

1. Најкратко објаснете ја мерката R-squared и како и кога би ја користеле.

R-squared уште се нарекува и коефициент на определување. Таа се користи така што ако моделот е добар како средна вредност  $y$ , тогаш  $R^2 = 0$ , ако моделот е перфектен  $R^2=1$  и ако  $R^2$  може да бид енегативен тоа настанува ако моделот е полош и а од проектот оваа најчесто се случува поради оценување во тест множество.

2. Која е мерка за сличност се употребува за да во KNN – класификација се предвиди даден примерок ?
  - растојание

The screenshot shows a Microsoft Photos window with a single image titled "Photos - Screenshot (264).png 106%". The image displays a digital exam interface. At the top, there's a toolbar with icons for search, delete, and other functions. Below the toolbar, the text "податоци-2020/2021/Z" is visible, followed by "Пример колоквиум". The main content area contains a question: "Која мерка за сличност се употребува за да во KNN - класификацијата се предвиди даден примерок во која класа припаѓа ?" Below the question, there's a "Select one:" label and three options:

- a. Растојание
- b. Предвидување на коефициенти
- c. Визуелизација

A "Clear my choice" link is also present. On the left side of the screen, there's a sidebar with the text "on 2" and "out of 1 question". At the bottom of the screen, there's a navigation bar with "ious page" and "Next page" buttons, along with a taskbar showing various application icons and system status.

3. Да се определи колку изнесува Gini Index за прва редица(R1) од табела колони ја означуваат класата , а редица регион.

Photos - Screenshot (265).png 87%

See all photos + Add to

БНП-20/21 / General / Пример колоквиум

**Question 3**  
Not yet answered  
Marked out of 1.00  
Flag question

Da se определи колку изнесува Чини индексот за првата редица (R1) од дадената табела каде колоните ја означуваат класата, а редиците регионот.

	Class 1	Class 2
R1	2	5
R2	6	4

Select one:

- a. 0.168
- b. 0.282
- c. 0.5
- d. 0.45

[Clear my choice](#)

[Previous page](#) [Next page](#)

- 0.42

Гини ( $i_1, j_1, t_1$ ) =  $1 - \sum p(z_i | R_i)^2$

1 -  $\left( \frac{2}{7} \right)^2 + \left( \frac{5}{7} \right)^2 = \frac{1 - \left[ \left( \frac{R_1 \text{Index}_1}{3\text{dip}} \right)^2 + \left( \frac{R_1 \text{Index}_2}{3\text{dip}} \right)^2 \right]}{3\text{dip}}$

$1 - \left[ \left( \frac{2}{7} \right)^2 + \left( \frac{5}{7} \right)^2 \right] =$

$= 1 - [(0,128)^2 + (0,171)^2] =$

$= 1 - [(0,0841) + (0,497)^2] =$

$= 1 - 0,5811 \approx 0,4189 \approx 0,42$

1 -  $\left( \frac{2}{7} \right)^2 + \left( \frac{5}{7} \right)^2$

$R_1 = ?$

1 -  $\left( \frac{2}{7} \right)^2 + \left( \frac{5}{7} \right)^2$

$2 - \text{Index}$

$2 - \text{Region}$

$2 + 5 = 7 - 3\text{dip}$

$3\text{dip} = R_1 \text{Index}_1 + R_1 \text{Index}_2$

4. Подреди ги командите по целосен редослед.

Question 4  
Not yet answered  
Marked out of 1.00  
Flag question

Подреди ги командите по редослед со цел на крај да се добие html кодот од дадената веб страница

import requests  
from bs4 import BeautifulSoup  
from IPython.display import HTML  
snapshot = requests.get('https://www.cnbc.com/finance/')

raw\_html = snapshot.text  
soup = BeautifulSoup(raw\_html, 'html.parser')

1 2 3 4 5 6

5. Со наредба Df.mean() што ќе се добие како резултат ?

Question 5  
Not yet answered  
Marked out of 1.00  
Flag question

Со наредбата df.mean() што ќе се добие како резултат ?

Select one:

a. За секоја од колоните на податочното множество df ќе се испримта средната вредност

b. Само за нумеричките колони од податочното множество df ќе се испримта средната вредност

c. Само за категориските колони од податочното множество df ќе се испримта средната вредност

[Clear my choice](#)

6. Со команда enc= OneHotEncoder(handle\_unknown='ignore')

Photos - Screenshot (271).png 96%

See all photos + Add to ... Edit & Create Share ...

↳ ВО науката за податоци-2020/2021/Z

:courses / ВНП-20/21 / General / Пример колоквиум

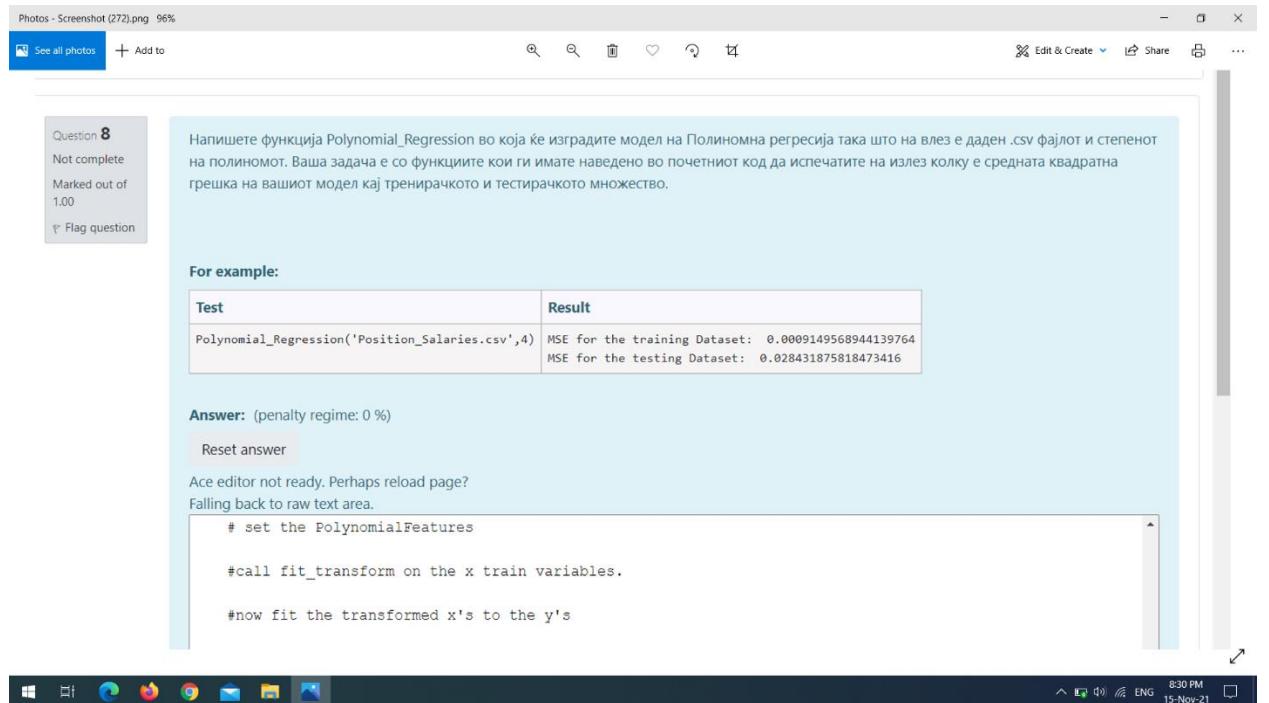
Question 7  
Not yet answered  
Marked out of 1.00  
Flag question

Со командата enc = OneHotEncoder(handle\_unknown='ignore') се  
Креира инстанца од OneHotEncoder , додека пак со команда  
enc.fit\_transform(X) се Се тренира моделот OneHotEncoder и се добива матрица за влезната колона X .

