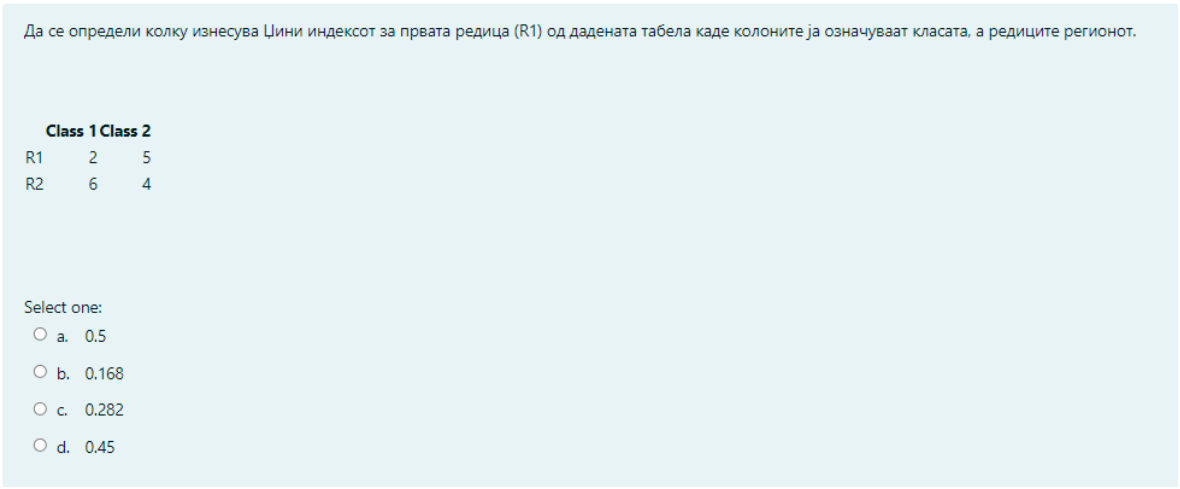
- ( ) a. seaborn.distplot(df['Bed Grade'], bins=3, kde=True, rug=True)
- ( ) b. df['Hospital_type_code'].hist(bins = 3)
- ( ) c. seaborn.distplot(df['Bed Grade'], bins=6, kde=True, rug=True)
- (•) d. df['Hospital_type_code'].hist(bins = 6)

Clear my choice

5. Ако се подели податочното множество на повеќе делови и потоа се остава едно за тестирање, а другите се користат за обука, за која техника на машинското учење станува збор.
    - Ласо регуларизација (LASSO Regularization)
    - Врстена валидација (Cross Validation) - ТОЧНО
    - Регуларизација по сртот (Ridge Regularization)
    - Ентропија



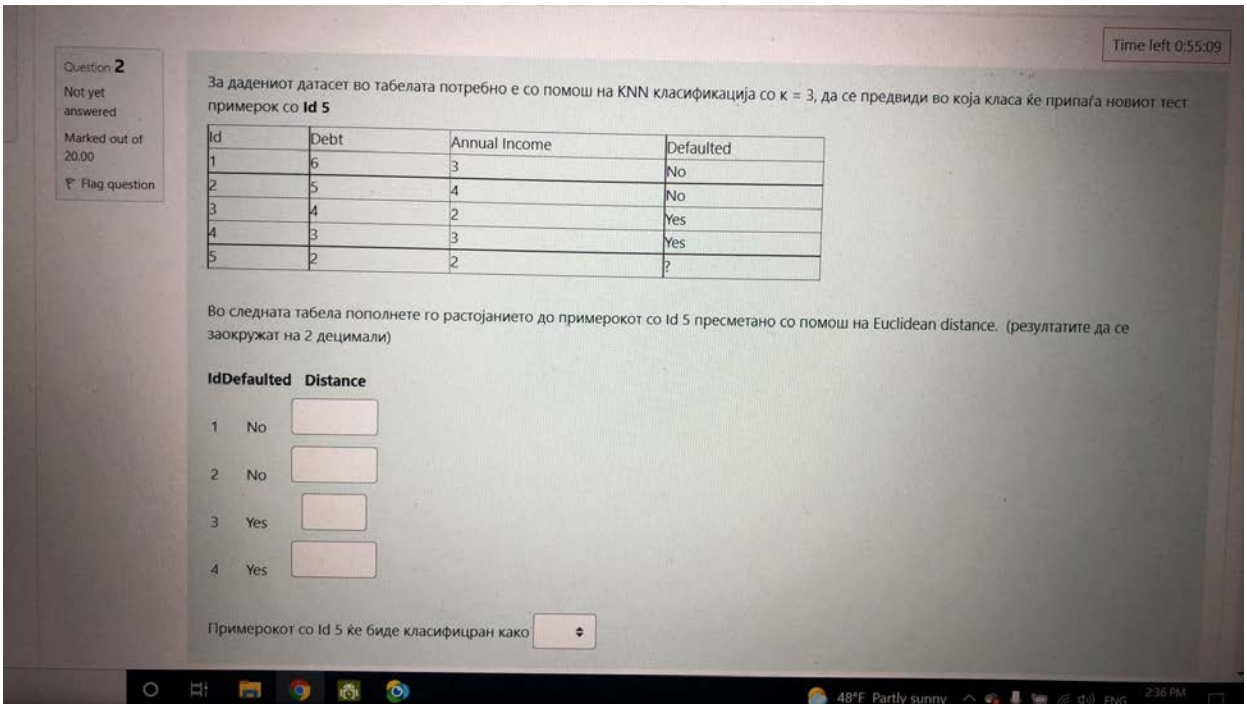
6. Да се определи колку изнесува Џини индексот за првата редица (R1) од дадената табела каде колоните ја означуваат класата, а редиците регионот.
    - 0.282
    - 0.45
    - 0.168
    - 0.5



7. With the command enc=OneHotEncoder(handle_unknown='ignore') _____, while the command enc.fit_transform(X).

With the command enc = OneHotEncoder(handle_unknown='ignore') Creates an instance from OneHotEncoder , while with the command enc.fit_transform(X) The OneHotEncoder model is trained and matrix is obtained from input column X .

8. За дадениот датасет во табелата потребно е со помош на KNN класификација со k=3, да се предвиди во која класа ќе припаѓа новиот тест примерок со ID 5.



Question 2
Not yet answered
Marked out of 20.00
Flag question

Time left 0:55:09

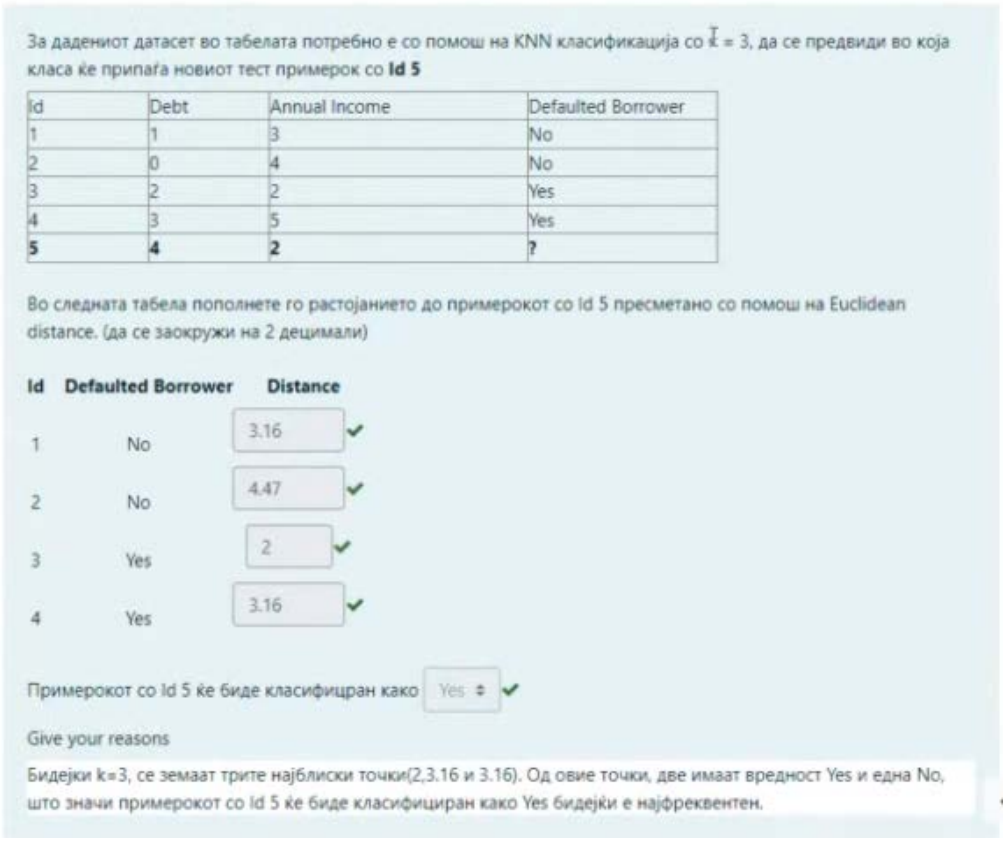
За дадениот датасет во табелата потребно е со помош на KNN класификација со k = 3, да се предвиди во која класа ќе припаѓа новиот тест примерок со **Id 5**

Id	Debt	Annual Income	Defaulted
1	6	3	No
2	5	4	No
3	4	2	Yes
4	3	3	Yes
5	2	2	?

Во следната табела пополнете го растојанието до примерокот со Id 5 пресметано со помош на Euclidean distance. (резултатите да се заокружат на 2 децимали)

Id	Defaulted	Distance
1	No	
2	No	
3	Yes	
4	Yes	

Примерокот со Id 5 ќе биде класифициран како



За дадениот датасет во табелата потребно е со помош на KNN класификација со $k = 3$, да се предвиди во која класа ќе припаѓа новиот тест примерок со **Id 5**

Id	Debt	Annual Income	Defaulted Borrower
1	1	3	No
2	0	4	No
3	2	2	Yes
4	3	5	Yes
5	**4**	**2**	**?**

Во следната табела пополнете го растојанието до примерокот со Id 5 пресметано со помош на Euclidean distance. (да се заокружи на 2 децимали)

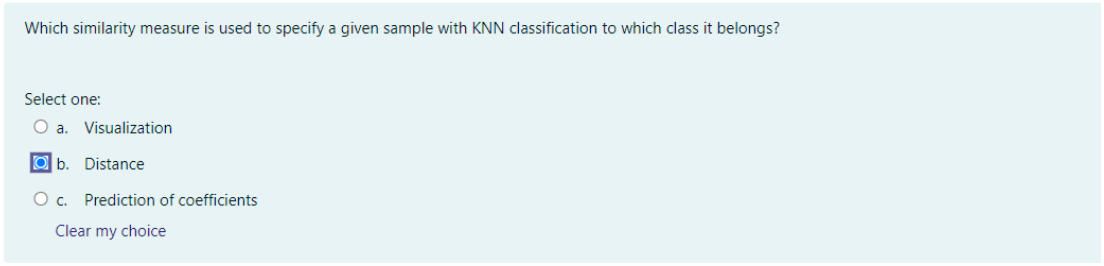
Id	Defaulted Borrower	Distance
1	No	3.16 ✔
2	No	4.47 ✔
3	Yes	2 ✔
4	Yes	3.16 ✔

Примерокот со Id 5 ќе биде класифициран како Yes ✔

Give your reasons

Бидејќи k=3, се земаат трите најблиски точки(2,3.16 и 3.16). Од овие точки, две имаат вредност Yes и една No, што значи примерокот со Id 5 ќе биде класифициран како Yes бидејќи е најфреквентен.

9. Which similarity measure is used to specify a given sample with KNN classification to which class it belongs?
    - Visualization
    - Distance
    - Prediction of coefficients



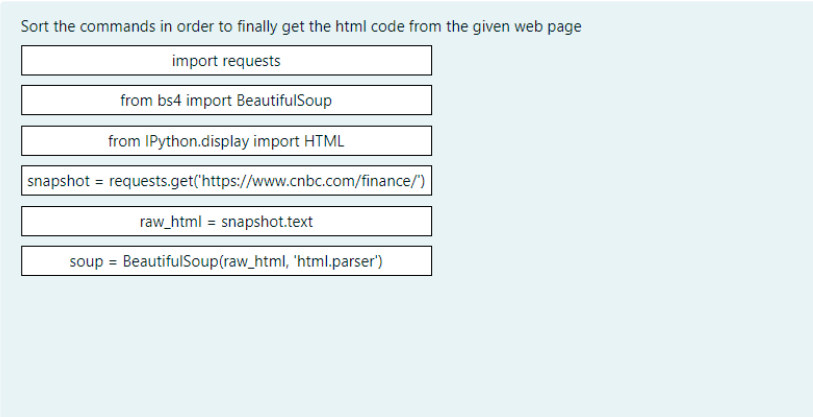
Which similarity measure is used to specify a given sample with KNN classification to which class it belongs?

Select one:

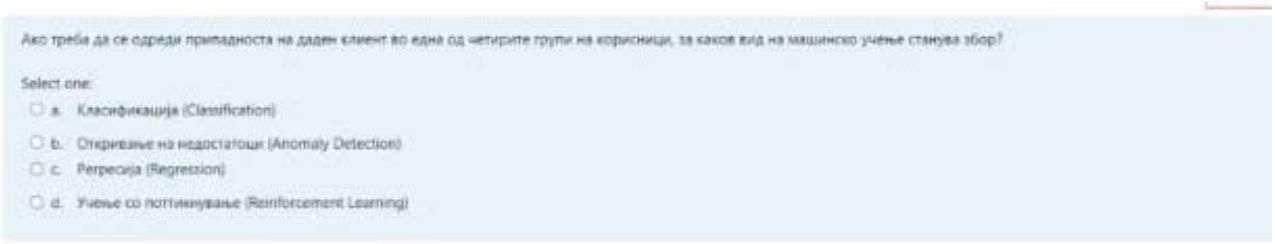
- a. Visualization
- b. Distance
- c. Prediction of coefficients

Clear my choice

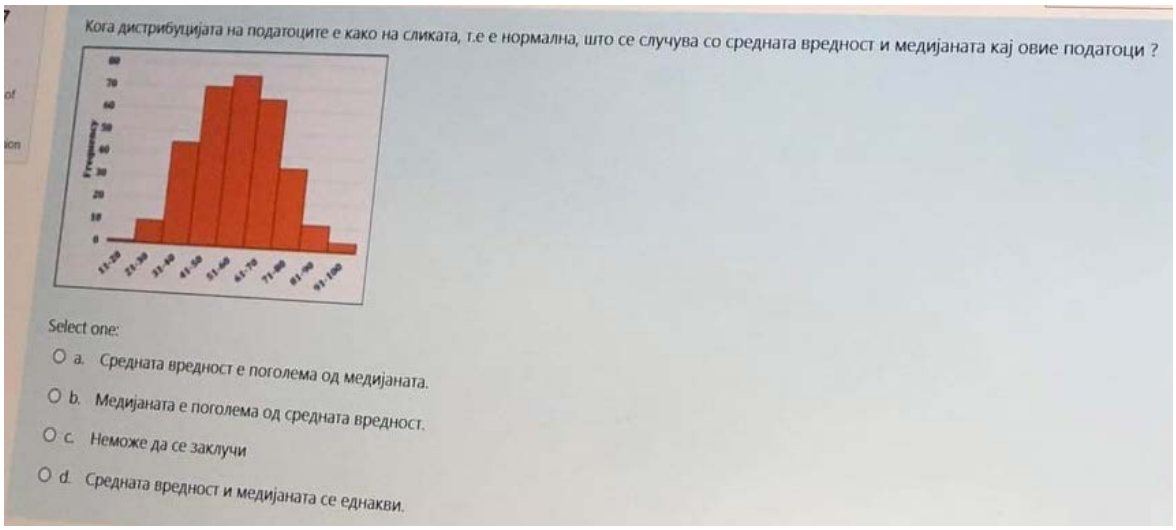
10. Sort the commands in order to finally get the html code from the given web page.