Previous page

◀ Анкета за присуство на првиот колоквиум Jump to... 01 Воведно предавање

7. Напишете функција `Polynomial_Regression` во која ќе изградите модел на Полиномна регресија.



```

import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error

import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split

def Polynomial_Regression(file,degree):

    data = pd.read_csv(file)
    X = data.iloc[:,1:2].values
    Y = data.iloc[:,2:3].values

```

```

X_train = X[0:int(len(X)*0.8)]
X_test = X[int(len(X)*0.8):]
Y_train = Y[:int(len(Y)*0.8)]
Y_test = Y[int(len(Y)*0.8):]

#use StandardScaler for standardization the data

from sklearn.preprocessing import StandardScaler

scaler=StandardScaler()

X_train=scaler.fit_transform(X_train)

X_test=scaler.fit_transform(X_test)

Y_train=scaler.fit_transform(Y_train)

Y_test=scaler.fit_transform(Y_test)

# set the PolynomialFeatures

from sklearn.linear_model import LinearRegression

from sklearn.preprocessing import PolynomialFeatures

from sklearn.metrics import

mean_squared_error poly_reg = PolynomialFeatures(degree)

#call fit_transform on the x train variables.

X_poly = poly_reg.fit_transform(X_train)

#now fit the transformed x's to the y's

poly_reg.fit(X_poly, Y_train)

#create an object of the class

LinearRegression lin_reg = LinearRegression()

#fit the LinearRegression model to our transformed

X lin_reg.fit(X_poly, Y_train)

#call fit_transform on the x test variables.

X_test_poly = poly_reg.fit_transform(X_test)

#make prediction for the testing data

```

```

y_pred=lin_reg.predict(X_test_poly)

training_score=mean_squared_error(Y_train,lin_reg.predict(X_poly))

#calculate the mean_squared_error for the training and testing data

test_score=mean_squared_error(Y_test,y_pred)

#print the scores

print("MSE for the training Dataset: " + str(training_score))

print("MSE for the testing Dataset: " + str(test_score))

```

8. За дадениот датасет во табела епотребно со помош на Classification Error да се одреди следна колона по која ќе се изврши разгранување на дрво на одлука. Fruit\_type

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 одлука. (Fruit\_type е таргет колона т.е по неа се врши класификацијата)

Fruit	Sweetness	Sourness	Fruit_type
Lemon	Extremely Low	High	Sour
Grapfruite	Low	Medium	Sour
Orange	Low	Medium	Sour
Rasberry	Medium	Medium	Sour
Cherry	Medium	Medium	Sweet
Banana	High	Low	Sweet
Watermelon	High	Low	Sweet
Mandarin	Extremely Low	Medium	None

(Закружи ги децималните места ако се повеќе на втората децимала)

Classification Error за колоната Sweetness изнесува:  ×

Classification Error за колоната Sourness изнесува:  ×

За следна поделба на дрвото на одлука се избира колоната  ×

Give your reasons

Задаваме вредност за секој елемент во датасетот. Extremely Low=0, Low=1, Medium=2, High=3. Сега пресметуваме Classification Error за секое овошје според неговиот тип.

Lemon:  $1 - \max(0/3, 3/3) = 1 - 1 = 0$ ; Grapfruite =  $1 - \max(1/3, 2/3) = 1 - 2/3 = 0.33$ ; истата пресметка е за Orange; Rasberry =  $1 - \max(2/4, 2/4) = 1 - 0.5 = 0.5$ . Classification Error за колоната Sourness изнесува  $\min(3/13 * 0 + 3/13 * 0.33 + 3/13 * 0.33 + 4/13 * 0.5) = 0.31$

Cherry :  $1 - \max(2/4, 2/4) = 1 - 0.5 = 0.5$ ; Banana =  $1 - \max(3/4, 1/4) = 1 - 0.75 = 0.25$ ; истата пресметка е

Sweetness = 0.25

Sourness = 0.25

За следна поделба се избира колон која има помал error.

	Sour	None	Sweet
Ext. low	1	1	0
low	2	0	0
Medium	1	0	1
High	0	0	2

W

8 - largest spot  
are low, medium,  
high

Error (Ext. low) =  $1 - \left( \frac{1}{2} \right) = 0,5$

Error (low) =  $1 - \left( \frac{2}{2} \right) = 0,0$

Error (Medium) =  $1 - \left( \frac{1}{2} \right) = 0,5$

Error (High) =  $1 - \left( \frac{1}{2} \right) = 0,0$

Sweetness

$\frac{2}{8} \cdot 0,5 + 0 + \frac{2}{8} \cdot 0,5 + 0 =$   
 $= 0,125 + 0,125 = \underline{\underline{0,25}}$

Sourness error too is worth  
so we're going to  
do this one

	Sour	None	Sweet
Ext. Low	0	0	0
Low	0	0	2
Medium	3	1	1
High	1	0	0

	Sour	None	Sweet
Ext. Low	0	0	0
Low	0	0	2
Medium	3	1	1
High	1	0	0

Sourness  $\Rightarrow \frac{0}{8} \cdot 1 + 0 + \frac{5}{8} \cdot 0,4 + 0 =$   
 $\Rightarrow 0 + 0 + 0,625 \cdot 0,4 + 0 =$   
 $\Rightarrow 0,625 \cdot 0,4 = \underline{\underline{0,25}}$

Error (Ext. low) =  $1 - \max(0, 0, 0) = 1 - 0 = 1$

Error (Low) =  $1 - \left( \frac{0}{2} + \frac{0}{2} + \frac{2}{2} \right) = 1 - 1 = 0$

Error (Medium) =  $1 - \left( \frac{0}{5} + \frac{1}{5} + \frac{1}{5} \right) = 1 - 0,6 = 0,4$

Error (High) =  $1 - \frac{1}{2} = 1 - 1 = 0$

9. Кога дистрибуција на податоците е како на сликата т.е нормална што се случува со средна вредност и медијана 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. Медијаната е поголема од средната вредност.
- b. Средната вредност е поголема од медијаната. X
- c. Средната вредност и медијаната се еднакви.

Give your reasons:

Средната вредност во ваков случај е поголема од медијаната. Пример ако ги имаме следните податоци: [1,2,2,3,3,3,4,4,4,4,4,4,5,5,5,6,6,7], просекот на овие податоци изнесува  $4.15$ , а медијата изнесува  $(4+4)/2 = 4$  што покажува дека средната вредност е поголема од медијаната.

Точен одговор с, средната вредност и медијана се еднакви.

10. Кои од наведените карактеристики е најдобро да се изберат ако податочно множество се состои од континуирани податоци ?

Question 1

Not yet  
answered

Marked out of  
1.00

Flag question

Кои од наведените дескриптивни статистики е најдобро да се изберат ако податочното множество се состои од континуирани податоци ?

Select one or more:

- a. Фреквенција
- b. Медиана
- c. Процент (редица, колона или вкупно)
- d. Средна вредност

- Медијана/средна вредност

11. Со која од дадените алгоритми за кластирање може да се добијат шест кластери

Question 1  
Not yet answered  
Marked out of 1.00  
Flag question

Со кој од дадените алгоритми за кластерирање може да се добијат шест кластери како што се прикажани на сликата?



a. K-means  
 b. Hierarchical Clustering  
 c. DBSCAN  
 d. K-means++

12. Која е промената кај нумеричките податоци при нормализација на истите ?

Question 6  
Not yet answered  
Marked out of 1.00  
Flag question

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](#)

[Previous page](#)

- a. Вредностите ќе бидат во опсегот помеѓу 0 и 1

13. За дадениот датасет во табела епотребно со помош на Classification Error да се одреди следна колона по која ќе се изврши разгранување на дрво на одлука.

Question 4  
not yet  
answered  
marked out of  
5.00  
Flag question

За дадениот датасет во табелата потребно е со помош на Classification Error да се одреди најдобриот начин за поделба на дрвото на одлука (Defaulted Borrower е таргет колона т.е по неа се врши класификацијата)

	Id	Marital Status	Annual Income	Defaulted Borrower
1	Single	High	No	
2	Married	Low	No	
3	Divorced	Low	Yes	
4	Married	Medium	Yes	
5	Divorced	High	No	
6	Single	Low	No	
7	Divorced	Medium	No	
8	Divorced	High	Yes	

(Задорожжи ги десичалните места ако се повеќе на втората десичала)

Classification Error за колоната Marital Status јанесува:

Classification Error за колоната Annual Income јанесува:

За следна поделба на дрвото на одлука се избира колоната

Give your reasons

Classification Error за колона Marital Status e : 0.375

Classification Error за колона Annual Income e : (0.24875) = 0.25

За следна поделба на дрво на одлука се избира колоната : Annual Income

#### 14. Кои се предности на Двонасочните LSTM мрежи (Bi-directional LSTM) ?

Question 8  
not yet  
answered  
marked out of  
1  
Flag question

Кои се предностите на Двонасочните LSTM мрежи (Bi-directional LSTMs)?

Select one or more

a. Обично се подобри од еднонасочните рекурентни и LSTM мрежи.

b. Побрзи се при обучувањето.

c. Го зимаат предвид поширокото значење на контекстот.

d. Не бараат пристап до сите податоци однапред.

Previous page

- A i c ( za c ne sum sigurna)

15. Каква дводимензионалност треба да е влезното тренирачко множество кај LSTM невронска мрежа ?

Каква димензионалност треба да е влезното тренирачко множество кај LSTM невронската мрежа?

- a. 2D - матрица
- b. 1D
- c. 3D

Pod c: 3D

16. Каков вид на учење се реализира кај Автоенкодерите ?

Каков вид на учење се реализира кај Автоенкодерите?

Select one or more:

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

- Moze b i c.

17. Што претставува инерција (momentum) при оптимизација на невронските мрежи ?

Question 7  
Not yet answered  
Marked out of 1.00  
Flag question

Што претставува инерција (momentum) при оптимизацијата на невронските мрежи?

Select one:

- a. Метод со кој оптимизацијата на тежините обезбедува глобален оптимум.
- b. Параметар со кој се одредува моменталната активација на невроните.
- c. Параметар со кој се влијае врз брзината на невронските мрежи.
- d. Дека е многу тешко да се обучи невронската мрежа

A

18. Кaj обработка на податоци се скрываат следниве задачи:

Question 4  
Answer saved  
Marked out of 1.00  
Flag question

Кaj Обработката на природни јазици се скриваат следниве задачи:

Select one or more:

- a. Категоризација на теми
- b. Препознавање на векторски претстави на зборовите (word embeddings)
- c. Извлекување на контекстни зборови (skip-grams)
- d. Препознавање на именувани нешта

B i C

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

Question 8  
Not yet answered  
Marked out of 1.00  
Flag question

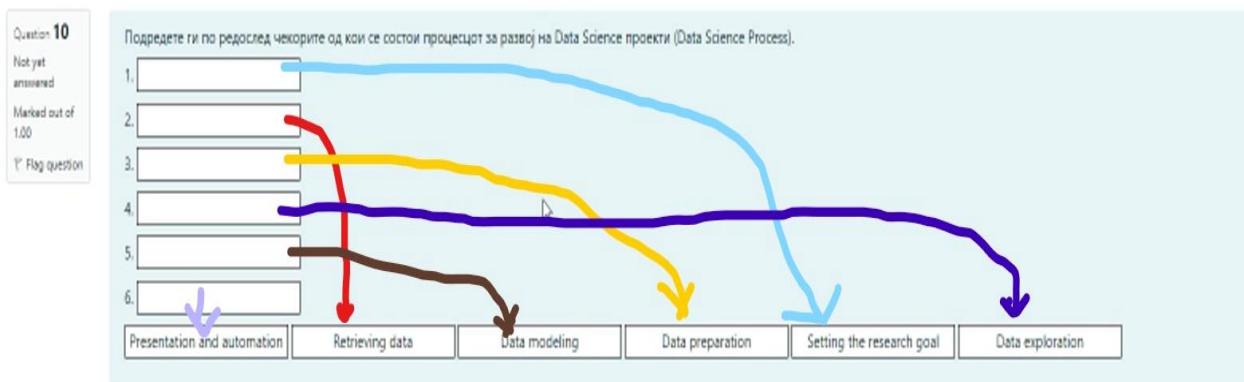
Кои од визуелизациите е најдобро да се изберат кога станува збор за датасет од категориски податоци ?

Select one or more:

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

Previous page

20. Подредете ги по редослед чекорите од кои се состои процесот за развој на Data Science Process.



Setting the research goal

Retrieving Data

Data preparation

Data exploration

Data modeling

Presentation and automation

21. За дадениот модел , кој параметар треба да се додаде како аргумент за да се користи ентропија како метрика за поделба на дрво за одлука

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.

b

a. metric = "entropy"  
 b. criterion="entropy"  
 c. splitter = "entropy"

Clear my choice

b e točno

22. 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 слови

b e točno

23. Во кој случај би било најдобро да се употреби Sigmoid како излезно ниво кај невронски мрежи.

Question 9  
Not yet answered  
Marked out of 1.00  
Flag question

Во кој случај би било најдобро да се употреби Sigmoid како излесно ниво кај невронските мрежи.

- a. Кога влезните во мрежата се дискретни предности
- b. Кога кое мрека за пресметка на загуба во мрежата се користи MSE (Mean Squared Error)
- c. Кога бројот на влезови е поголем од бројот на излези во невронската мрежа
- d. Кога сакаме да добиеме по број процесирани на резултатите на GPU
- e. Кога имаме бинарна класификација

24. Кои мерки можеме да ги користиме за сличност помеѓу два кластера ?

Question 10  
Not yet answered  
Marked out of 1.00  
Flag question

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.

25. Skip-gram со големина на прозорец три на зборот day е :

Question 6  
Not yet answered  
Marked out of 1.00  
Flag question

Time

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. a bright cold
- b. in April, and
- c. was bright cold April clocks were
- d. a bright cold in April and

Д е точно

26. Кое од наведените параметри се дел од хиперпараметрите за тренирање на XGBoost модел ?

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

a. n\_estimators i d. max\_Depth

27. Што резултат враќа дадениот код: df.isnull()

Question 7

Not yet  
answered

Marked out of  
1.00

Flag question

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

Previous page

28. Која активацијска функција е претставена на графикот ?

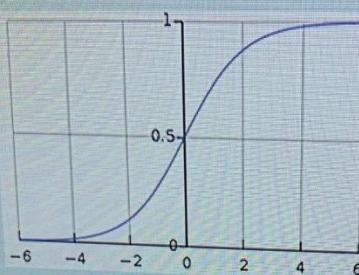
Question 9

Not yet  
answered

Marked out of  
1.00

Flag question

Која активацијска функција е претставена на графикот?



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

C sigmoid

29. Што е точно за моделот seq2seq ?

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.

Tocni se : a –sigurno , za c ne sum sigurna

30. Кои карактеристики треба да ги има активацијска функција кај невронските мрежи ?

Question 2  
Not yet answered  
Marked out of 1.00  
Flag question

Кои карактеристики треба да ги има активацијската функција кај невронските мрежи?

Select one or more:

- a. Да има некаква нелинеарност.
- b. Да овозможи градиентите да останат доволно големи и преку неколку скриени слоја.
- c. Да дава активација само за позитивни влезови.
- d. Да е заоблена.

Tocni se : a i b

31. Кои од наведените карактеристики се новитети кај Трансформер моделите ?

Кои од наведените карактеристики се новитети кај Трансформер моделите?

- a. Positional embeddings
- b. Self Attention layer
- c. Feedforward Network
- d. Tokenization

Tocni se : a i b

32. Кај случајните шуми кои хипер-параметри можат да се нагодуваат

**Question 11**

Not yet  
answered  
Marked out of  
1.00

Flag question

Кај Случайните шуми, кои хипер-параметри можат да се најодуваат

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

Previous page

Tocni se : a i b

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

**Question 5**

Not yet  
answered  
Marked out of  
1.00

Flag question

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

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

Што ќе биде излозот за дадениот код:

`model.coef_`

- a. Три коефициенти (десимални вредности) за секој од влезните податоци
- b. Еден коефициент (десимална вредност) за сите влезни податоци
- c. Четири коефициенти (десимални вредности) за секој од влезните податоци плюс интерцептот

34. Еден од најдобрите јазични модели BERT се потпира на трансформер архитектура. Кој дел од трансформер архитектура се користи во BERT ?

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

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

C. ENCODER

35. Што претставува поимот отфрање (dropout) во контекст на невронски мрежи ?

Што претставува поимот отфрање (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 отфрање.

C ili d - ne sum sigurna

36. Кои од следниве репозиториуми/библиотеки се користат за едноставно споделување на претренирани NLP модели ?

The screenshot shows a question from an online exam. The header indicates it's February Exam / T1 - first part (16.02.2022). The question number is 4, and it's marked as not yet answered. It's worth 1.00 point and has a flag option. The question text asks which of the following repositories/libraries are used for easily sharing pre-trained NLP models. The answer options are: a. HuggingFace Transformers library, b. PyTorch Hub, c. GitHub, and d. TensorFlow-Hub. The user has selected option b.

Февруари Испит / Т1 - прв дел (16.02.2022)

Question 4  
Not yet  
answered  
Marked out of  
1.00  
Flag question

Кои од следниве репозиториуми/библиотеки се користат за едноставно споделување на претренирани NLP модели.

Select one or more:

- a. HuggingFace Transformers library
- b. PyTorch Hub
- c. GitHub
- d. TensorFlow-Hub

- b. PyTorch Hub
- d. TensorFlow – Hub

37. На прикажана слика е дадена шема за кој вид на учење со ансамбли ?

Февруари Испит / Т1 - прв дел (16.02.2022)

Question 3  
Not yet answered  
Marked out of 1.00  
Flag question

На прикажаната слика е дадена шема за кој вид на учење со ансамбли?

Step 1: Create Multiple Data Sets  
Step 2: Build Multiple Classifiers  
Step 3: Combine Classifiers

Select one:

- a. Bagging
- b. Boosting

Tocno e : pod a Bagging

38. Кое од наведените можат да се користат како критериуми за прекин на понатамошното делење на јазли кај дрвата за одлучување (Shopping Conditions) ?

Question 5  
Not yet answered  
Marked out of 1.00  
Flag question  
Time left 0:13:1

Кои од наведените можат да се користат како критериуми за прекин на понатамошното делење на јазлите кај дрвата за одлучување (Stopping Conditions)?