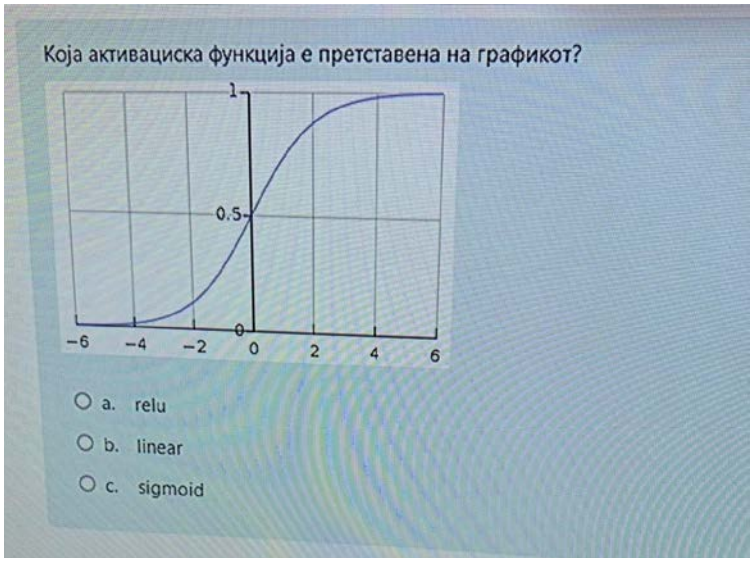
11. Ако треба да се одреди припадноста на даден клиент во една од четирите групи на корисници, за каков вид на машинско учење станува збор?
    - Класификација (Classification) - ТОЧНО
    - Откривање на недостатоците (Anomaly Detection)
    - Регресија (Regression)
    - Учење со поттикнување (Reinforcement Learning)



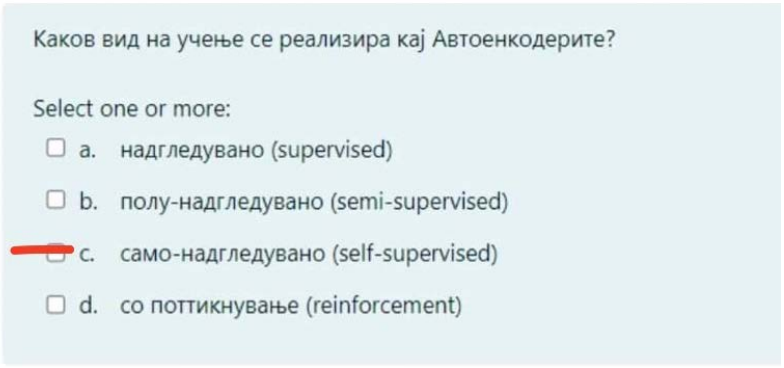
12. Кога дистрибуцијата на податоците е како на сликата, т.е. е наклонета на десно, што се случува со средната вредност и медијаната кај овие податоци?
    - Средната вредност е поголема од медијаната
    - Медијаната е поголема од средната вредност
    - Средната вредност и медијаната се еднакви – ВОА Е ТОЧНО
    - Не може да се заклучи од дадениот график



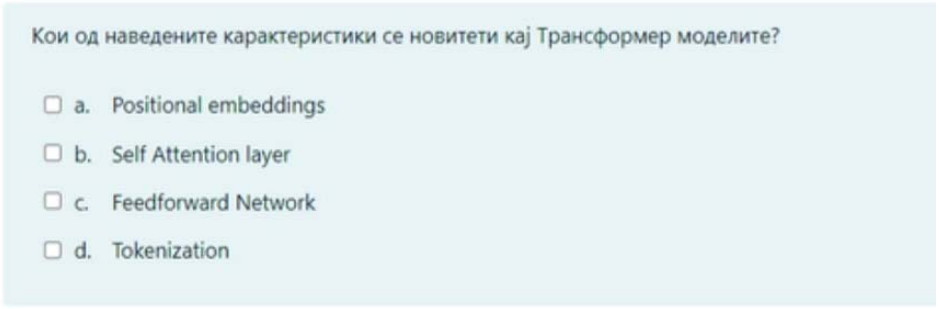
13. Која активациска функција е претставена на графикот?
    - relu
    - linear
    - sigmoid – ВОА Е ТОЧНО



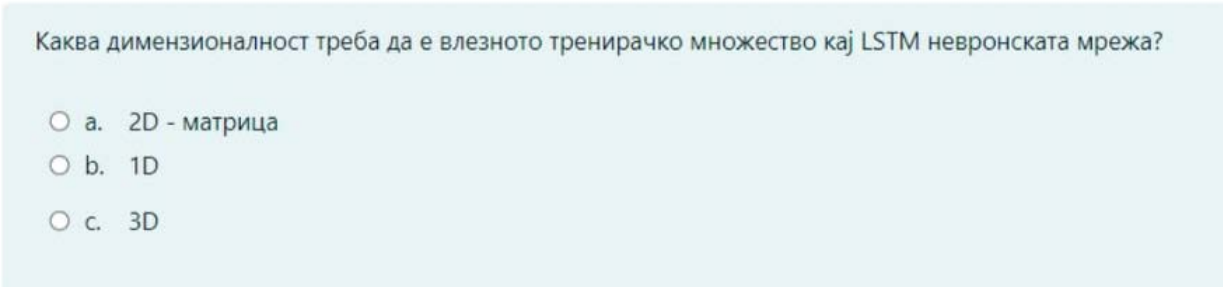
14. Каков вид на учење се реализира кај Автоенкодерите?
    - Надгледувано (supervised)
    - Полу-надгледувано (semi-supervised) - ТОЧНО
    - Само-надледувано (self-supervised) – ТОЧНО
    - Со поттикнување (reinforcement)



15. Кои од наведените карактеристики се новитети кај Трансформер моделите?
    - Positional embeddings – ВОА - ТОЧНО
    - Self Attention layer - ТОЧНО
    - Feedforward Network
    - Tokenization - ВОА - ТОЧНО



16. Каква димензионалност треба да е влезното тренирачко множество кај LSTM невронската мрежа?
    - 2D – матрица – А МОЖЕ И ВОА ДА Е
    - 1D
    - 3D – ТОЧНО



17. Што е точно за моделот seq2seq?

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


Што е точно за моделот seq2seq?

Select one or more:

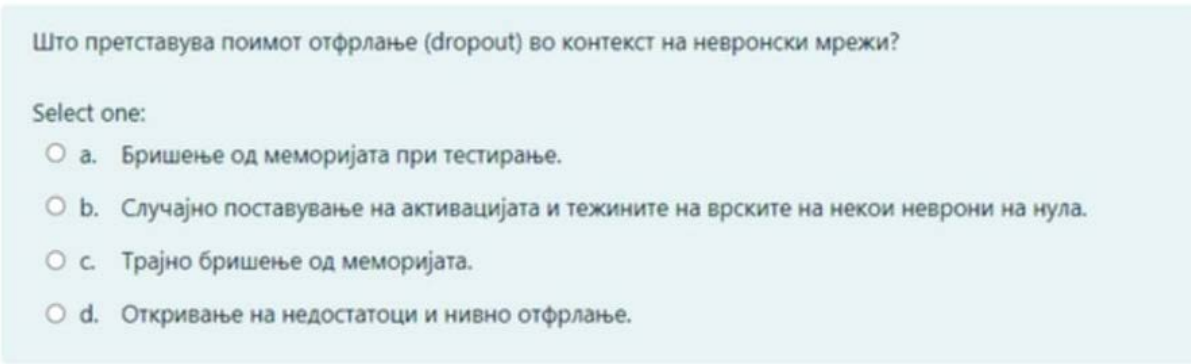
a. Крајниот скриен слој на енкодерскиот дел е влезен слој за декодерскиот дел.

b. Обуката се одвива како и кај другите Рекурентни невронски мрежи.

c. Предноста на seq2seq е што целото значење на реченицата е претставено во крајниот скриен слој на енкодерскиот дел.

d. При тестирањето се генерираат збор по збор, сè додека не се добие на излез знак за крај на реченицата.

18. Што претставува поимот отфрлање (dropout) во контекст на невронски мрежи?
    - Бришење од меморијата при тестирање. – ВОА Е ТОЧНО
    - Случајно поставување на активацијата и тежините на врските на некои неврони на нули.
    - Трајно бришење од меморијата.
    - Откривање на недостатоци и нивно отфрлање.


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

Select one:

a. Бришење од меморијата при тестирање.

b. Случајно поставување на активацијата и тежините на врските на некои неврони на нула.

c. Трајно бришење од меморијата.

d. Откривање на недостатоци и нивно отфрлање.

19. Кои од наведените параметри се дел од хиперпараметрите за тренирање на XGBoostмоделот?
    - n_estimators – ВОА Е ТОЧНО
    - min_depth
    - learning_rate - ВОА Е ТОЧНО
    - max_depth - ВОА Е ТОЧНО

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

20. Што резултат враќа дадениот код: df.isnull()
    - Целата табела (df) со True/False вредности во зависност дали на дадената позиција има/нема NAN вредност
    - Целата табела (df) само со позициите каде има NAN вредност
    - Целата табела (df) само со позициите каде нема NAN вредност
    - Број на NAN вредности по колона

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

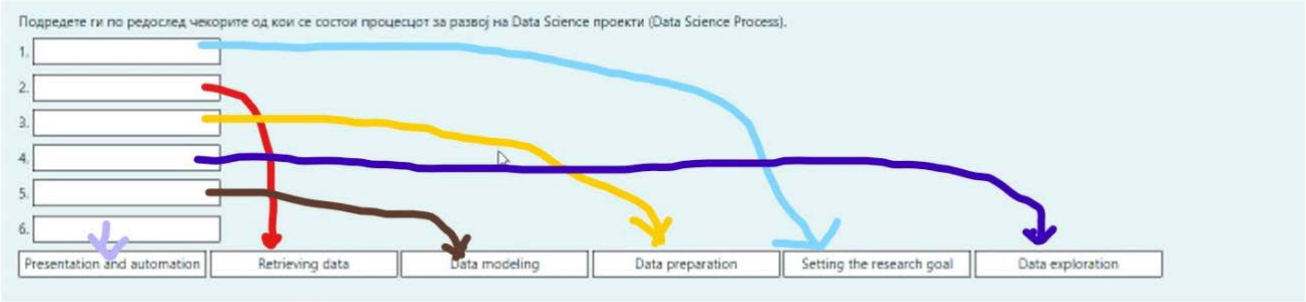
21. Кои од визуелизациите е најдобро да се изберат кога станува збор за датасет од категориски податоци?
    - Dot plot
    - Scatter plot
    - Histogram
    - Bar chatts

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

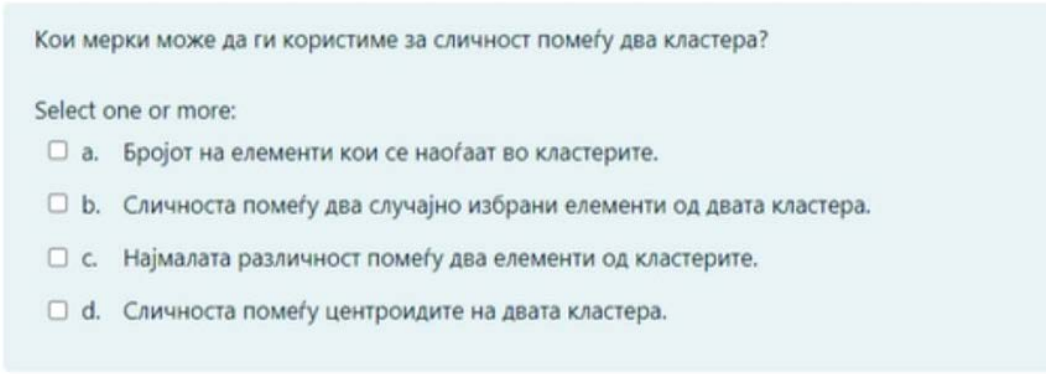
Select one or more:

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

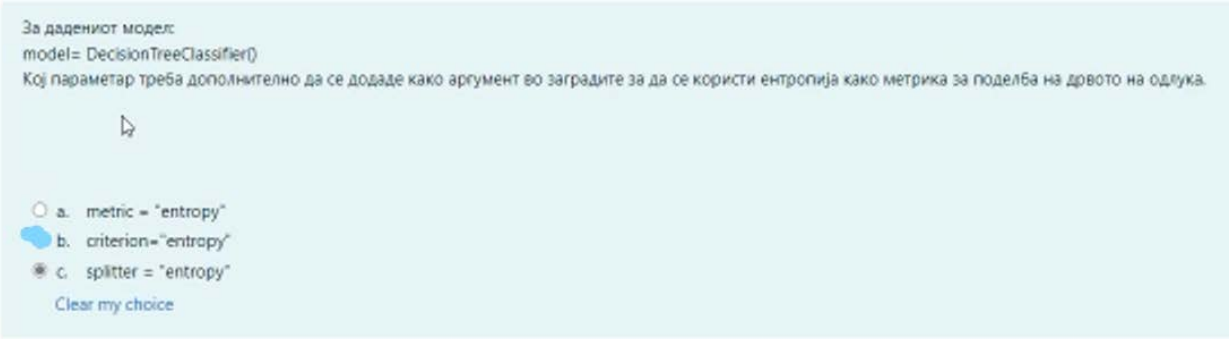
22. Подредете ги по редослед чекорите од кои се состои процесот за развој на Data Science проекти (Data Science Process).
    - Setting the research goal
    - Retrieving data
    - Data preparation
    - Data exploration
    - Data modeling
    - Presentation and automation



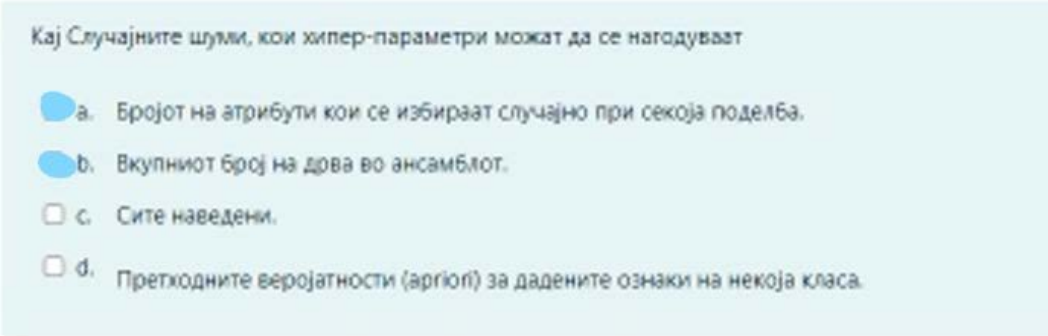
23. Кои мерки може да ги користиме за сличност помеѓу два кластера?
    - Бројот на елементи кои се наоѓаат во кластерите.
    - Сличноста помеѓу два случајно избрани елементи од двата кластера.
    - Најмала различност помеѓу два елементи од кластерите. - ТОЧНО
    - Сличноста помеѓу центроидите на двата кластера. - ТОЧНО



24. За дадениот модел:
    model = DecisionTreeClassifier()
    Кој параметар треба дополнително да се додаде како аргумент во заградите за да се користи ентропијата како метрика за поделба на дрвото на одлука.
    - metric = “entropy”
    - criterion = “entropy”
    - spitter = “entropy”



25. Кај Случајните шуми, кои хипер-параметри можат да се нагодуваат.
    - Бројот на атрибути кои се избираат случајно при секоја поделба.
    - Вкупниот број на дрва на ансамблот.
    - Сите наведени
    - Претходните веројатности (aprion) за дадените ознаки на некоја класа.



26. 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. површина на земјиштето

    Што ќе биде излезот на дадениот код:

    model.coef_

    - Три Коефициенти (децимални вредности) за секој од влезните податоци
    - Еден коефициент (делимална вредност) за сите влезни податоци
    - Четирите коефициенти (децимални вредности) за секој од влезните податоци плус интерцелптот.