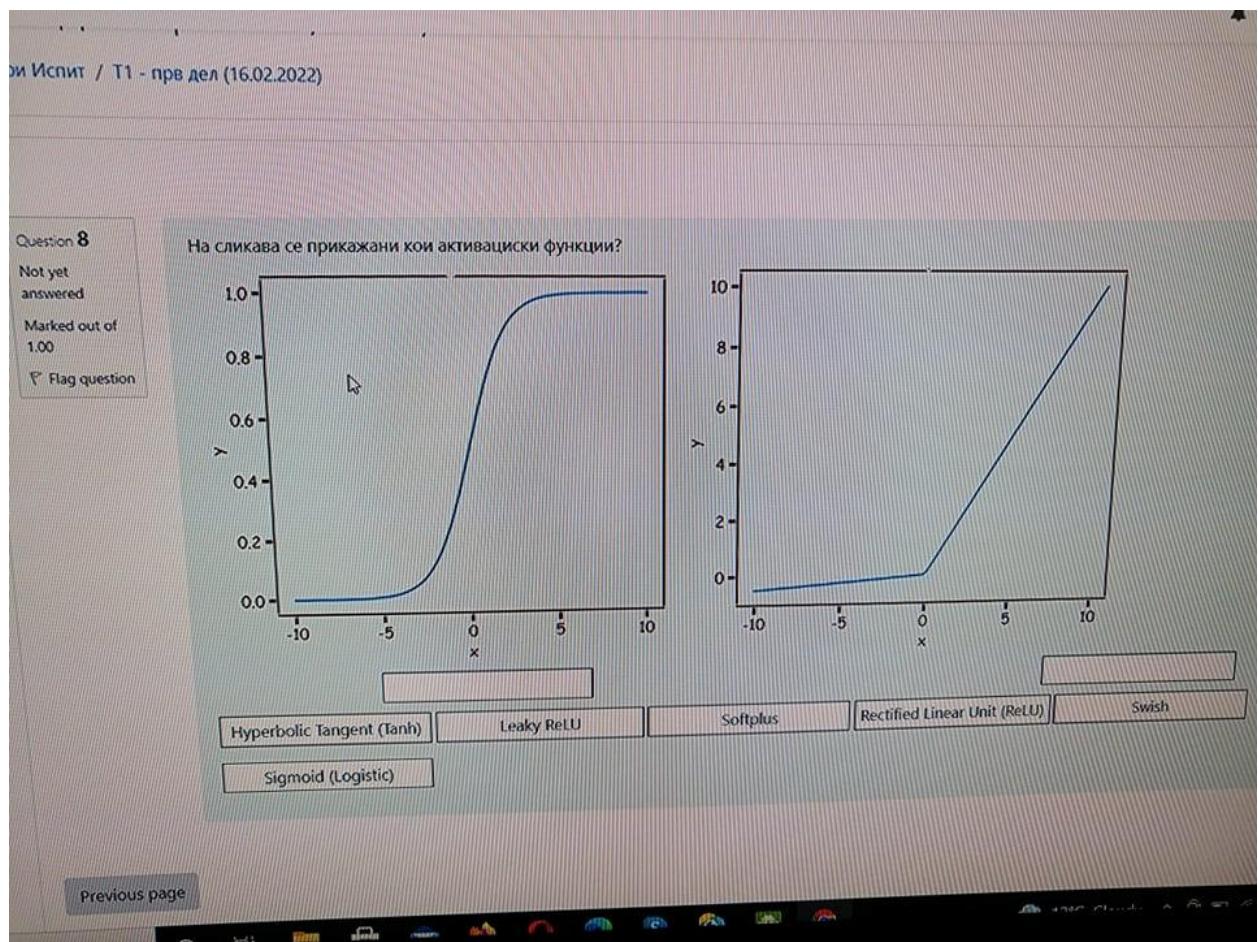
Select one or more:

- a. Ако бројот на примероци што припаѓаат на дадена класа го надмине дозволениот број.
- b. Ако сите примероци во јазелот припаѓаат на истата класа
- c. Ако бројот на циклуси надмине даден праг
- d. Ако бројот на примероци во под-јазлите спадне под даден праг (`min_samples_leaf`)
- e. Ако бројот на јазлите во дрвото надмине даден праг.

Previous page

Za d i b sum sigurna

39. На сликава се прикажани кои активациски функции ?



1 слика – Sigmoid (Logistic)

2 слика – Leaky ReLU

40. Кај случајните шуми кои хипер-параметри можат да се нагодуваат ?

Question 3  
Not yet answered  
Marked out of 1.00  
Flag question

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.

41. За даден 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

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**

<b>Id</b>	<b>Debt</b>	<b>Annual Income</b>	<b>Defaulted Borrower</b>
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)

<b>Id</b>	<b>Defaulted Borrower</b>	<b>Distance</b>
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 најфrekventен.

42. За даден модел е дадена информацијата дека има "accuracy" од 95%. Опишете ги потребните преуслови за да можете да кажете дека моделот не е добар.

Question 3  
Complete  
Mark 15.00 out of 15.00  
Flag question  
Edit question

За даден модел е дадена информацијата дека има "accuracy" од 95%. Опишете ги потребните предуслови за да можете да кажете дека моделот **не** е добар.

Еден предуслов за ова е ако датасетот е небалансиран. На пример ако се работи за нумерички податоци, ако голем дел од податоците се позитивни, моделот при секое тестирање ќе заклучува дека "classifiers" ќе бидат позитивни и со тоа не може да се заклучи дали навистина е добар моделот. Исто така во accuracy не спаѓаат false positive примероците(тие кои биле точни а сме рекле дека се грешни), а тие податоци што фалат можат да значат многу за моделот.

Comment:

Макар комплетно некомпетентно

43. Што се случува кога ентропијата за одредена колона кај даден датасет тежне кон нула ?

Question 1  
Correct  
Mark 5.00 out of  
0.0  
Flag question  
Edit  
Question

Што се случува кога ентропијата за одредена колона кај даден датасет тежне кон нула ?

Select one:

- a. Податоците се добро поделени.
- b. Податоците се несредени т.е немаат добра поделба.

Give your reasons

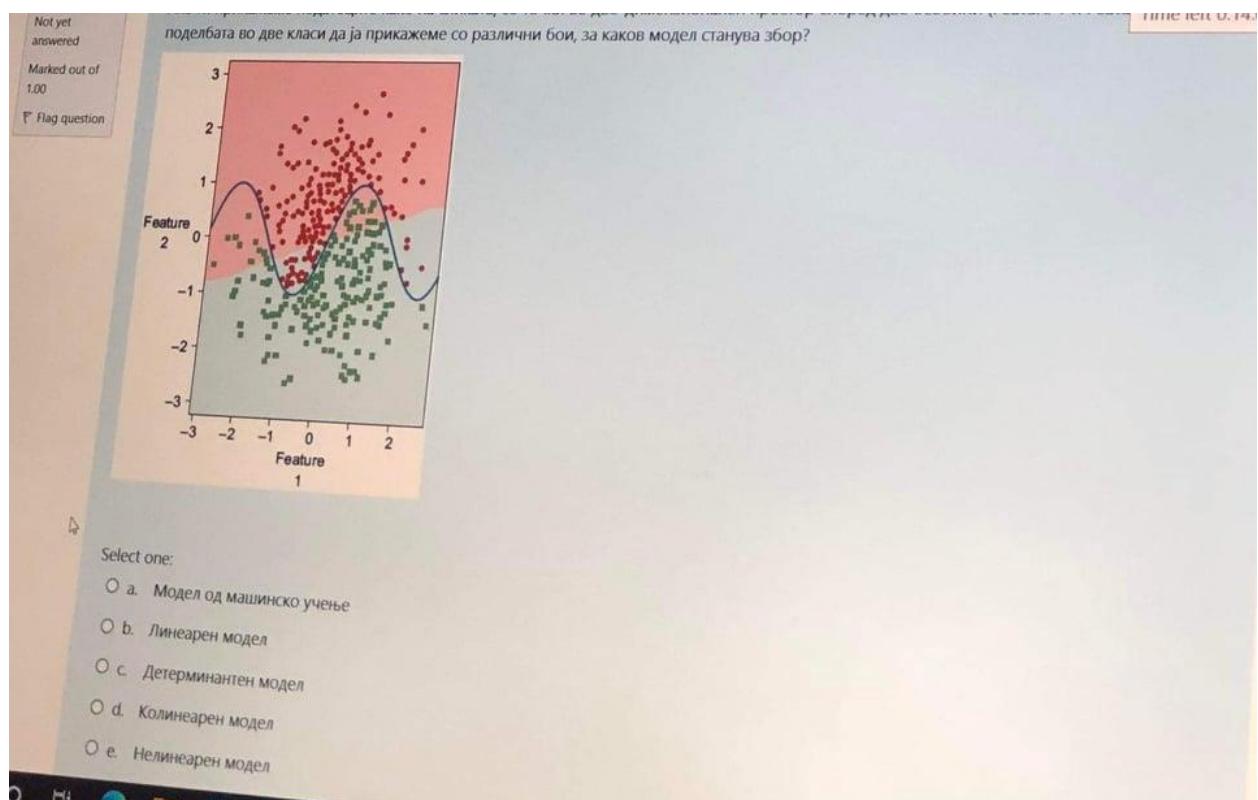
Кога ентропијата тежне кон нула, податоците се полесни за претпоставка и имаат поистакнати "пикови".

Your answer is correct.

The correct answer is: Податоците се добро поделени.

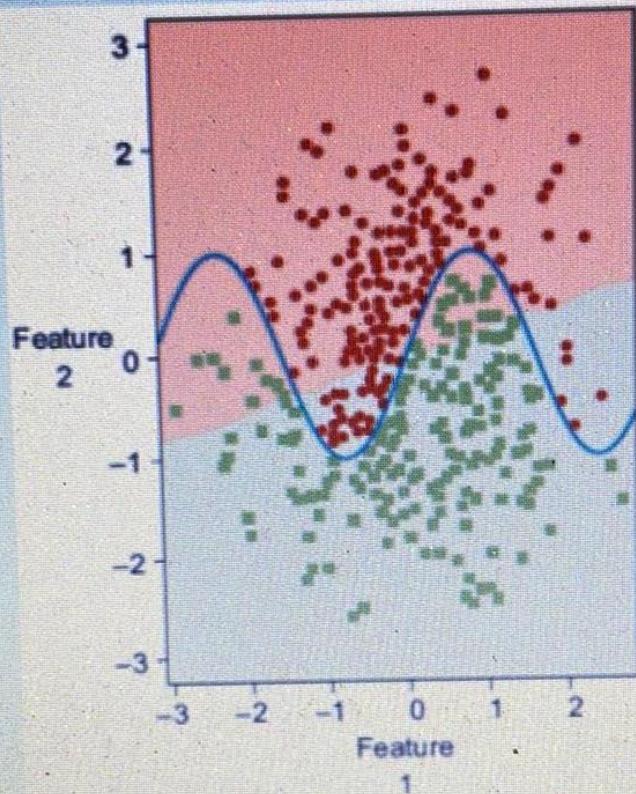
44. Ако ги прикажеме податоците, за кој модел станува збор ?

Tosno e pod b: Линеарен модел



Not yet  
answered  
Marked out of  
0.00  
Flag question

Ако ги прикажеме податоците како на сликата, со точки во дво-различни бои, за каков модел станува збор?

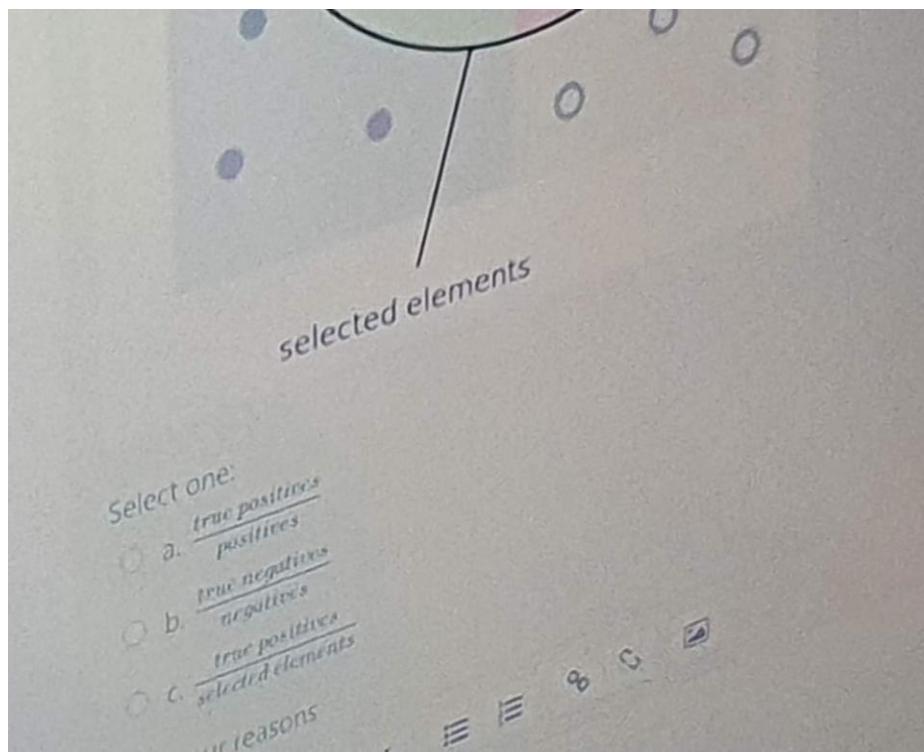
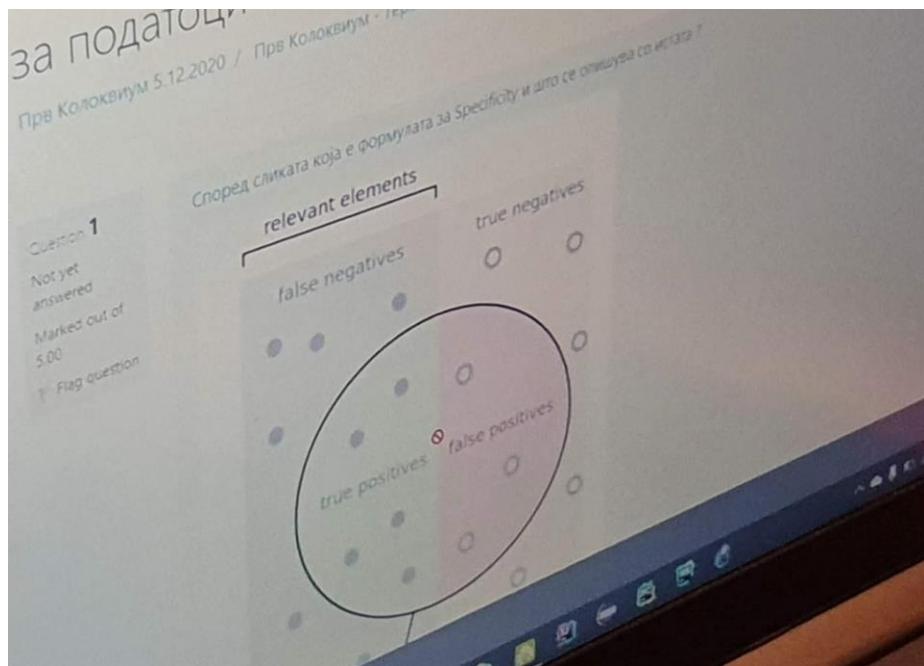


Select one:

- a. Детерминантен модел
- b. Линеарен модел
- c. Нелинеарен модел
- d. Модел од машинско учење
- e. Колинеарен модел

Точно е под b: Линеарен модел

45. Според слика која е формулата за Specificity и што се описува со истата ?



C. True positives / selected elements

Описува колку од негативните селектирани елементи се навистина негативни .

46. За дадена табела во прилог ако се енкодира колоната Class со помош на One-Hot encoding, како ќе изгледа ново добиената табела ?

10 Time left: 0:1

За дадената табела во прилог ако се енкодира колоната **Class** со помош на One-Hot encoding, како ќе изгледа ново добиената табела?

Select one:

a. **Id Company\_name Class**

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

b. **Id Company\_name Class\_Drink Class\_Clothes Class\_Technology Class\_Food**

Id	Company_name	Class	Drink	Clothes	Technology	Food
1	Pepsi	1	1	0	0	0
2	Zara	2	0	1	0	0
3	Coca - Cola	3	1	0	0	0
4	Apple	4	0	0	1	0
5	Mcdonald's	5	0	0	0	1

c. **Id Company\_name Class\_1 Class\_2 Class\_3**

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

Time left

Finish

ТОСНО Е: **под б** - One – Hot encoding – се конвертира секоја од вредностите во нова колона и се додава 1 или 0 ( дали е во колона).

Ако е дадено во барањето Binary Encoding – Тогаш значи дека го намлува бројот на колоните и решнието би било под С.

47. За даден датасет во табелата потребно е со помош на KNN класификација со  $k=3$ , да се предвиди во која класа ќе припаѓа ноциот тест примерок со id 5

Question 5  
Correct  
Mark 15.00 out of 15.00  
Flag question  
Edit question

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 заокругли 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 класифицран како

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.

48. Кaj Наивните Баесови класификатори, за атрибути  $A_i$  за дадена класа  $C$  може да претпоставиме:

Time left: 0:18:47

Question 5  
Not yet answered  
Marked out of 1.00  
Flag question

Кaj Наивните Баесови класификатори, за атрибутите  $A_i$  за дадена класа  $C$  може да претпоставиме:

- a. Условна зависност меѓу атрибутите, за таа класа
- b.  $P(A_1, A_2, \dots, A_n | C) = P(A_1 | C)P(A_2 | C) \dots P(A_n | C)$
- c. Условна независност меѓу атрибутите, за таа класа
- d.  $P(A_1, A_2, \dots, A_n | C) = P(A_1 | C) + P(A_2 | C) + \dots + P(A_n | C)$

Previous page Next page

b (Ne sum sigurna )

49. Кои од наведените кодови ќе ги избрише сите редици кои имаат NAN вредности ?