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

27. Ако треба да се предвидува вредноста на температурата во даден пластеник во текот на ноќта, за каков вид на машинско учење станува збор?
    - Откривање на недостатоци (Anomaly Detection)
    - Учење со поттикнување (Reinforcement Learning)
    - Класиикација (Classification)
    - Регресија (Regression)

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)

28. Кога дистрибуцијата на податоците е како на сликата, какви се наклонетоста (bias) и варијансата кај овие податоци?
    - Мала наклонетост и мала варијанса
    - Голема наклонетост и мала варијанса
    - Мала наклонетост и голема варијанса
    - Голема наклонетост и голема варијанса

Кога дистрибуцијата на податоците е како на сликата, какви се наклонетоста (bias) и варијансата кај овие податоци?

Select one:

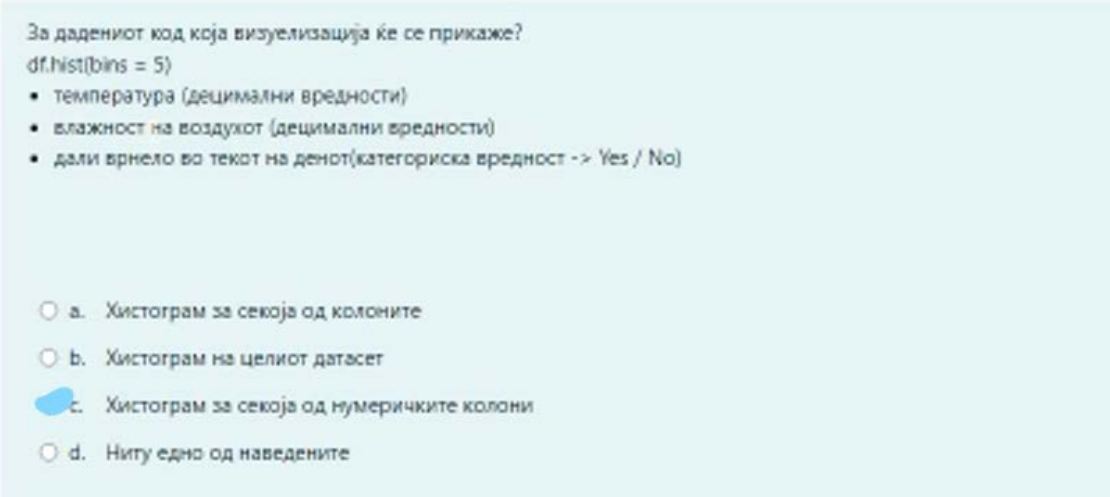
a. Мала наклонетост и мала варијанса

29. 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 (категирусја вредност -> Yes / No)

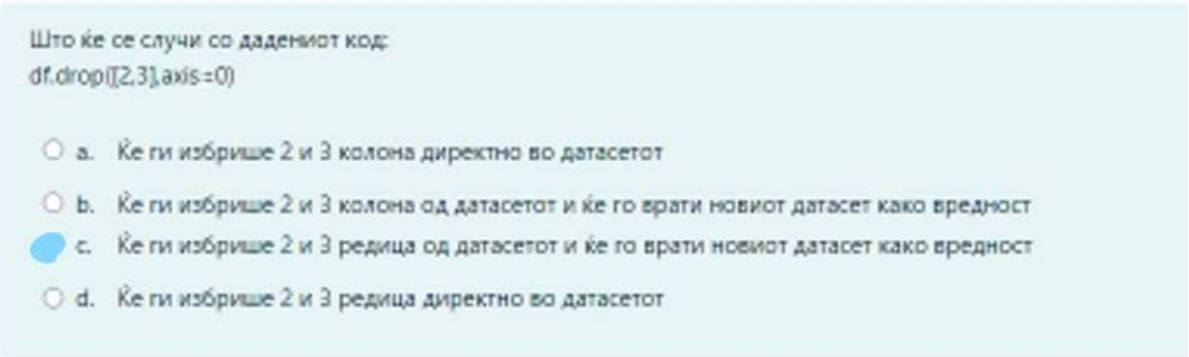
- Хистограм за секоја од колоните
- Хистограм на целиот датасет
- Хистограм за секоја од нумеричките колони
- Ниту едно од наведените



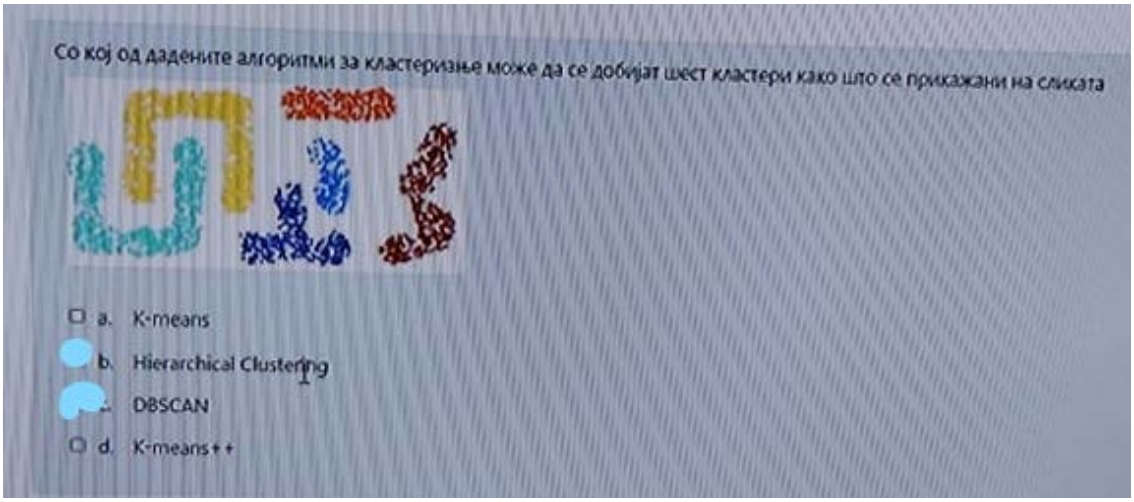
30. Што ќе се случи со дадениот код:

df.drop([2,3], axis=0)

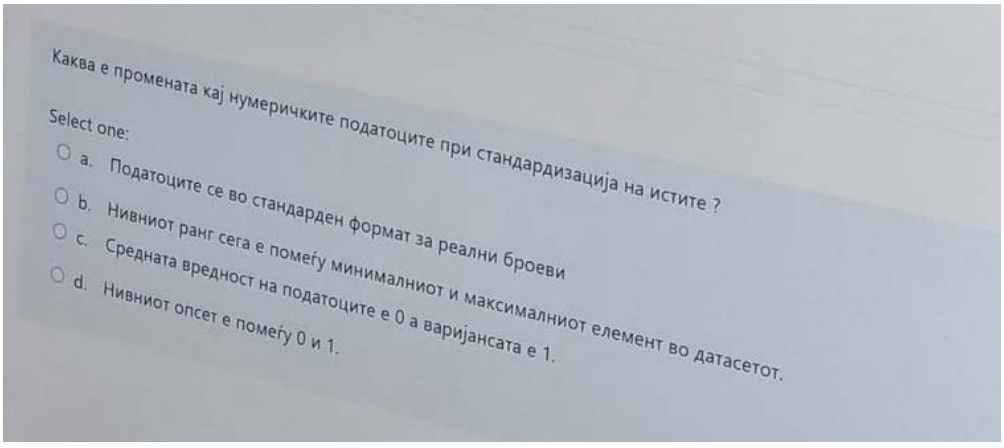
- Ќе ги избрише 2 и 3 колона директно во датасетот
- Ќе ги избрише 2 и 3 колона од датасетот и ќе го врати новиот датасет како вредност
- Ќе ги избрише 2 и 3 редица од датасетот и ќе го врати новиот датает како вредност
- Ќе ги избрише 2 и 3 редица директно во датасетот



31. Со кој од дадените алгоритми за кластерирање може да се добијат шест кластери како што се прикажани на сликата
    - K-means
    - Hierarchical Clustering
    - DBSCAN
    - K-means++



32. Каква е промената кај нумеричките податоци при стандардизација на истите?
    - Податоците се во стандарден формат за реални броеви
    - Нивниот ранг сега е помеѓу минималниот и максималниот елемент во датасетот
    - Средната вредност на податоците е 0, а варијансата е 1 – ИСТО ТОЧНО АЛИ НЕ СИМ СИГУРЕН
    - Нивниот опсег е помеѓу 0 и 1 - ТОЧНО



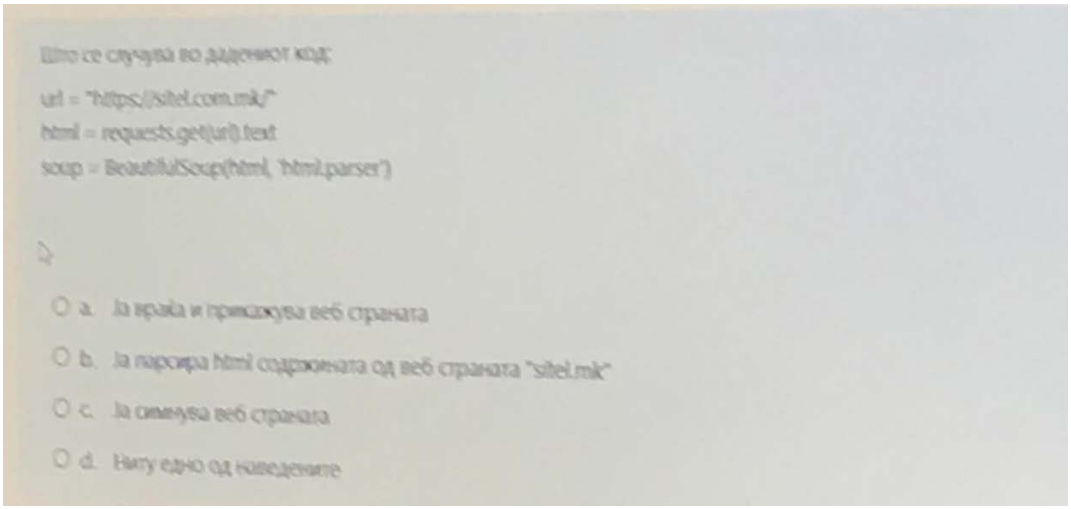
33. Што се случува во дадениот код?

url = https://sitel.com.mk/

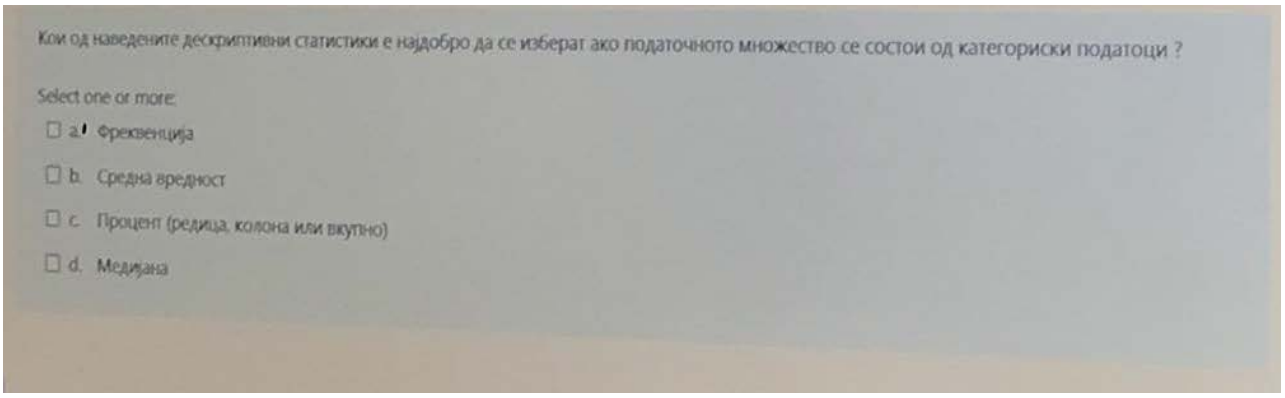
html = requests.get(url).text

soup = BeautifulSoup(html, ‘html.parser’)

- Ја враќа и прикажува веб страната
- Ја парсира html содржината од веб страна “sitel.mk” – МИСЛАМ ДЕКА Е ВИА АЛИ НЕ СУМ СИГУРЕН
- Ја симнува веб страната
- Ниту едно од наведените



34. Кои од наведените дескриптивни статистики е најдобро да се изберат ако податочното множество се состои од категориски податоци?
    - Фреквенција
    - Средна вредност
    - Процент (редица, колона или вкупно)
    - Медијана

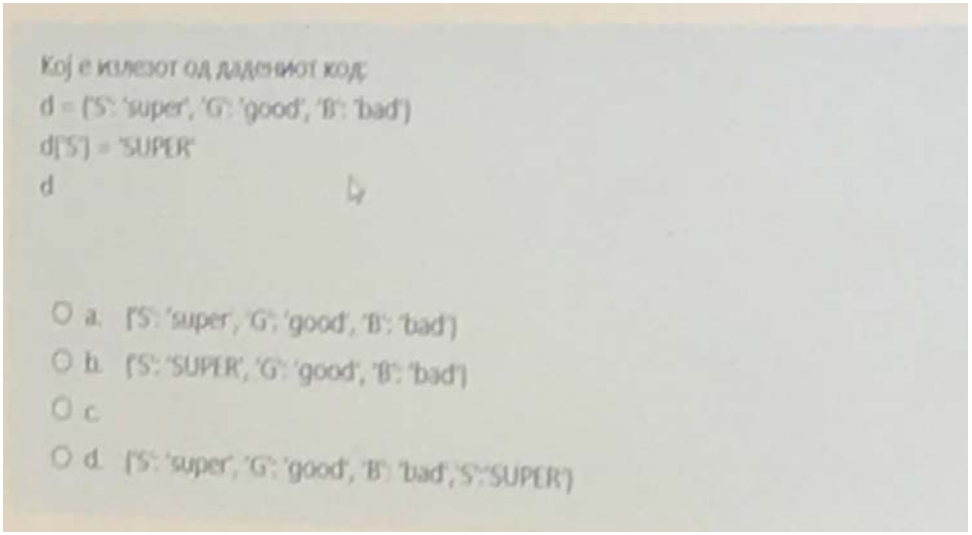


35. Кој е излезот од дадениот код:

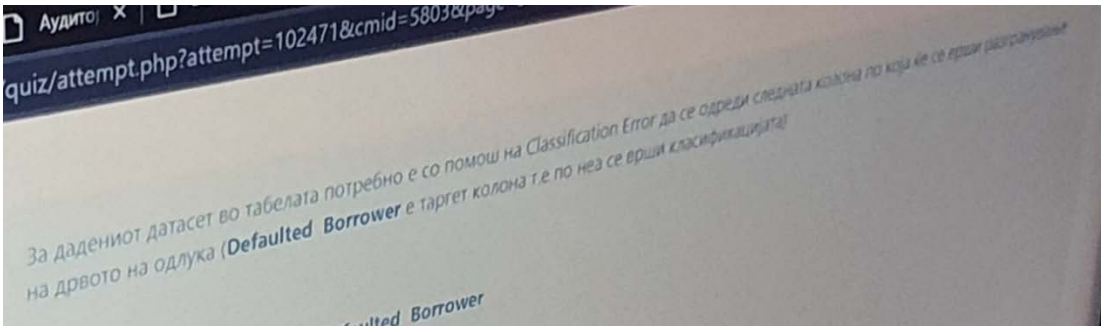
d = [‘S’: ‘super’, ‘G’: ‘good’, ‘B’: ‘bad’]