Time left: 0:18:25

Question 8  
Not yet answered  
Marked out of 1.00  
Flag question

Кои од наведените кодови ќе ги избрише сите редици кои имаат NAN вредности?

- a. df.deletena()
- b. df.dropna(axis = 0)
- c. df.delete
- d. df.dropna(axis = 1)
- e. df.drop()

Previous page

Tocen odgovor : pod b

50. Што печати следниот код:

Pod e

Time left: 0:18:00

Question 1  
Not yet answered  
Marked out of 1.00  
Flag question

Што печати дадениот код:

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

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

- a. Ниту едно од наведените
- b. Ја симнува веб страницата
- c. Ја враќа и прикачува веб страницата
- d. Ги принта сите сликовите од веб страницата "site.mk".
- e. Ги принта сите извори на сликовите 'img' под 'src' тагот на веб страницата "site.mk".

Previous page Next page

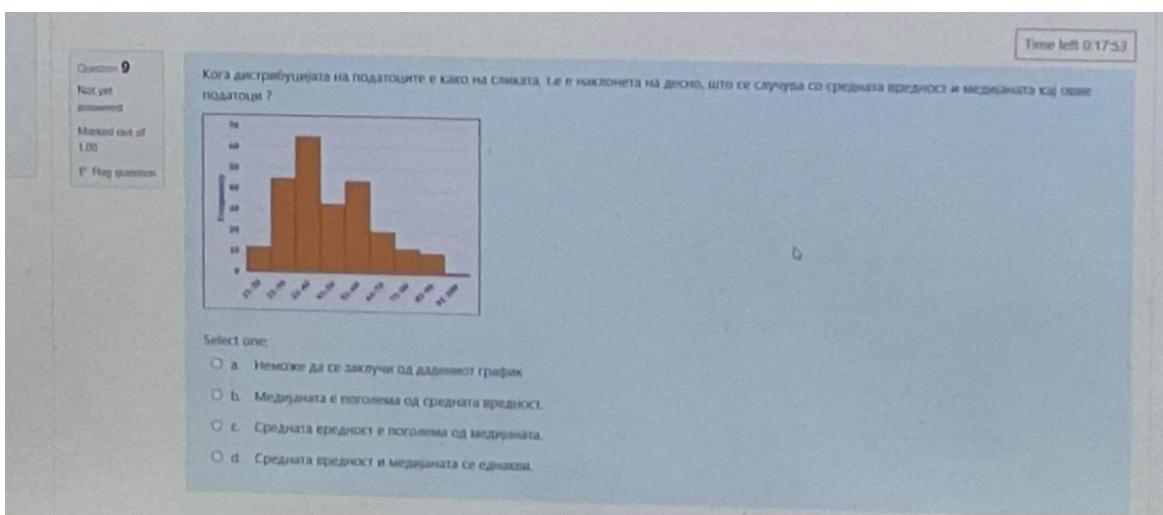
Што печати дадениот код:

```
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 претвора сите слики од веб страницата "sitel.mk"
- b. Ја враќа и прикажува веб страницата
- c. Ги претвора сите извори на слики 'img' под 'src' тагот на веб страницата "sitel.mk"
- d. Ја симнува веб страницата
- e. Ниту едно од наведените

[Clear my choice](#)

51. Кога дистрибуција на податоци е како на слика т.е. наклонетоста на десно, што се случува со средна вредност и медијаната кај овие податоци ?



Д. Средна вредност и медиана се еднакви

52. Нека се дадени следниве променливи X и Y

The screenshot shows a question titled 'Question 7' with a note 'Not yet answered' and a mark of 'Marked out of 1.00'. A 'Flag question' button is present. The question text reads: 'Нека се дадени следниве променлив X и Y.' Below the text is a table:

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. 3
- b. 4
- c. Неможе да се определи
- d. 2

Не може да се определи

X	Y
1	3
2	3
<b>Null</b>	4
4	3
1	2

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

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

53. Ако треба да се идентификува кога ќе се случи некоја необична трансакција при плаќањето за каков вид на машинско учење станува збор ?

Time left: 0:19:32

Question 2  
Not yet answered  
Marked out of 1.00  
Flag question

Ako треба да се идентификува кога ќе се случи некоја необична трансакција при плаќањето, за каков вид на машинско учење станува збор?

Select one:

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

Previous page

Announcements

Jump to...

Next page

You are logged in as Студентски Задато Јовановски

Csvora 1 ►

- Регресија , Откривање на недостатоци , Класификација, Учење со потикнување (не знам кој е точен), мислам дека е Регресија

54. Даден е модел на KNN класификација за предвидување дали куќата ќе се продаде или не 0 или 1 соодветно(class колона), ако влезните променливи се следниве:

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](#)

55. За дадениот код:  $x = [i \text{ for } i \text{ in range}(0,15)]$  кој тип ќе биде променливата x и кои вредности ќе ги содржи

Question 11  
Not yet answered  
Marked out of 1.00  
[Flag question](#)

За дадениот код:  
 $x = [i \text{ for } i \text{ in range}(0,15)]$

Кој тип ќе биде променливата x и кои вредности ќе ги содржи?

- a. x е од тип tuple и ги содржи вредностите: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
- b. x е од тип list и ги содржи вредностите: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
- c. x е од тип list и ги содржи вредностите: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
- d. x е од тип set и ги содржи вредностите: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
- e. x е од тип set и ги содржи вредностите: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

- Одговор –под е - **X е од тип set и ги содржи вредностите 0,1,2,3,4,5,6,7,8,9,10,11,12,13,14**

56. Селектирајте ги особините кои се карактеристиките за Data Science , а не и за Machine Learning

Question 6  
Not yet answered  
Marked out of 1.00  
[Flag question](#)

Селектирајте ги особините кои се карактеристични за Data Science а не и за Machine Learning

Time left 0:18:38

- a. Develop/use tools that can handle massive datasets
- b. Publish a paper
- c. Take action!
- d. Improve/validate on a few, relatively clean, small datasets

[Previous page](#) [Next page](#)

Tocen odgovor pod : c , a

-Take action! , Develop/use tools that can handle massive datasets

57. За табелата df:

Под: а

Question 4  
Not yet answered  
Marked out of 1.00  
Flag question

Time left 0:19:04

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 на цената за секој тип на куќа  
 b. Средната вредност типот на куќи (колоната house\_type)  
 c. Ниту едно од наведените  
 d. Средната вредност на цената за сите куќи

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](#)

58. Потребно е да се креира модел за предвидување на дневната температура, ако како влезните податоци ги зема:

Question 3  
Not yet answered  
Marked out of 1.00  
Flag question

Time left

Потребно е да се креира модел за предвидување на дневната температура, ако како влезни податоци ги зима:

1. влажноста на воздухот
2. температурата од претходниот ден
3. дали време или не во моментот

Кој од дадените модели е точен?

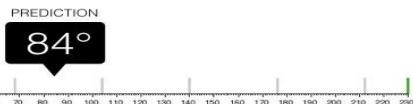
a. `model = DecisionTreeClassifier()`  
 b. `model = DecisionTreeRegressor()`

Под а



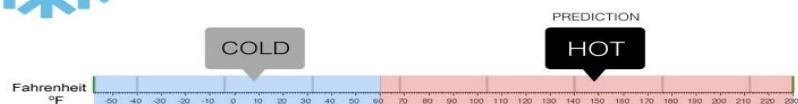
## Regression

What is the temperature going to be tomorrow?



## Classification

Will it be Cold or Hot tomorrow?



59. Кои се предности на Long Short-term Memory (LSTM) мрежите ?

Question 7  
Answer saved  
Marked out of 1.00  
Flag question

Кои се предностите на Long Short-term Memory (LSTM) мрежите?

Select one:

a. Можноста за учење на долги низи.

b. Потреба од мала меморија.

c. Краткотрајно бришење од меморијата.

d. Брзо учење при обука.

Clear my choice

60. Што претставува хиперпараметарот n\_estimators = 5 во XGBoost модел ?

Што претставува хиперпараметарот n\_estimators = 5 во XGBoost моделот?

a. 5 процесори да се искористат за тренирање на моделот

b. 5 листа на дрвото на одлука

c. 5 дрва на одлука кои паралелно ќе се изградат

d. 5 внатрешни јазли во дрвото на одлука

- Tocen odgovor e pod: С 5 дрва на одлука кои паралелно ќе се изградат

61. Кога дистрибуцијата на податоци е како на слика, какви се наклонетоста(bias) и варијанса кај овие податоци ?

Question 12  
Not yet answered  
Marked out of 1.00  
Flag question

Time left 0:17:07

Кога дистрибуцијата на податоците е како на slikata. Какви се наклонетоста (bias) и варијансата кај овие податоци?



Select one:

- a. Неможе да се заклучи од дадечниот график
- b. Голема наклонетост и мала варијанса
- c. Мала наклонетост и голема варијанса
- d. Голема наклонетост и голема варијанса

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

В i e ( no ne znam dali e tocen odgovorot)

63. На кој начин би требало да сго избереме параметарот ламбда во техника за Регуларизација ?

Question 3  
Answer saved  
Marked out of 15.00  
Flag question

На кој начин би требало да го избереме параметарот ламбда во техниката за Регуларизација?