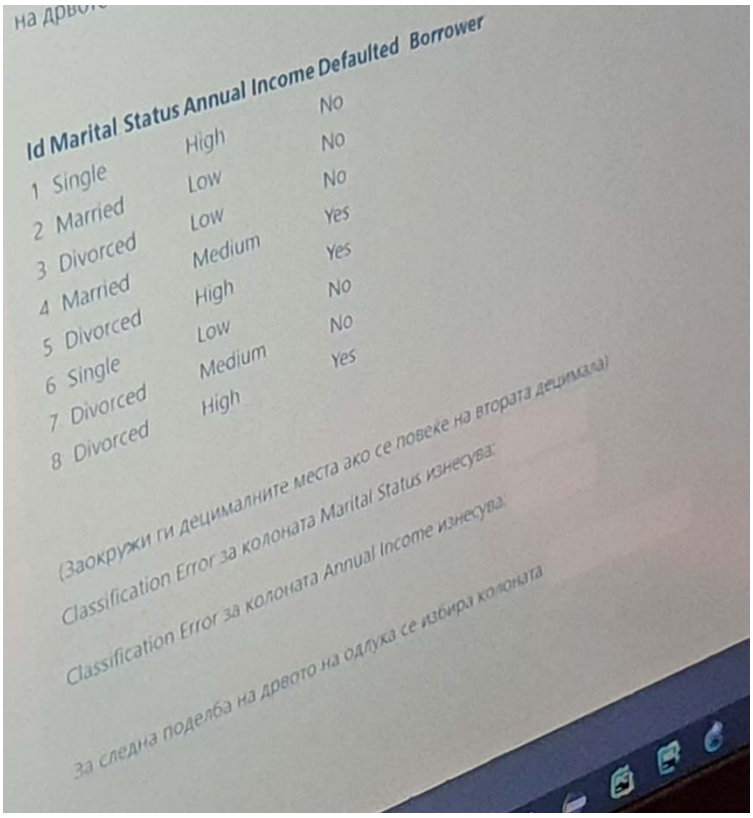
d[‘S’] = ‘SUPER’d

- [‘S’: ‘super’, ‘G’: ‘good’, ‘B’: ‘bad’]
- [‘S’: ‘SUPER, ‘G’: ‘good’, ‘B’: ‘bad’] - ТОЧНО
-
- [‘S’: ‘super’, ‘G’: ‘good’, ‘B’: ‘bad’, ‘S’: ‘SUPER]

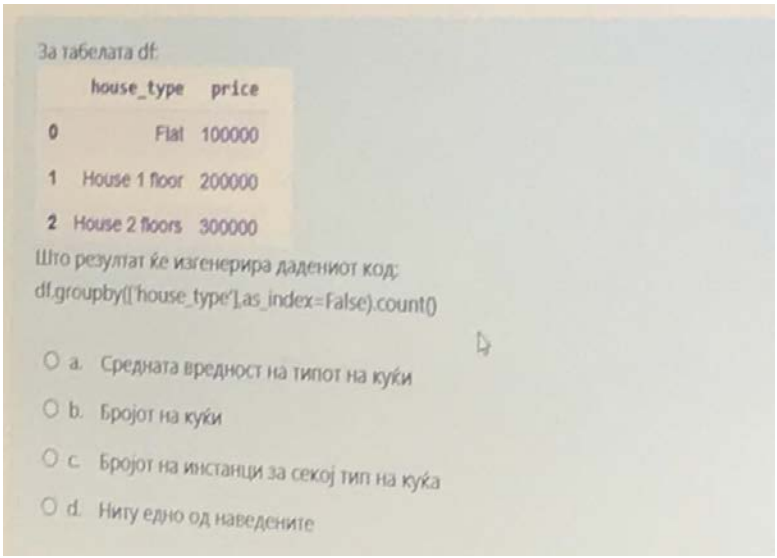


36. За дадениот датасет во табелата потребно е со помош на Classification Error да се одреди следната колона по која ќе се врши разгранување на дрвото на одлика (Defaulted Borrower е таргет колона т.е. по неа се врши класификациите).

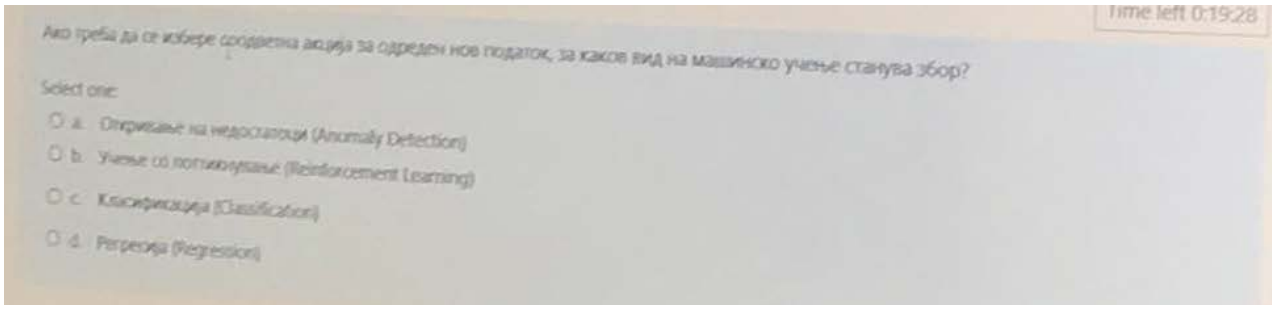




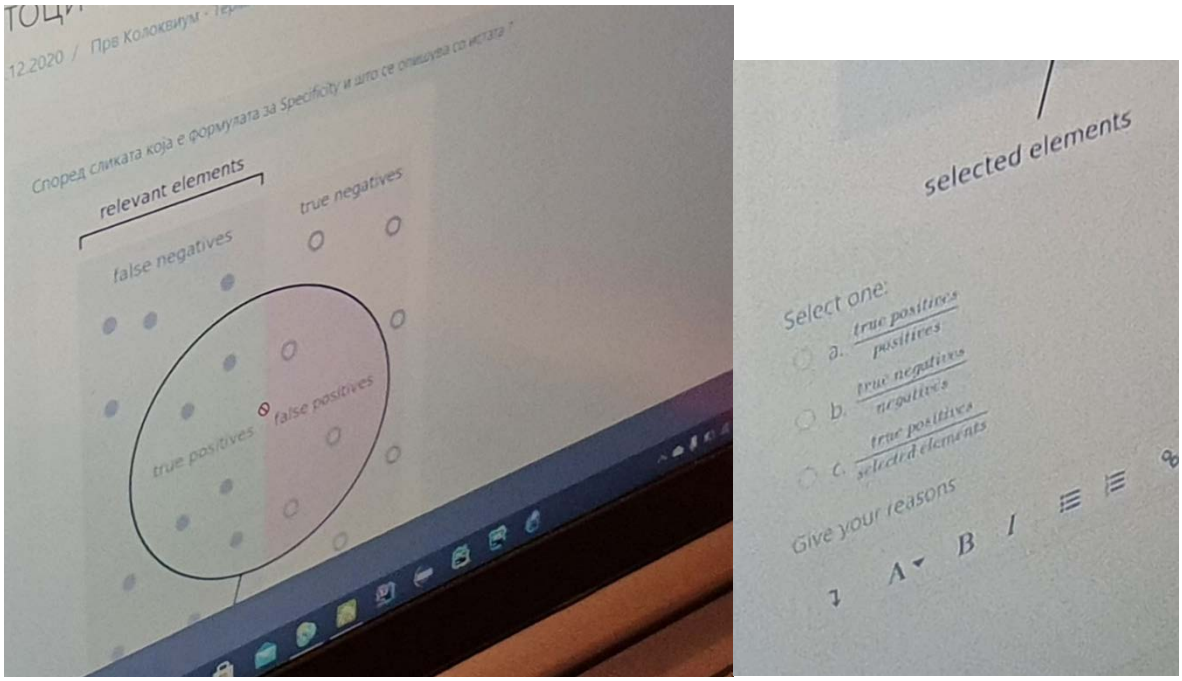
37. За табелата df, што резултат ќе изгенерира дадениот код:
df.groupby([‘house_type’], as_index=False).count()
    - Средната вредност на типот на куќи
    - Бројор на куќи
    - Бројор на инстанци за секој тип на куќа - ТОЧНО
    - Ниту едно од наведените



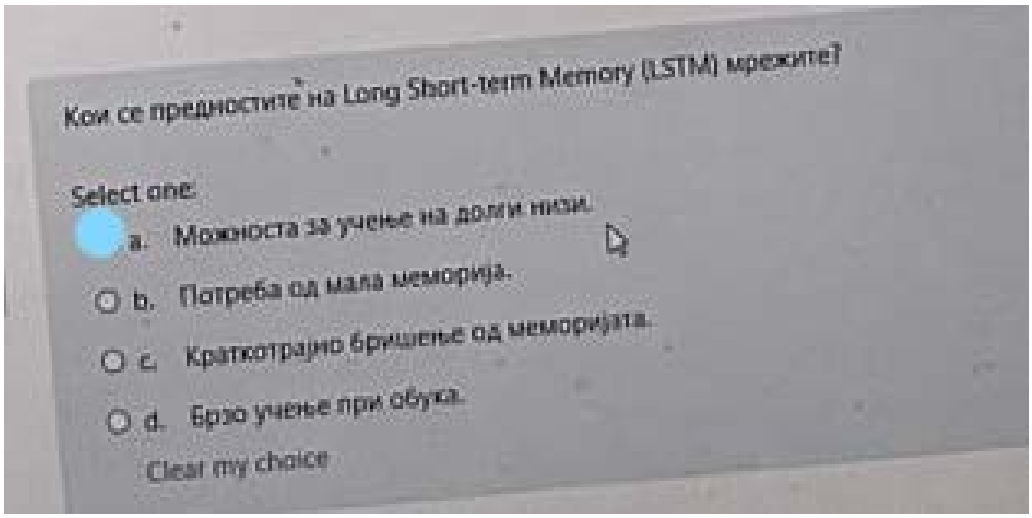
38. Ако треба да се избере соодветна акција за одреден нов податок, за каков вид на машинско учење станува збор?
    - Откривање на недостатоците (Anomaly Detection)
    - Учење со поттикнување (Reinforcement Learning)
    - Класификација (Classification)
    - Регресија (Regression)



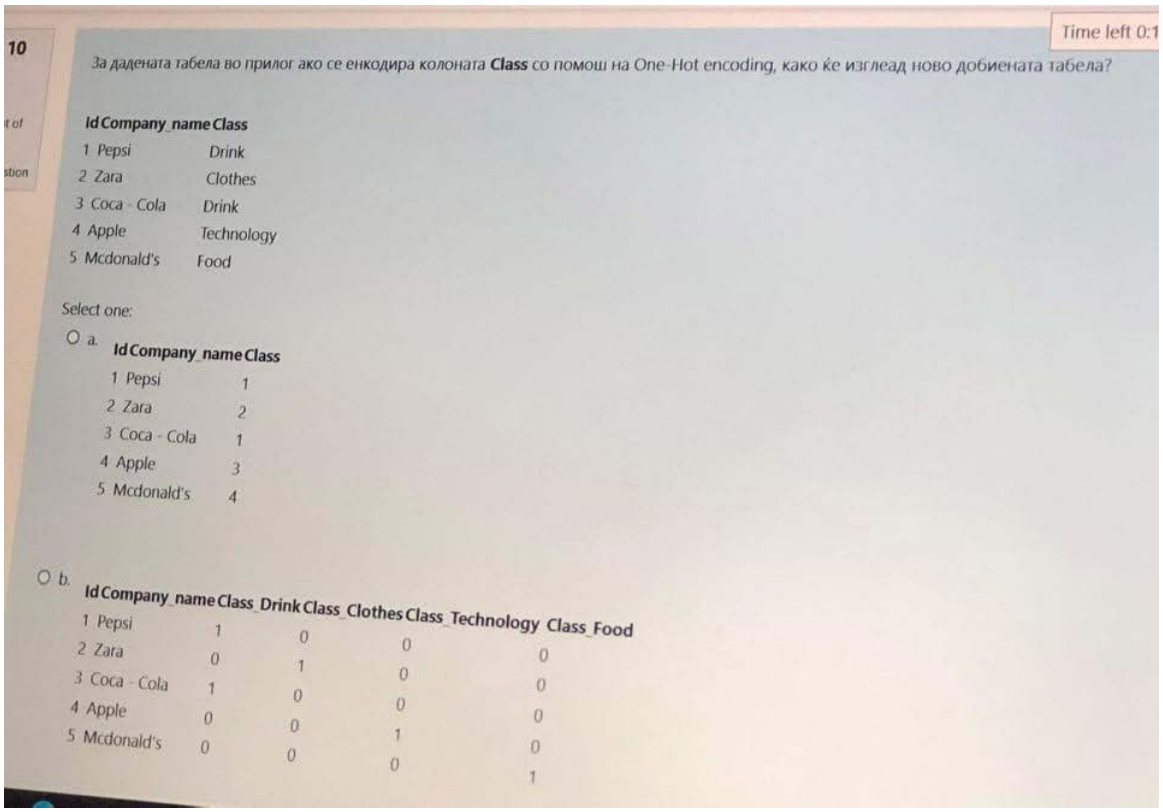
39. Според сликата која е формулата за Specificity и што се опишува со истата?
    - true positives / positives
    - true negatives / negatives – НЕ СУМ СИГУРЕН АМА МИСЛАМ ДЕКА Е ВОА
    - true positives / selected elements

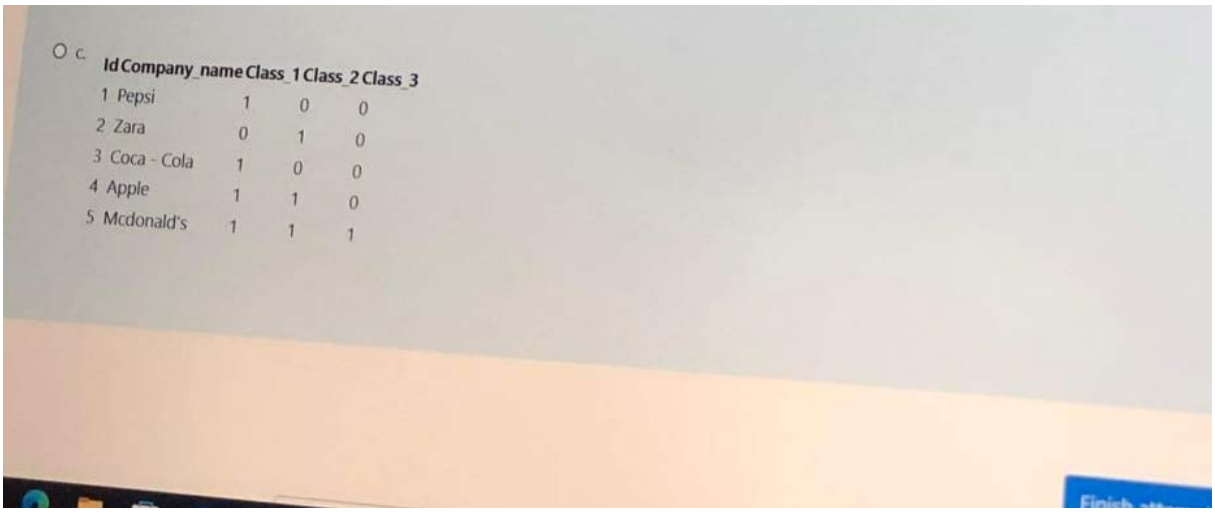


40. Кои се предностите на Long Short-term Memory (LSTM) мрежите?
    - Можноста за учење на долги низи
    - Потреба од мала меморија
    - Краткотрајно бришење од меморијата
    - Брзо учење при обука

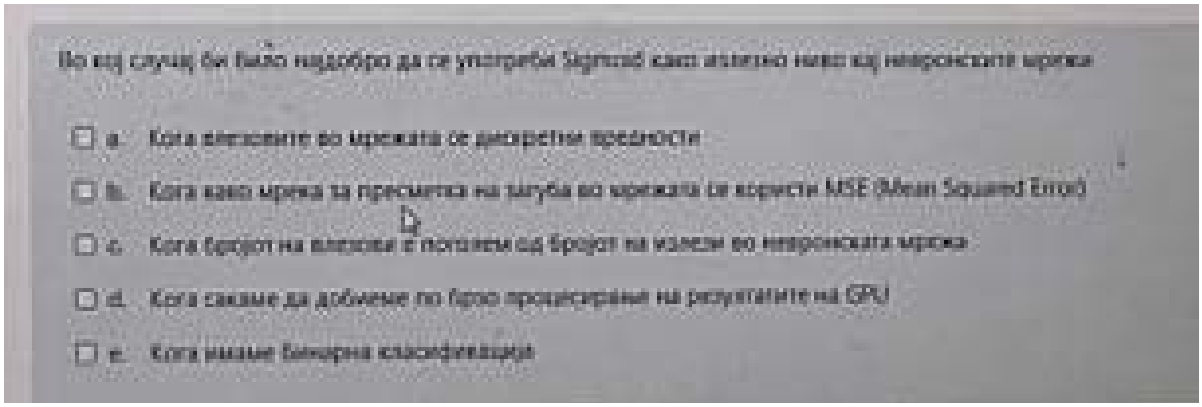


41. За дадената табела во прилог ако се енкодира колоната Class со помош на Binary Encoding како ќе изгледа ново добиенатаа табела? - б е ТОЧЕН

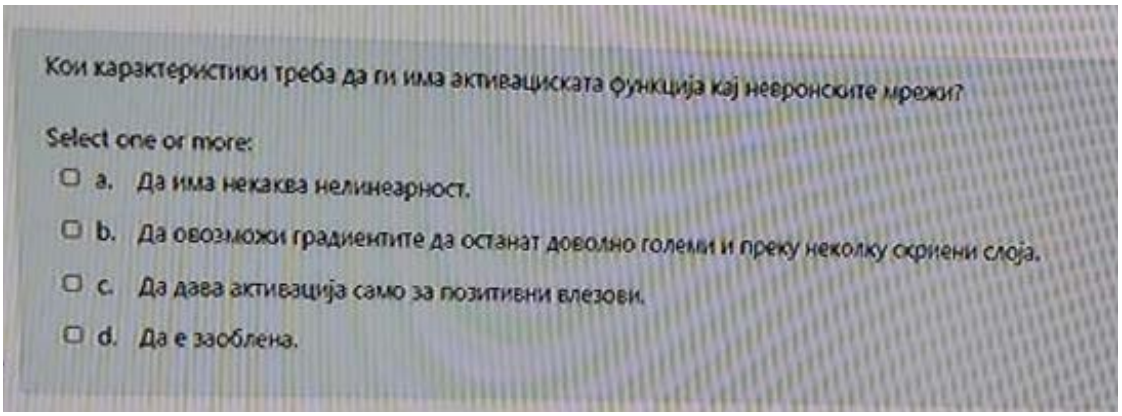




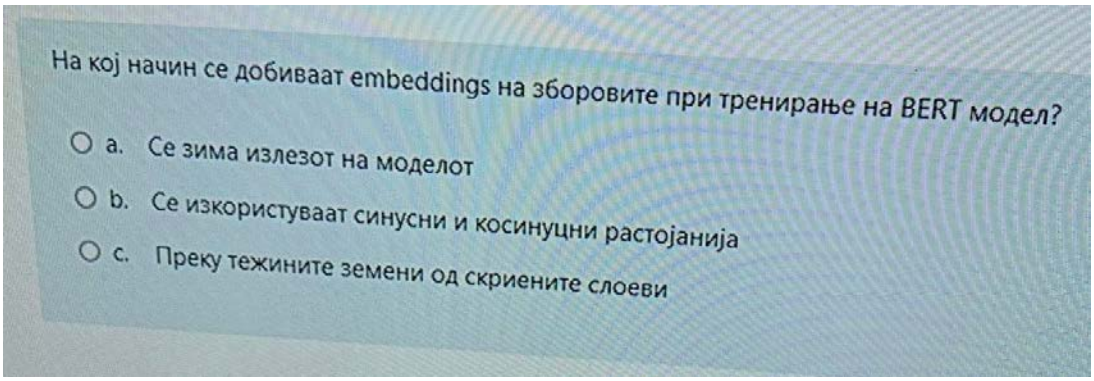
42. Во кој случај би било најдобро да се употреби Si.. како излезно ниво кај невронските мрежи?
    - Кога влезовите во мрежата се дискретни вредности
    - Кога како мрежа за пресметка на загуба во мрежата се користи MSE (Mean Squared Error)
    - Кога бројот на влезови е поголем од бројот на излези во нервонската мрежа
    - Кога сакаме да добиеме побрзо процесирање на резултатите на GPU
    - Кога имаме бинарна класификација
    - У ГРУПАТА ПИШЕ ДЕКА СЕ Ц И Е



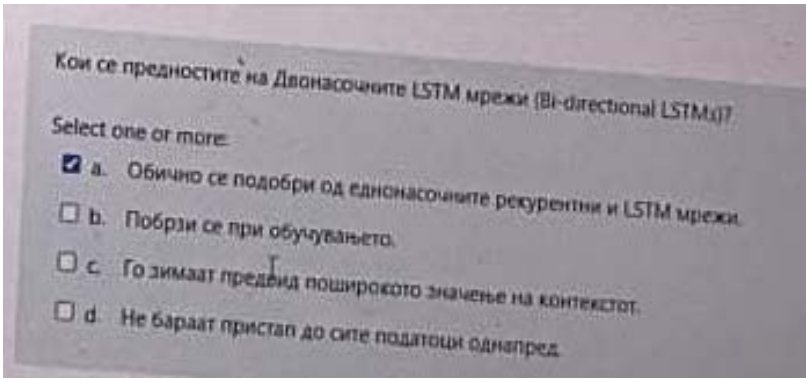
43. Кои карактеристики треба да ги има активациската функција кај невронските мрежи?
    - Да има некаква нелинеарност - ТОЧНО
    - Да овозможи градиентите да останат доволно големи и преку неколку скриени слоја
    - Да дава активација само за позитивни влезови
    - Да е заоблена



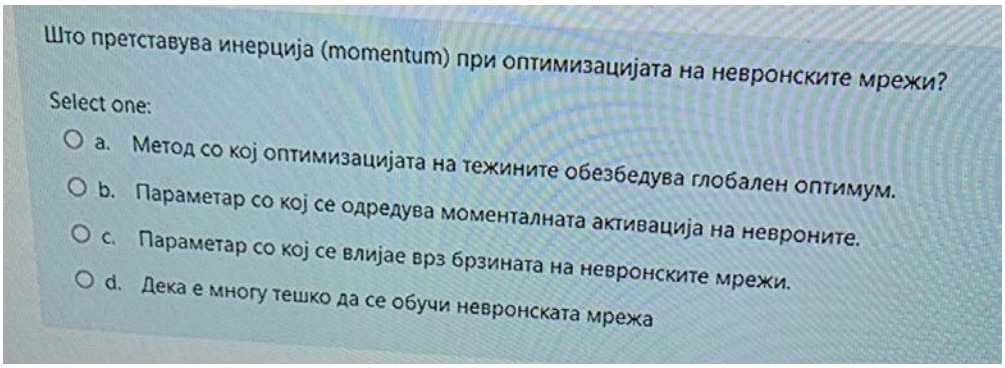
44. На кој начин се добиваат embedding на зборовите при тренирање на BERT модел?
    - Се зима излезот на моделот - ТОЧНО
    - Се искористуваат синусни и косинусни растојанија
    - Преку тежините земени од скриените слоеви



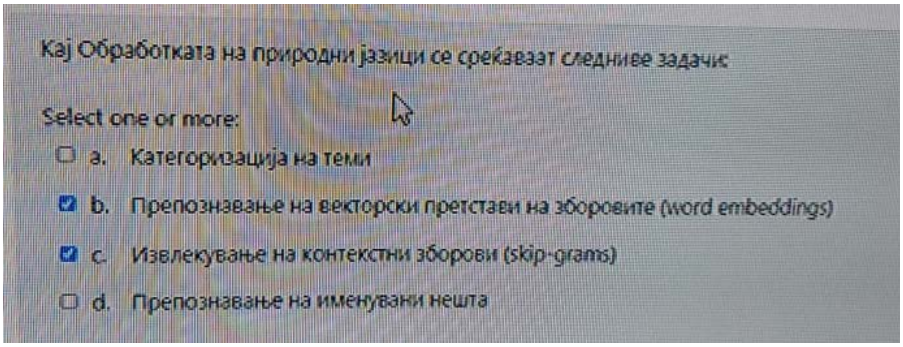
45. Кои се предностите на Двонасочните LSTM мрежи (Bi-directional LSTMx)?
    - Обично се подобри од еднонасочните рекурентни и LSTM мрежи - ТОЧНО
    - Побрзи се при обучувањето
    - Го земаат предвид поширокото значење на контекстот – ТОЧНО (НЕ СУМ СИГУРЕН)
    - Не бараат пристап до сите податоци однапред



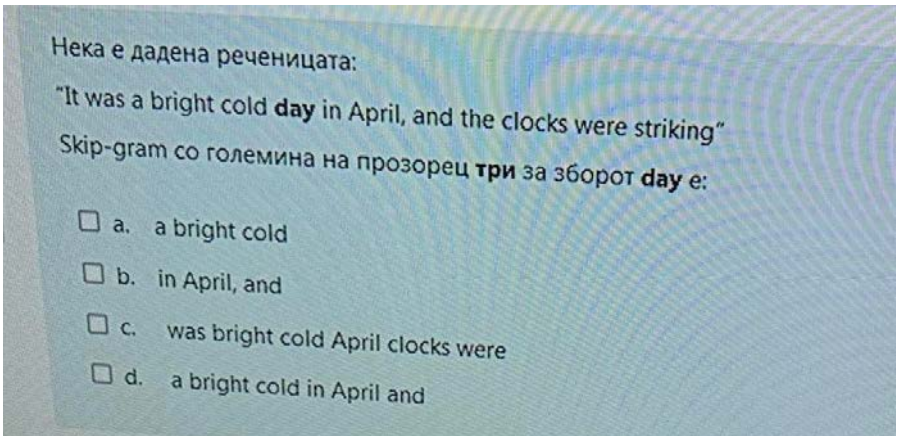
46. Што претставува инерција (momentum) при оптимизацијата на невронските мрежи?
    - Метод со кој оптимизацијата на тежините обезбедува глобален оптимум.
    - Параметар со кој се одредува моменталната активација на невроните.
    - Параметар со кој се влијае врз брзината на невронските мрежи.
    - Дека е многу тешко да се обучи невронската мрежа.



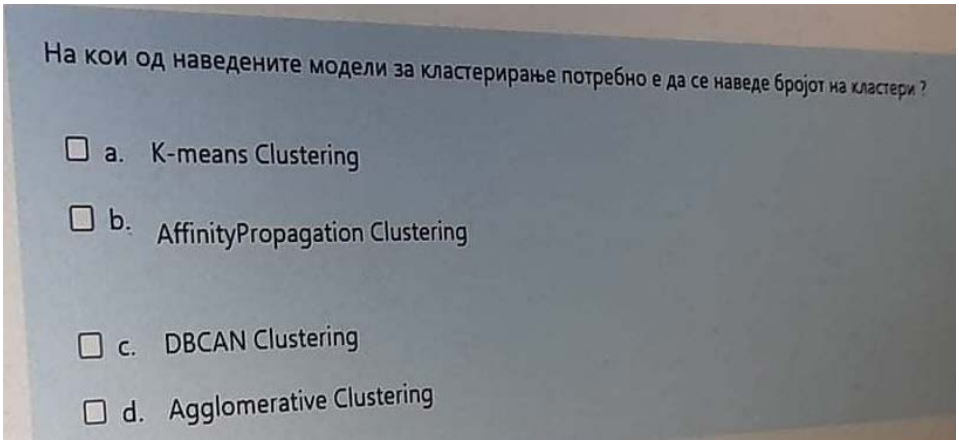
47. Кај Обработката на природни јазици се среќаваат следниве задачи:
    - Категоризација на теми
    - Препознавање на векторски претстави на зборовите (word embeddings)
    - Извлекување на контекстни зборови (skip-grams)
    - Препознавање на именувани нешта – И ВОА Е ТОЧНО



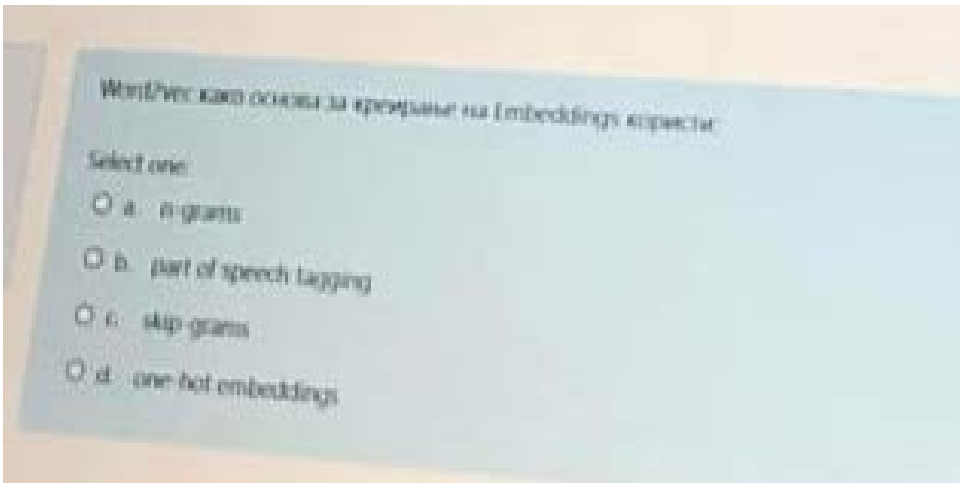
48. Нека е дадена реченицата:
    "It was a bright cold day in April, and the clocks were striking"
    Skip-gram со големина на прозорец три за зборот day е:
    - a bright cold
    - in April, and
    - was bright cold April clocks were
    - a bright cold in April and - ТОЧНО



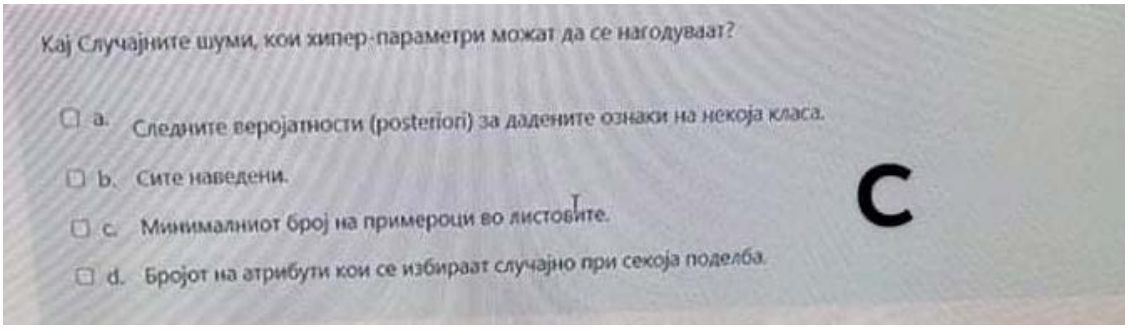
49. На кои од наведените модели за кластерирање потребно е да се наведе бројот на кластери?
    - K-means Clustering - ТОЧНО
    - AffinityPropagation Clustering
    - DBCAN Clustering
    - AggomerativeClutering - ТОЧНО