За да се избере ламбда, потребно е да се користи крос валидација бидејќи се земе преголема или премала ( $=>0$ ) вредност, би добивале отстапувања од стандардната грешка. За да се избегне тоа, се користи формула, така што се избира ламбда, а потоа се пресметува коефициент за корелација  $R^2$ . Се зема во предвид најголемиот пресметан коефициент на корелација кој одговара и тоа ламбда што се користело во таа формула, тоа се користи како крајно.

64. Што може да се заклучи кога податочното множество има висока варијанса и висок баес?

Question 1  
Answer saved  
Marked out of 5.00  
Flag question

Што може да се заклучи кога податочното множество има висока варијанса и висок баес ?

Select one:

- a. Претпоставките за таргет функцијата нема да бидат со голема точност, воедно ќе се зголеми прецизноста поради големите промени кај податоците.
- b. Претпоставките за таргет функцијата ќе бидат со голема точност, воедно ќе се намали прецизноста поради големите промени кај податоците.
- c. Претпоставките за таргет функцијата ќе бидат со голема точност, воедно ќе се зголеми прецизноста поради големите промени кај податоците.
- d. Претпоставките за таргет функцијата нема да бидат со голема точност, воедно ќе се намали прецизноста поради големите промени кај податоците.

[Clear my choice](#)

Give your reasons



65. Кои се предности а кои недостатоци на Convolutional Neural Network ?

Question 2  
Incomplete  
Mark: 10.00 out of 10.00  
Flag question  
Edit question

Кои се предностите а кои недостатоците на Convolutional Neural Network

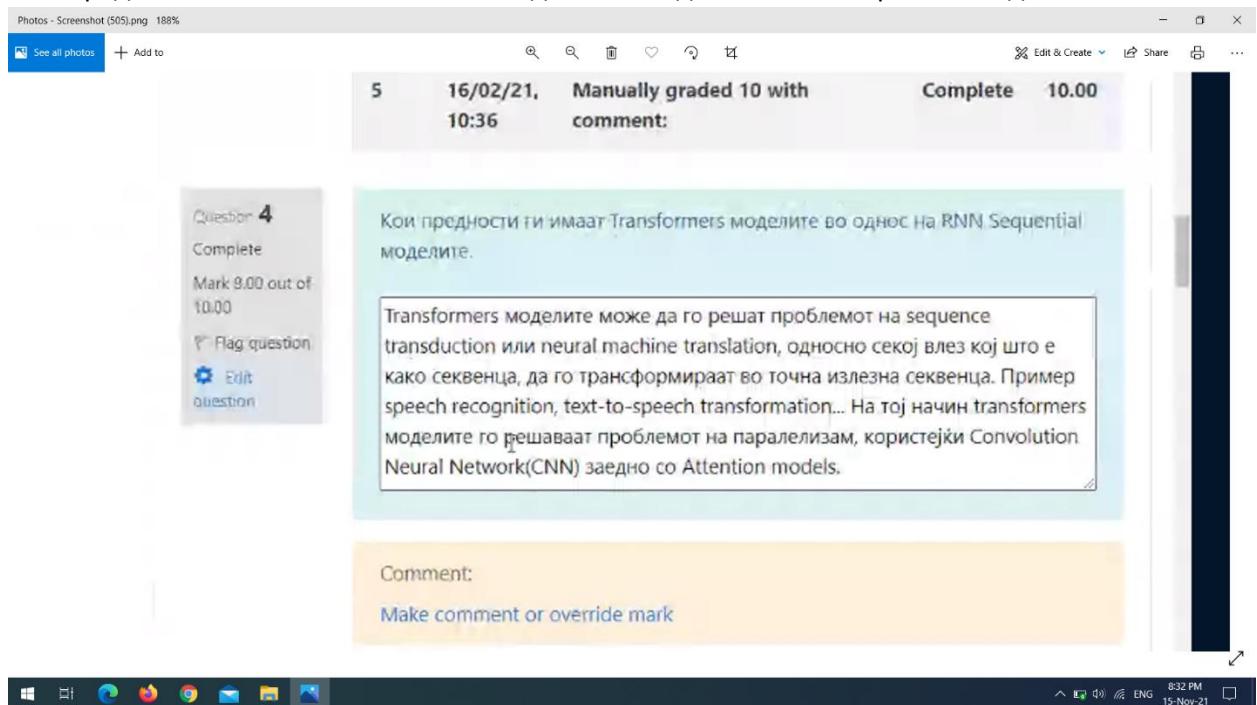
Во конволуциски мрежи предност е што секој слој е поврзан со константен број на единици од слојот пред него. Исто така единиците ги споделуваат тежините за овие поврзаности, односно конекции. Друга предност на овие мрежи е што нивните скриени единици се поврзани со одреден број на единици од слојот кој што следи после нив, со што се намалува бројот на параметри. Конволуциските мрежи како влезови прифаќаат матрици. Овие матрици се користат за процесирање на слики и текст. Недостаток или негативна страна на овој тип на мрежи е следниов: за да се изградиме една мрежа за учење на различни врски меѓу зборови, ќе ни требаат многу филтери со различни висини. Друг недостаток е што конволуциското јадро не е доволно за да се детектираат различни видови на features, кои ќе бидат користени во класификацијата.

- Проба да се воведат структурни подобрувања на мрежата.
- При fully connected има многу параметри кои треба да се научат и оптимизираат.
- Конвулциско ниво. Нема да ги поврзува сите неврон во мрежата, туку на еден неврон ќе ги земе само неговите три соседни влеза. Така се намалува бројот на конекции.

Се заклучува дека пример ако гледаме слика нема потреба целата одеднаш да се види туку може парче по парче. Специјален тип на множење на матрици каде се множат вредностите обратно (првиот со последниот, итн...). Се додава и падинг за да добиеме иста слика.

Така се креира CNN и има три фази : фаза на собирање на конволуција, нелинеарност и pooling.

## 66. Кои предности ги имаат Transformers моделите во однос на RNN Sequential модели



RNN - Излезот од една мрежа се праќа како излез од неа, но се прави копија од тој излез и потоа се праќа на следната мрежа

Transformers

- input | output encoder | decoder
- за да може да учи повеќе, делот со енкодери е спакуван еден врз друг на повеќе нивоа, кој што понатаму излезот од последниот енкодер ќе носи кај декодерот и тие пробуваат да декодираат. Додавањето на нивоа намерно, бидејќи сакаме моделот да знае повеќе и треба да има повеќе параметри. Колку повеќе нивоа – толку попаметен модел

- Во секој енкодер има селф атентион ниво и над него ФФНН. Поентата е што ФФНН ја даваат нелинеарноста бидејќи селф атентион е линеарно.
- На влезот на нивото доаѓа влезот кај е енкодиран . Потоа низ селф атентион ќе се добие самиот збор колку тој е поврзан со останатите. Тоа оди во невронската мрежа и се добива првата трансформација на зборот. Значењето на зборот се модификува во зависност од другите зборови во реченицата.

67. Објаснете го принципот на кој работат autoencoder-ите.

**Question 3**  
Complete  
Mark 9.00 out of 10.00  
[Flag question](#)  
[Edit question](#)

Објаснете го принципот на кој работат autoencoder-ите.

Аутоенкодерите работат на тој начин што автоматски пронаоѓаат најдобар начин за енкодирање на влезот, така што резултатот при декодирање да биде најблизок со оригиналниот влез. Енкодерите се претставници на supervised learning. Начинот на кој што тие работат е што најпрво правиме train на две мрежи преку минимизирање на loss function (cross entropy loss). Аутоенкодерите се користат за енкодирање на слики, видеа и аудио звуци, со цел да се намали нивната големина за да можат полесно да се испратат низ телекомуникациските мрежи. При кодирањето се губат податоци, и затоа постојат lossy и lossless енкодирање. Што поголема е разликата помеѓу оригиналниот податок и декодираниот податок, тогаш станува збор за lossy енкодирање. Овој тип на енкодирање најчесто се користи кај енкодирањето на слики, видеа, а поретко кај звуци и текстови. Бидејќи ако премногу ја намалиме големината на еден текст тогаш ќе ја изгубиме смислата на самиот текст, ќе изгубиме важни податоци.

68. Објаснете ја мерката која овозможува да се направи баланс помеѓу Precision и Recall.

**Question 1**  
Complete  
Mark 7.00 out of 10.00  
[Flag question](#)  
[Edit question](#)

Објаснете ја мерка која овозможува да се направи баланс помеѓу Precision и Recall.

Precision означува колку (колкав процент) од селектираниот елементи се релевантни, точни. A Recall се означува колкав процент се релевантни од вкупниот број на релевантни елементи.

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

дел од трансформер архитектурата се користи во GPT-2

### Decoder

70. 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**

71. 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)**

72. Кои од наведените дескриптивни статистики е најдобро да се изберат ако податочното множество се состои од континуирани податоци и при тоа е потребно да се прикаже варијација на истите?

**Ранг , Интер-квартална разлика, Стандардна Варијација**

73. Кои од наведените може да се користат како критериум за поделба (Splitting Criterion) на јазлите кај дрвата за одлучување?

- **Грешка при класификација , Ентропија , Индексот Чини**

74. Кои од следните репозиториуми/библиотеки се користат за едноставно споделување на претренираните NLP модели.

- **HuggingFace transformsrs library, PyTorch Hub, TensorFlow-Hub**

75. Кои од следните се називи на алгоритми за оптимизација кај невронските мрежи?

- **Adam , Adagrad**

76. 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**.

77.

1. Да се објасни за Gradient boosting и зошто е тој ефикасен?

Gradint Boosting - се проучува нов learner за остаток од целокупниот модел.

Приближување на преостанатиот дел и додавање на претходниот резултат е во суштинска форма на спуштање на градиент.На пример: XGBoost

$$T = \sum_h \lambda_h T_H$$