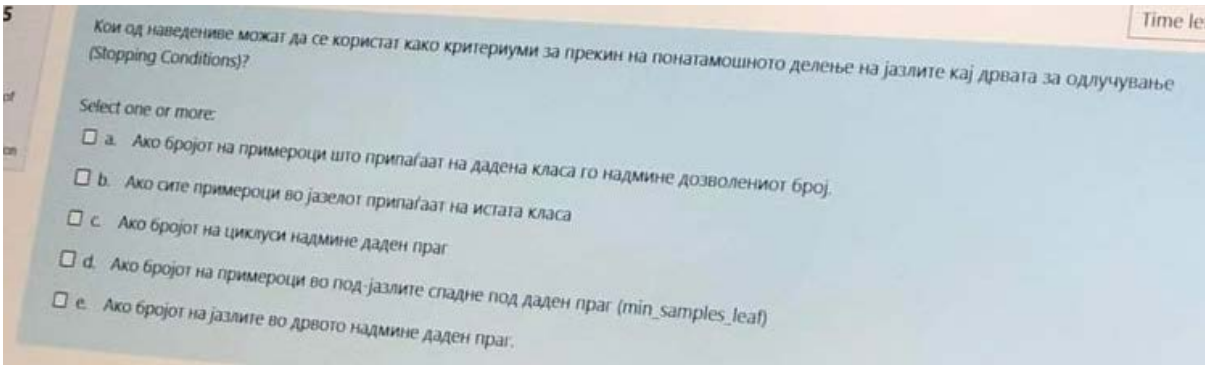
50. Word2vec како основа за креирање на Embeddings користи:
    - n-grams
    - part of speech tagging
    - skip-grams – ТОЧНО
    - one-hot embeddings



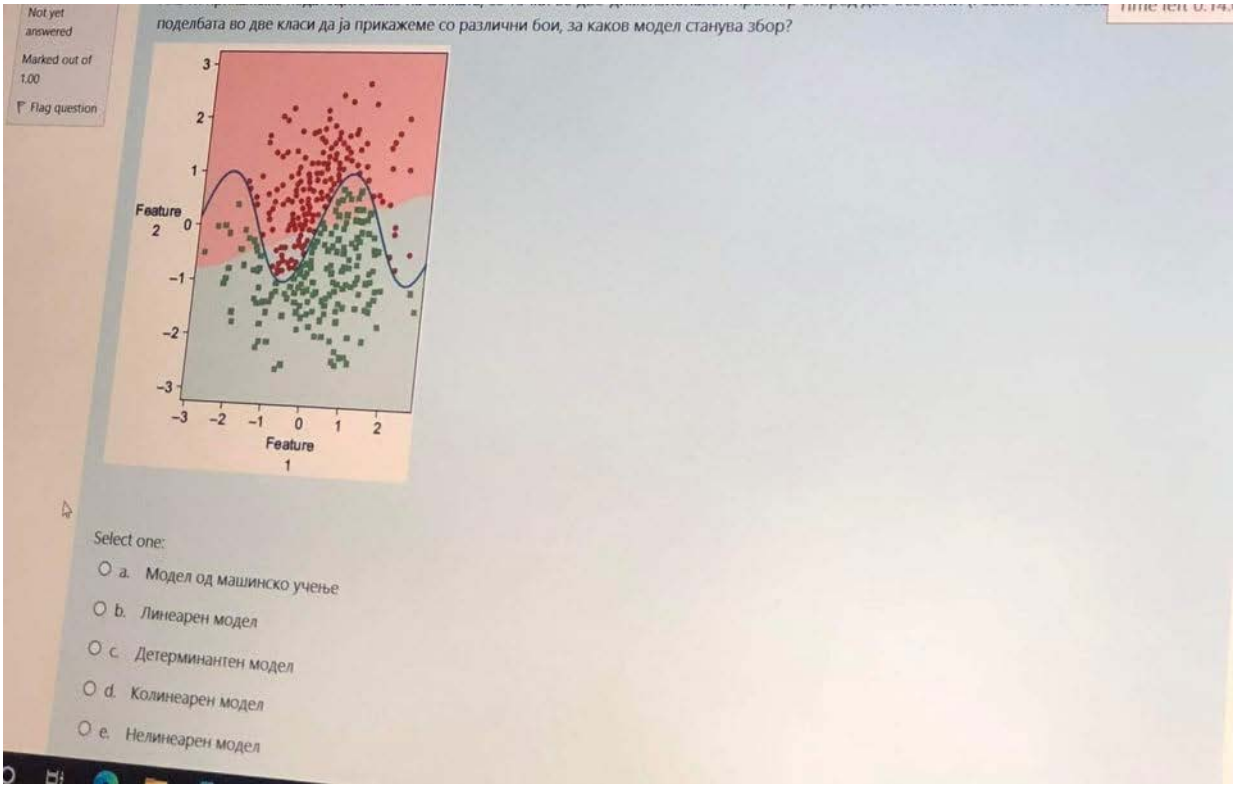
51. кај Случајните шуми, кои хипер-параметри можат да се нагодуваат? – ПОД А И Д АМА НЕ СУМ СИГУРЕН
    - Следните веројатности (posteriori) за дадените ознаки на некоја класи
    - Сите наведени
    - Минималниот број на примероци во листовите
    - Бројот на атрибути кои се избираат случајно при секоја поделба. – И ВОА Е ТОЧНО МИСЛАМ СПОРЕД ТОА ШО ПИШЕЛЕ У ГРУПАТА



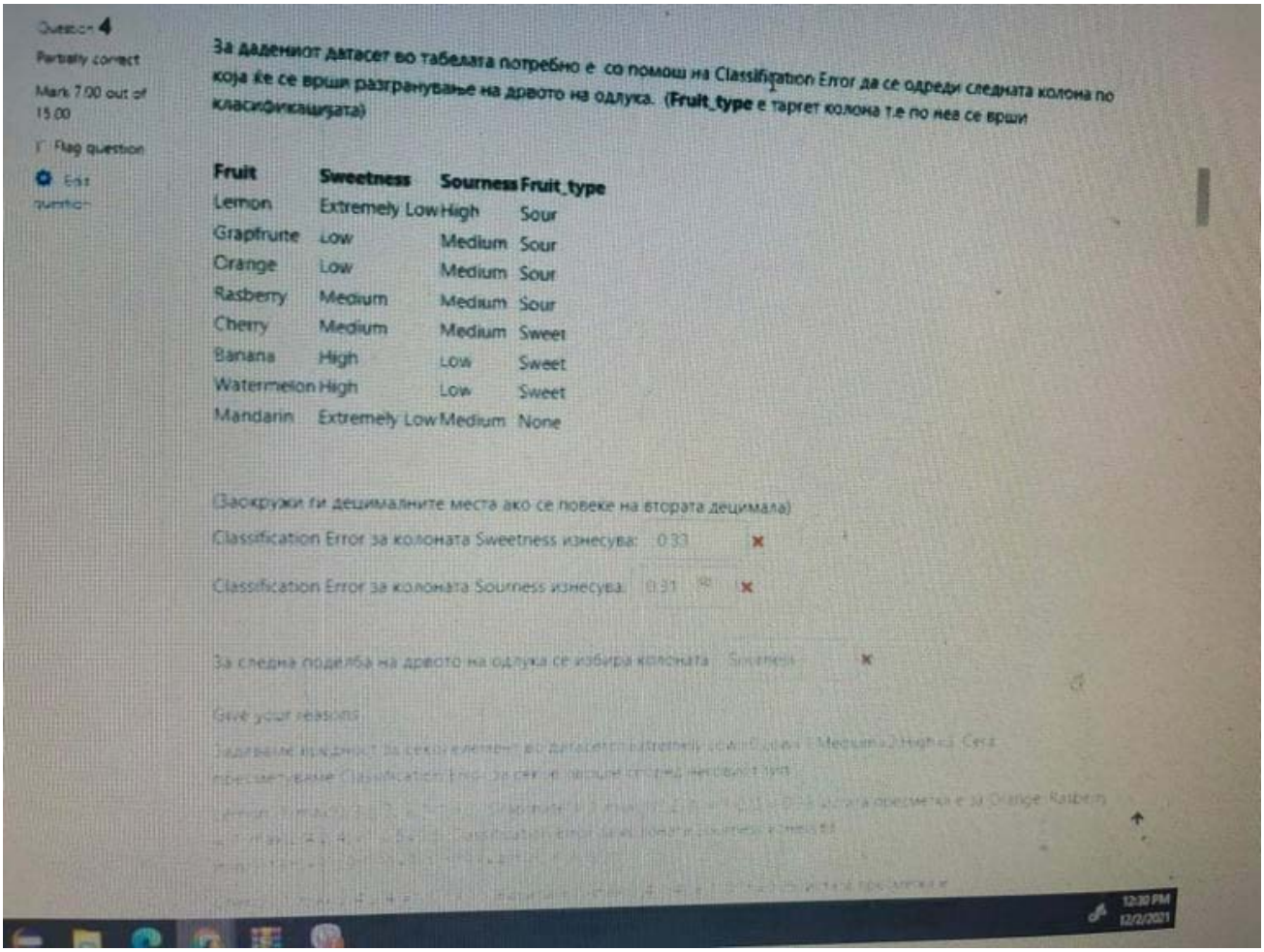
52. Кои од наведените можат да се користат како критериуми за прекин на понатамошното делење на јазлите на дрвата за одлучување (Stopping Conditions)?
    - Ако бројот на примероци што припаѓаат на дадена класа го надмине дозволениот број
    - Ако сите примероци во јазелот припаѓаат на истата класа
    - Ако бројот на циклуси надмине даден праг
    - Ако бројот на примероци во под-јазлите спадне под даден праг (min_samples_leaf)
    - Ако бројот на јазлите во дрвото надмине даден праг



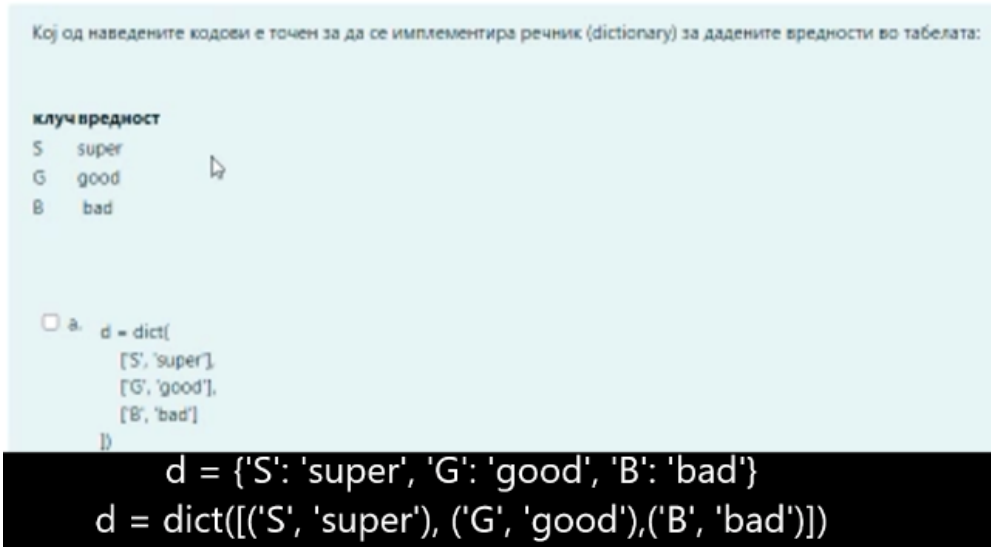
53. Ако и прикажеме податоците како на сликата, со точки во дво-димензионален....
    - Модел од машинското учење
    - Линеарен модел – ТОЧНО
    - Детерминантен модел
    - Колинеарен модел
    - Нелинеарен модел



54. За дадениот датасет во табелата потребно е со помош на Classification Error да се одреди следната колона по која ќе се врши разгранување на одлука. (Fruit_typee таргет колона т.е. по неа се врши класификацијата). Sweetnes – 0.25, Sourness – 0.25, Sourness



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



56.

Овој дел е само за проверка на точноста!
Задачата мора да се прикачи во блокот погоре!

На променливата **your_answers** доделете и листа со предвидени вредности на индексот DJIA врз основа на податоците од dataset3_test.csv. Листата треба да има 126 вредности. За проверка на точноста се користи MSE.

Кликнете **[Check]** за да се провери точноста.

Answer: (penalty regime: 0 %)

Reset answer

```
1 your_answers = []
2
3
```

Check

Finish attempt ...

57.

Даден е dataset (dataset1.csv) во која има податоци поврзани со здравјето на фетусот (плодот) кај мајките. Целта е да се направи невронска мрежа кој ќе овозможува предвидување на состојбата на плодот дадена во колоната **fetal_health.** Состојбата може да има три различни врендости (класи):

- 1=Normal
- 2=Suspect
- 3=Pathological

Невронската мрежа треба да ги содржи следниве слоеви:

1. Конволуциски слој составен од 32 неврона, со големина на прозорец 3 и ист padding.
2. Конволуциски слој со 16 неврона, со големина на прозорец 1 и без padding
3. Dense слој кој ќе има само 10 неврони.
4. Сите слоеви имаат relu активациска функција.

Останатите параметри на мрежата како што се излезниот слој, излезната активациска функција, функцијата на загуба (loss function), отпимизатор, метрики за точност, како и останати параметри ако има потерба, потребно е да ги дефинирате вие.

Дополнително потребно е да се внесе некој механизам за спречување на overfitting.

Бројот на епохи на кои треба да ја тренирате невронската мрежа е 6 со големина на batch од 32.

Урнек за задачата на Colab или .py скрипта

Решената задача потребно е да ја прикачите на оваа задача како notebook датотека (**.ipynb**) или python скрипта (**.py**).

Ако имате било каков дополнителен коментар внесете го во текстуалното поле.

2. Конволуциски слој со 16 неврона, со големина на прозорец 1 и без padding
3. Dense слој кој ќе има само 10 неврони.
4. Сите слоеви имаат relu активациска функција.

Останатите параметри на мрежата како што се излезниот слој, излезната активациска функција, функцијата на загуба (loss function), отпимизатор, метрики за точност, како и останати параметри ако има потерба, потребно е да ги дефинирате вие.

Дополнително потребно е да се внесе некој механизам за спречување на overfitting.

Бројот на епохи на кои треба да ја тренирате невронската мрежа е 6 со големина на batch од 32.

Урнек за задачата на Colab или .py скрипта

Решената задача потребно е да ја прикачите на оваа задача како notebook датотека (**.ipynb**) или python скрипта (**.py**).

Ако имате било каков дополнителен коментар внесете го во текстуалното поле.

Maximum file size: 256MB, maximum number of files: 2

Files

58.

Даден е dataset (dataset2.csv) кој содржи твитови со информации поврзани со берза. Вашата задача е да ги кластрирате дадените твитови.

1. Најпрво употребете **Universal Sentence Encoder** за креирање на embeddings за секој твит.
2. Потоа со помош на **PCA** да се изврши намалување на димензионалноста на добиените embeddings така што сега секој од векторите ќе има 10 димензии.
3. Применете **KMeans** како метод за кластерирање
4. Најдете го оптималниот број на кластери
5. Направете 2D визуелизација на кластерите

Урнек за задачата на следните линкови: Colab или .py скрипта
Решената задача потребно е да ја прикачите на оваа задача како notebook датотека (**.ipynb**) или python скрипта (**.py**).
Ако имате било каков дополнителен коментар внесете го во текстуалното поле.

Maximum file size: 256MB, maximum number of files: 1

59.

Потребно е да се направи модел за предвидување на финалната вредност (Close) на Dow Jones Industrial Average (DJIA) индексот.

Даден е dataset (dataset3_train.csv) во кој за периодот од **2008-08-08** до **2015-12-31** се дадени следниве информации: **Open** - почетната вредност на DJIA на дадениот датум, десетте најчитани вести во тој ден (**Top1-Top10**) според Reddit WorldNews Channel и колоната **Close** во која е финалната вредноста на индексот за даден ден.

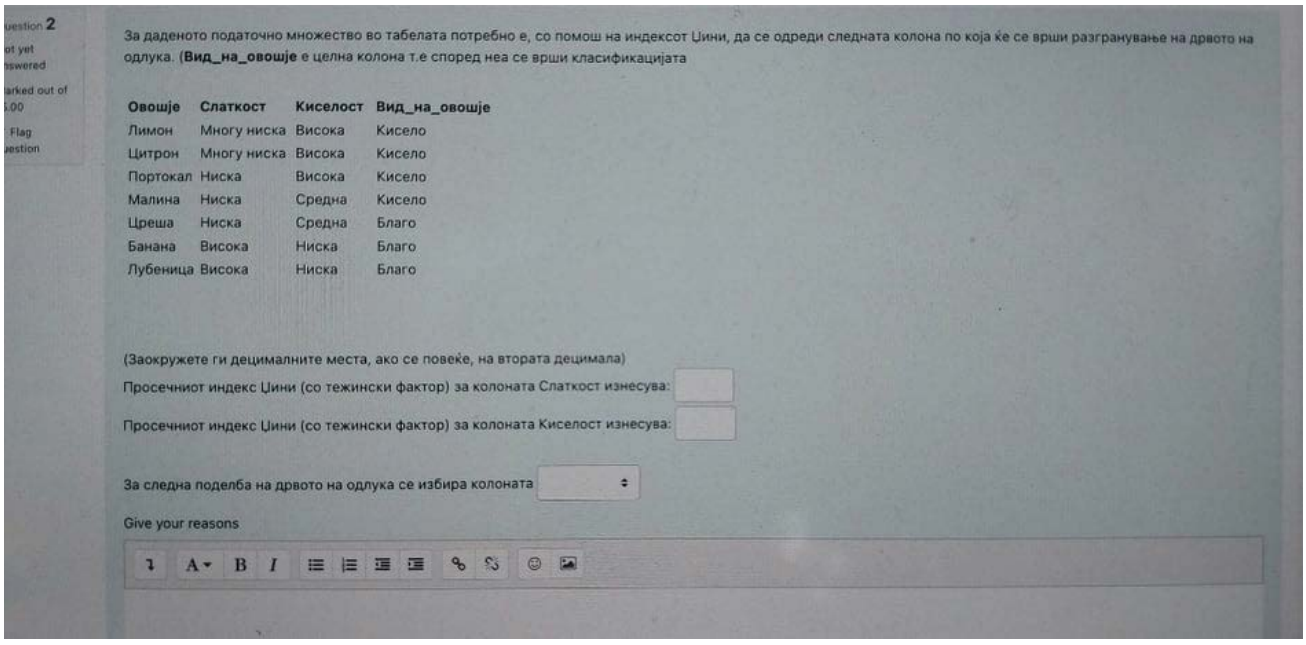
Потребно е да изберете модел/модели кои најмногу би одговарале за поставениот проблем (може да користите било која од моделите кои се изучувани во рамките на курсот). Влезните параметри, како и хипер параметрите на моделот дефинирајте ги така за да добиете најдобри предвидувања **.**

На крај треба да ја прикачите датотеката .ipynb или .py од вашето решение според дадениот урнек на Colab или .py скрипта.

*За оценување на оваа задача, покрај квалитетот на кодот ќе се зема во предвид и точноста на вашето предвидување во споредба со најдобриот резултат од колоквиумот.

**Ако сакате да ја проверите точноста на вашиот модел, потребно е да направите предвидувања на индексот за 2016 година врз основа на податоците кои се дадени во dataset3_test.csv. Во оваа датотека ги има инфорамциите за почетната вредноста на DJIA, како и соодветните вести по датум. Резултатот од предвидувањата во форма на листа треба да ги ставите во your_answer листата во CodeRunner задачата подолу и да ја извршите.

60. Слаткост: 0.1928, Киселост:0.1428. Се избира колоната киселост



Zadacha 1

Објаснете зошто Attention нивото ги подобрува резултатите кај seq2seq моделите.

Zadacha 2

Даден е dataset (dataset.csv) кој содржи податоци за пациенти. Целта е да се предвидат вредностите од колоната class, која означува дали пациентот ќе има или нема да има срцев удар - инфаркт (1/0) според дадените карактеристики на пациентите.

Поделете го dataset-от на множество за тренирање и тестирање во сооднос 70/30. Направете претпроцесирање на податоците Креирајте Boosting базиран модел со соодветно подесени хипер параметри прикладен за овој dataset. Направете евалуација на моделот со употреба на F1 Score. Урнек за задачата на Colab или .py скрипта Датотеката со кодот од решението потребно е да ја прикачите на оваа задача како notebook датотека (.ipynb) или python скрипта (.py) со име во формат Задача1_индекс.

Ако имате било каков дополнителен коментар внесете го во текстуалното поле.

```
# This is formatted as code
```

Zadacha 3

Даден е dataset (dataset.csv) кој содржи податоци за пациенти. Целта е да се предвидат вредностите од колоната class, која означува дали пациентот ќе има или нема да има срцев удар - инфаркт (1/0) според дадените карактеристики на пациентите.

Невронската мрежа треба да ги содржи следниве слоеви:

Конволуциски слој составен од 50 неврона, со големина на прозорец 10 и ист padding. Конволуциски слој со 16 неврона, со големина на прозорец 1 и валиден padding MaxPooling слој со pool size 8 и ист padding. Сите скриени слоеви имаат sigmoid активациска функција. Останатите параметри на мрежата како што се излезниот слој, излезната активациска функција, функцијата на загуба (loss function), отпимизатор, метрики за точност, како и останати параметри ако има потерба, потребно е да ги дефинирате вие.

Дополнително потребно е да се внесе некој механизам за спречување на overfitting.

Бројот на епохи на кои треба да ја тренирате невронската мрежа е 4 со големина на batch од 12.

Урнек за задачата на Colab или .py скрипта

Решената задача потребно е да ја прикачите на оваа задача како notebook датотека (.ipynb) или python скрипта (.py) со име во формат Задача2_индекс.

Ако имате било каков дополнителен коментар внесете го во текстуалното поле.

Zadacha 4

Дадено е податочното множество train_3.csv, каде за даден датум (колона date) се дадени твитови (колона text) за вакцинација против корона вирусот и бројот на вакцинирани случаи во Европа (колона total_cases)

Ваша задача е да истренирате модел со кој ќе одредува колку ќе биде бројот на вакцинирани случаеви.

Потребно е да изберете модел/модели кои најмногу би одговарале за поставениот проблем (може да користите било која од моделите кои се изучувани во рамките на курсот). Проверката на перформансите на моделот ја правите со некоја од метриките изучувани за евалуација.

Потоа истрениранот модел треба да го искористите за да добиете embedding на зборовите дадени во test_3.csv (колона word). кои е потребно да ги кластерирате со Agglomerative Clustering во 3 кластери и истите да ги визуелизирате во 3D график.

Урнек за задачата на Colab или .py скрипта

Решената задача потребно е да ја прикачите на оваа задача како notebook датотека (.ipynb) или python скрипта (.py) во формат Задача3_индекс

Ако имате било каков дополнителен коментар внесете го во текстуалното поле.

```
[ ]
```

61.

Time left 0:52:53

Question **3**

Not complete

Marked out of 15.00

Flag question

Даден е dataset со огласи за работа во кој се наоѓа описот за работната позиција (колона: description) и дали огласот е лажен или не - 1/0 (колона: fraudulent). Ваша задача е да ја предвидите fraudulent таргет колоната, каде на влез на моделот ќе бидат проследени обработените текстови од description колоната.

За таа цел потребно да имплементирате две функционалности:

1. претпроцесирање на влезните податоци. Во овој сегмент потребно е да ги процесирате текстовите така што ќе добиете вектор од една или повеќе нумерички вредности.
2. имплементација на KNN моделот со 3 најблиски соседи

Датасетите е поставен на курсот на ispiti со име "train1.csv и "test1.csv" и истите можете од таму да ги симнете!

For example:

Test	Result

`train = pd.read_csv('train1.csv')`
`test = pd.read_csv('test1.csv')`
`print(f'{build_model(train,test):.3f}')`
	0.522

Answer: (penalty regime: 0 %)

Reset answer

```
import pandas as pd
from sklearn.metrics import f1_score


def build_model(train,test):
```

48°F Partly sunny ENG 2:38 PM 12/4/2021

```
def split_data(x,y,test_size):
  X_train = x[:int(len(x)*(1-test_size))]
  X_test = x[int(len(x)*(1-test_size)):]
  Y_train = y[:int(len(y)*(1-test_size))]
  Y_test = y[int(len(y)*(1-test_size)):]
  return X_train, X_test, Y_train, Y_test

def build_model(df,test_size):
  #Handle missing values


  #Normalize the values on whole dataset input

  #Split the data with the use of train_split function


  # Create Decision Tree model
  y_pred =
  return roc_auc_score(Y_test,y_pred)

df = pd.read_csv('dataset1_2.csv')

```

Precheck Check

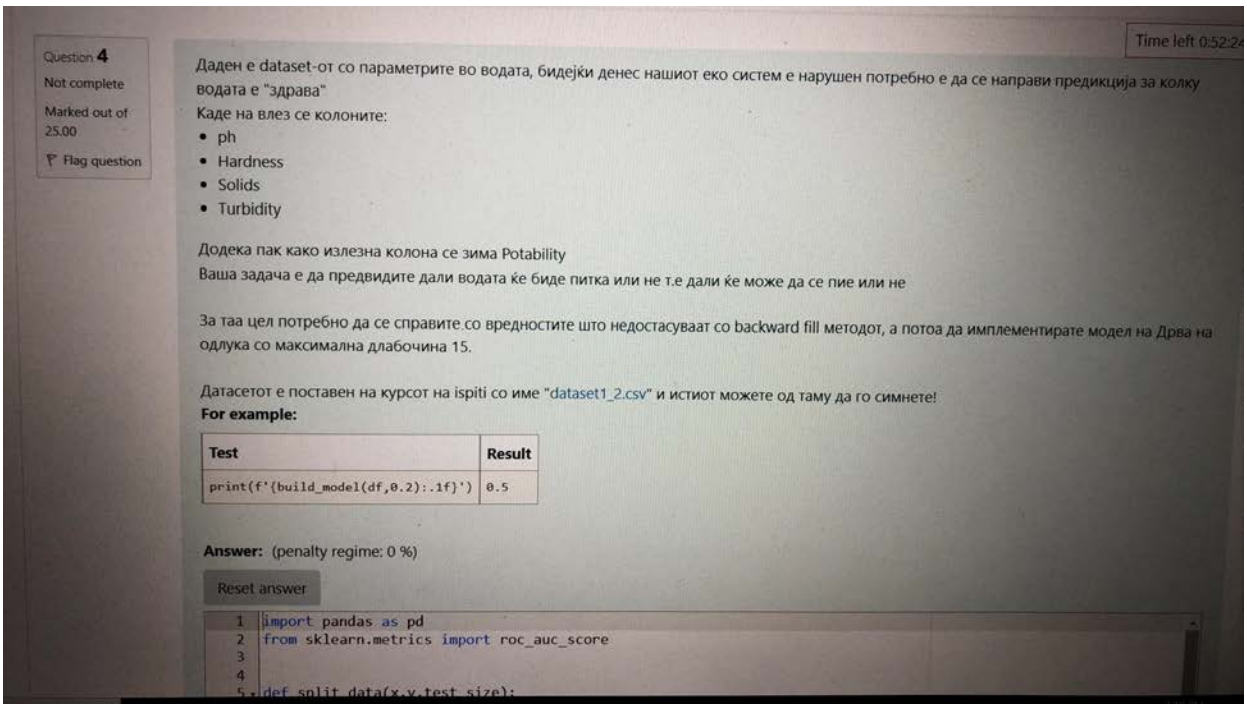
Answer: (penalty regime: 0 %)

Reset answer

```
import pandas as pd
from sklearn.metrics import roc_auc_score


def split_data(x,y,test_size):
  X_train = x[:int(len(x)*(1-test_size))]
  X_test = x[int(len(x)*(1-test_size)):]
  Y_train = y[:int(len(y)*(1-test_size))]
  Y_test = y[int(len(y)*(1-test_size)):]
  return X_train, X_test, Y_train, Y_test

def build_model(df,test_size):
  #Handle missing values


  #Normalize the values on whole dataset input

  #Split the data with the use of train_split function


  # Create Decision Tree model
  y_pred =
```

Precheck Check

62.



Question **4**

Not complete

Marked out of 25.00

Flag question

Time left 0:52:2

Даден е dataset-от со параметрите во водата, бидејќи денес нашиот еко систем е нарушен потребно е да се направи предикција за колку водата е "здрава"

Каде на влез се колоните:

- ph
- Hardness
- Solids
- Turbidity

Додека пак како излезна колона се зима Potability

Ваша задача е да предвидите дали водата ќе биде питка или не т.е дали ќе може да се пие или не

За таа цел потребно да се справите со вредностите што недостасуваат со backward fill методот, а потоа да имплементирате модел на Дрва на одлука со максимална длабочина 15.

Датасетот е поставен на курсот на ispiti со име "dataset1_2.csv" и истиот можете од таму да го симнете!

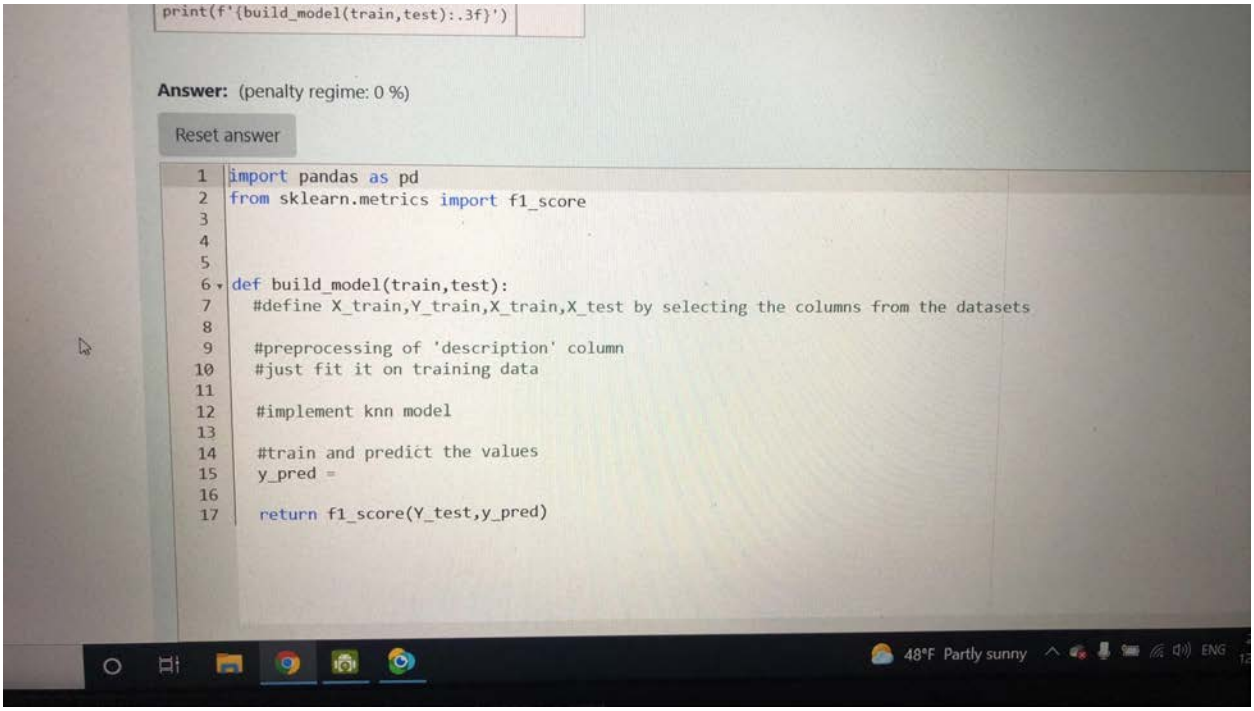
For example:

Test	Result
`print(f'{build_model(df,0.2):.1f}')`	0.5

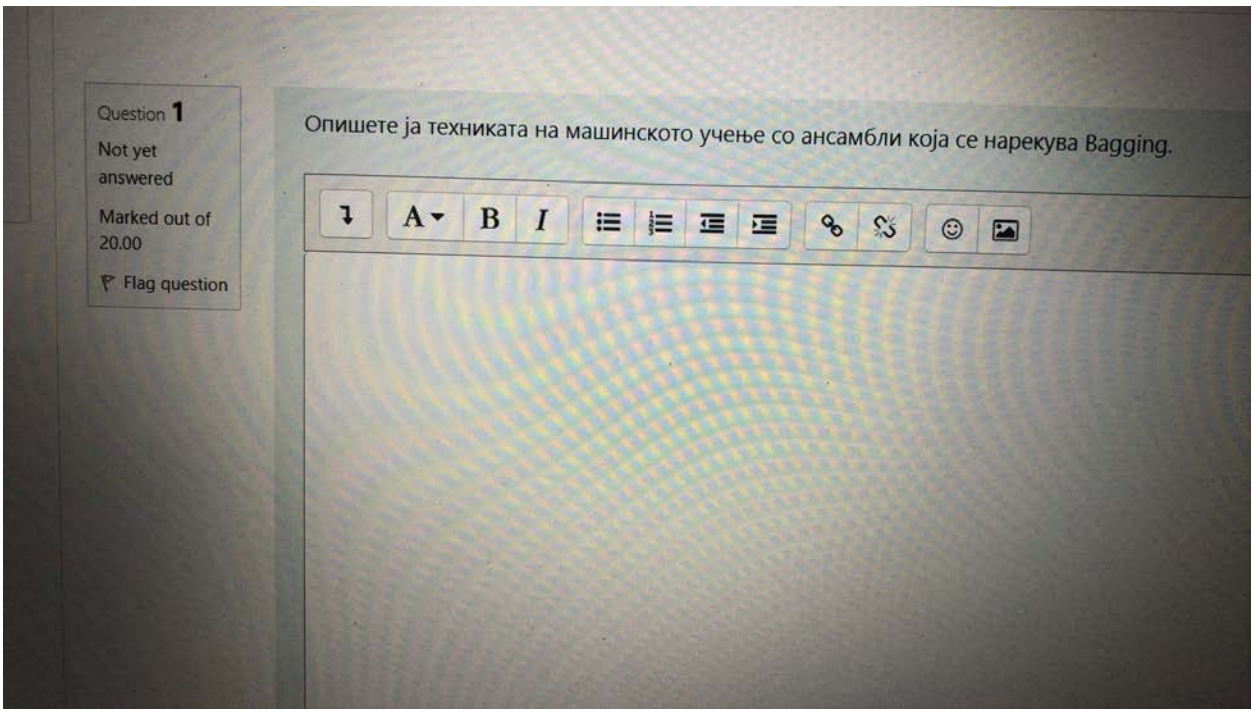
Answer: (penalty regime: 0 %)

Reset answer

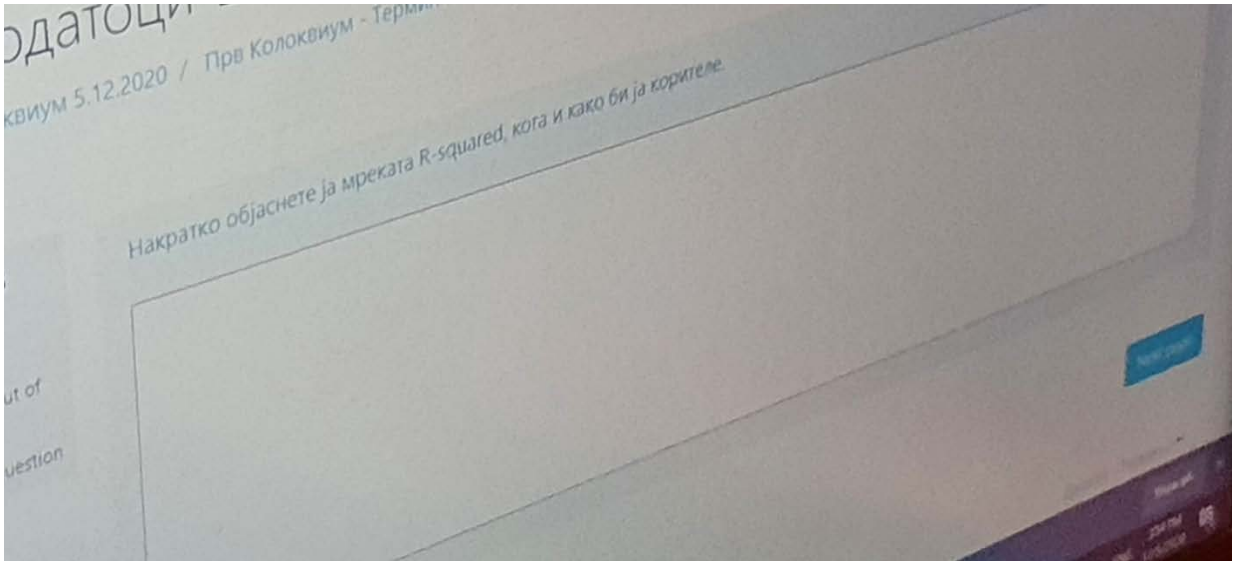
```
import pandas as pd
from sklearn.metrics import roc_auc_score


def split_data(x,y,test_size):
```


print(f'{build_model(train,test):.3f}')
Answer: (penalty regime: 0 %)
Reset answer
import pandas as pd
from sklearn.metrics import f1_score
def build_model(train,test):
#define X_train,Y_train,X_train,X_test by selecting the columns from the datasets
#preprocessing of 'description' column
#just fit it on training data
#implement knn model
#train and predict the values
y_pred =
return f1_score(Y_test,y_pred)
48°F Partly sunny
ENG

63.


Question 1
Not yet answered
Marked out of 20.00
Flag question
Опишете ја техниката на машинското учење со ансамбли која се нарекува Bagging.

64.



65.

Во оваа задача е потребно да напишете три функции:

- **encoding(data, columns)**: **data** е од тип pandas **DataFrame**, **columns** е **листа од имиња на колоните** кои се лабелирани и е потребно да се енкодираат. Целта на оваа функција е да се енкодираат колоните во дадениот датасет
- **handling_missing_values(data, column, degree)**: **data** е од тип pandas **DataFrame**, **column** е **колоната** во која недостасуваат вредности. Целта на оваа функција е да се заменат вредностите кои недостасуваат од колоната со вредностите кои ќе се предвидат со помош на **KNeighborsClassifier**, **degree** е степенот во **KNeighborsClassifier**
- **prediction(file, labeled_columns, missing_value_column, target_column, knn_degree, max_depth,n_estimators,learning_rate): Оваа е главната функција во која ги повикувате претходните две функции. Целта на оваа функција е со помош на XGBRegressor да се предвидат вредностите на таргет колоната (Y).**

1. **file:** патеката до csv фајлот
2. **labeled_columns: листа од имињата на колоните** кои се лабелирани, аргумент за во **encoding** функцијата
3. **missing_value_column: колоната** во која недостасуваат вредности, аргумент за во **handling_missing_values** функцијата
4. **target_column: Y колоната**
5. **knn_degree: степенот во KNeighborsClassifier, аргумент за во handling_missing_values функцијата**
6. **max_depth: хипер-параметар за XGBRegressor**
7. **min_child_weight: хипер-параметар за XGBRegressor**
8. **n_estimators: хипер-параметар за XGBRegressor**
9. **learning_rate: хипер-параметар за XGBRegressor**

Изгледот на датасетот кој ќе се користи за тренирање на моделот:

gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
female	group B	bachelor's degree	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's degree	standard	none	90	95	93
male	group A	associate's degree	free/reduced	none	47	57	44
male	group C	some college	standard	none	76	78	75

Целта во оваа задача е предикција на **writing_score** кај студентите.

Датасетот е поставен на курсот на ispiti со име "Датасет - Термин 2" и истиот можете од таму да го симнете!

For example:

Пр. За даден dataset, со помош на classification error да се одреди следната колона по ше се среди разгранување на дрвото на одлуки.

Fruit	sweetness	seazne's	fruit_type
Lemon	extremly low	high	sour
Grapfruit	low	medium	sour
Orange	low	medium	sour
Raspery	medium	medium	sour
Chezzy	medium	medium	sweet
Banana	high	low	sweet
Watermelon	high	low	sweet
Mandarin	extremly low	medium	none

Classification error за колоната sweetness изнесува 0,75

Classification error за колоната seazne's изнесува 0,38

Решение: $1 - MAX\left(\frac{K}{R}\right)$

Classification error за sweetness

$1 - \left(\left(\frac{2}{8}\right), \left(\frac{3}{8}\right), \left(\frac{2}{8}\right)\left(\frac{1}{8}\right)\right) = 1 - 0,25 = 0,75$

extremly low, low, medium, high

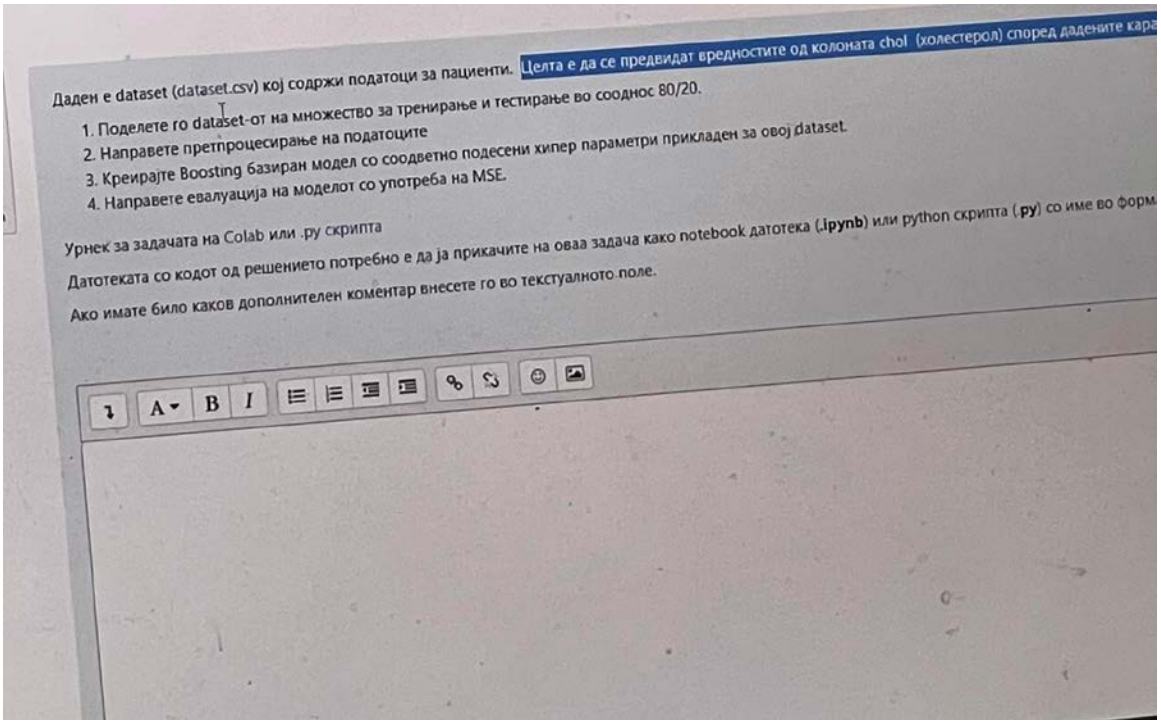
Classification error за seazne's:

$1 - \left(\left(\frac{1}{8}\right), \left(\frac{5}{8}\right)\left(\frac{1}{8}\right)\right) = 1 - 0,63 = 0,38$

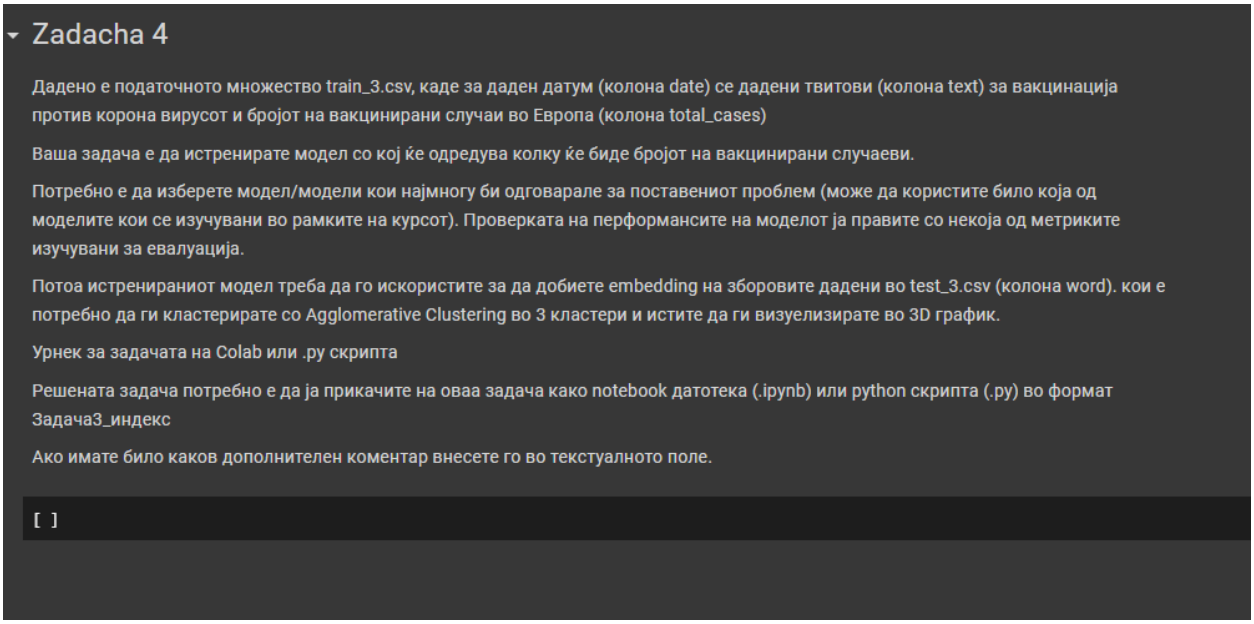
high, medium, low

(c) boolstar (d) entities+zoyet

66.


Даден е dataset (dataset.csv) кој содржи податоци за пациенти. Целта е да се предвидат вредностите од колоната chol (холестерол) според дадените кара
1. Поделете го dataset-от на множество за тренирање и тестирање во сооднос 80/20.
2. Направете претпроцесирање на податоците
3. Креирајте Boosting базиран модел со соодветно подесени хипер параметри прикладен за овој dataset.
4. Направете евалуација на моделот со употреба на MSE.
Урнек за задачата на Colab или .py скрипта
Датотеката со кодот од решението потребно е да ја прикачите на оваа задача како notebook датотека (.ipynb) или python скрипта (.py) со име во форм
Ако имате било каков дополнителен коментар внесете го во текстуалното поле.

67.


Zadacha 4
Дадено е податочното множество train_3.csv, каде за даден датум (колона date) се дадени твитови (колона text) за вакцинација против корона вирусот и бројот на вакцинирани случаи во Европа (колона total_cases)
Ваша задача е да истренирате модел со кој ќе одредува колку ќе биде бројот на вакцинирани случаеви.
Потребно е да изберете модел/модели кои најмногу би одговарале за поставениот проблем (може да користите било која од моделите кои се изучувани во рамките на курсот). Проверката на перформансите на моделот ја правите со некоја од метриките изучувани за евалуација.
Потоа истрениранот модел треба да го искористите за да добиете embedding на зборовите дадени во test_3.csv (колона word). кои е потребно да ги кластерирате со Agglomerative Clustering во 3 кластери и истите да ги визуелизирате во 3D график.
Урнек за задачата на Colab или .py скрипта
Решената задача потребно е да ја прикачите на оваа задача како notebook датотека (.ipynb) или python скрипта (.py) во формат Задача3_индекс
Ако имате било каков дополнителен коментар внесете го во текстуалното поле.
[ ]

68.

3. Да се определи колку изнесува Gini (Џини) индексот за првата редица R1 од дадената табела каде колоните ја означуваат класата, а редиците регионот.

	Class 1	Class 2	
R1	2	5	1-(2/7^2 + 5/7^2) =1-(0.082 + 0.51) = 0.408
R2	6	4	1-[(6/10)^2 + (4/10)^2] = 1-(0.36 + 0.16)=0.48

-0.168 За R1 7/17 * 0.408 = 0.168

-0.282 За R1 и R2 7/17*0.408 + 10/17*0.48 =0.45

-0.5

-0.45