Клучната интуиција зад засилувањето е дека може да се земе ансамбл од едноставни модели  $\{T_h\}_{h \in H}$  и дополнително комбинирајте ги во единствена, посложена модел.

- Секој модел  $T_h$  може лошо да одговара на податоците, но линеарна комбинација на ансамбл
- може да биде експресивен/флексибилен.

Gradient Boosting е метод за итеративно градење на комплексен регресивен модел  $T$  со додавање едноставни модели. Секој нов едноставен модел додаден во ансамблот ги компензира слабостите на актуелниот ансамбл.

#### 78. 2. Да се објасни за концептот positonal encoding кај трансформерите.

Positional encoding - Трансформаторот додава вектор на секое влезно вградување. Овие вектори следат специфично модел што го учи моделот, што му помага да ја одреди позицијата на секој збор или на растојание помеѓу различни зборови во низата.

Интуицијата овде е дека додавањето на овие вредностите на вградувањето обезбедува значајни растојанија помеѓу векторите за вградување откако ќе бидат проектирани во Q/K/V вектори и за време на вниманието на производ-точка.. Но, со користење на бинарни вредности за позиционо кодирање

би било губење простор во светот на floats.. Така, наместо тоа, можеме да го користиме sinusoidal functions.

Positional embedding – Значењето на зборот и влијанието негово врз контекстот на реченицата зависи од позицијата на зборот во самата реченица. За да се постигне ова се додава Positional embedding. Бидејќи користење на вредност за позиционо енкодирање би било трошење на време и простор, затоа се користат синусоидни функции (син и кос). Место сигнали кои ги прави бинарниот енкодинг. Ќе се земаат врз база на кое место во секвенцата се појавуваат. Се прави одредена фреквенција и се претставуваат како син и кос. Вредностите што ќе се добијат како резултати од син и кос се додаваат како уште едно бројче во ембединг кое ќе ни кажува на кое место се наоѓа зборот во реченицата. И вака се решава проблемот за редоследот.

#### 79. 3. Објасни ги разликите помеѓу GINI и Entropy метриките.

Gini index ја пресметува количината на веројатност за одредена карактеристика што е погрешно класифицирана кога е избрана по случаен избор. Ако сите елементи се поврзани со една класа, тогаш таа може да се нарече чиста.

Како својствата на ентропијата, Gini index варира помеѓу вредностите 0 и 1, каде што 0 ја изразува чистотата на класификацијата, односно сите елементи припаѓаат на одредена класа или постои само една класа таму. И 1 означува случајна дистрибуција на елементи низ различни класи. Вредноста од 0,5 на Џини Индексот покажува еднаква дистрибуција на елементите низ некои класи.

$$\text{Gini}(i|j, t_j) = 1 - \sum_k p(k|R_i)^2$$

• 

Ентропија - Еден начин да се измери јачината на сигналот во одреден регион е да се анализира

распределбата на паралелките во рамките на регионот. Ја пресметуваме ентропијата на ова дистрибуција.

- За случајна променлива со дискретна распределба, ентропијата се пресметува со:
- Повисока ентропија значи дека распределбата е униформа (рамен хистограм) и на тој начин вредностите земени од него се „помалку предвидливи“ (сите можни вредности се подеднакво веројатно).

Пониска ентропија значи дека дистрибуцијата има подефинирани врвови и долини и така што вредностите земени од него се „попредвидливи“ (вредностите околу врвовите се повеќе веројатно).

Ентропијата се дефинира како случајност или мерење на нередот на информациите што се обработуваат во машинското учење. Понатаму, со други зборови, можеме да кажеме дека ентропијата е метрика за машинско учење што ја мери непредвидливоста или нечистотијата во системот.

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x)$$

#### 80. Кои се предностите на LSTM во однос на обичниот Vanilla RNNs.

Невронот во LSTM е специјален со тоа што во него има мемориска клетка која памети што се случувало и следен пат кога ќе треба да се пресметува нов влезен податок може да се види што се случило претходно и така аналогно се ажурира.

- за дадена мрежа од секој влез  $x_1, x_2$  одат тежините до контролни точки на LSTM мрежата и врз база на тие сигнали одлучуваат што ќе се зема во мрежата, како ќе се зема во мрежата, како ќе се пресметува тоа што ќе се запамти во мрежата, како ќе се пресметува нелинеарност, што ќе се даде на излез, дали ќе се памети или заборави нешто во мрежата. Се додаваат 4 пати повеќе параметри на влезот на мрежата, но тие се помалце отколку ако копираме реченица од 100 збора.

Тип на RNN кој е дизајниран за подобро справување со long-range dependencies

- Во „vanilla“ RNN, скриената состојба постојано се препишува
  - Покрај традиционалната скриена состојба  $h$ , да имаме посветено мемориска ќелија с за долгочочни настани. More power to relay sequence info
- Речиси секогаш ги надминува перформансите на vanilla RNN

- Шокантно добро гиоловува зависностите на долг дострел  
Има повеќе тежини за учење од vanillaRNN; на тој начин,
- Потребна е умерена количина на податоци за обука (во спротивно, vanilla RNN се подобри)
- Сè уште може да страда од наклони кои исчезнуваат/експлодираат

81. 5. Ако се занемарат 5 димензии ако PCA покрива околу 90% од варијансата.

PCA се користи за намалување на димензионалности

Pca = PCA(n\_components=5) – број на димензии

Result = pca.fit\_transform() – ќе ги учи и трансформира податоци

pca - прави сумеризација на податоците

82. 6. Кои се разликите помеѓу N-grams и skip-grams ?

#### N-grams

- Подобро од BoW. Кај н-грамс за разлика од BoW битен е редоследот. Не треба да се гледа збор по збор, туку да има групирање –комбинирање на зборови. Прво два по два, па три по три итн. Тоа се прави за да има реченицата некаков контекст, но бројот на комбинации значително се зголемува. А и голем дел од комбинациите се невозможни . Исто така некои n-grams се многу фреквентни, а некои не се.
- unigrams имаат повисоко бројење и може да се детектираат слаби влијанија. Додека Bi и trigrams фаќаат силни влијанија кои се поспецифични.

#### Skip-grams

N-grams пробуваат да ги најдат сите можни комбинации. Тоа не е секогаш добро. Треба да се дефинира контекстот во реченицата. За тоа да се постигне се јавуваат Skip-grams. Тоа е контекст во кој се појавува одреден збор и ги гледа неколку зборови пред и после него. Ова е ефикасно бидејќи го фаќа контекстот на зборот пред и после појавувањето. Основа е за успешна имплементација на првиот NLP јазичен модел Word2vec.

83. Precision and Recall

Precision - под ова мислиме на процентот на вистински позитиви, NTP, меѓу сите примери што класификаторот ги означи како позитивни: NTP + NFP . Вредноста е добиени со следнава формула:

$$Pr = \frac{N_{TP}}{N_{TP} + N_{FP}}$$

Recall - под ова мислиме на веројатноста дека ќе биде позитивен пример правилно препознаен од класификаторот. Вредноста се добива со делење на бројот на вистински позитиви, NTP , со бројот на позитивните во даденото множество: NTP + NFN

$$Re = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

84. За дадена визуелизација определи која од наведените команди одговара.

The screenshot shows a Microsoft Photos app window with a histogram visualization and a corresponding multiple-choice question.

**Photos - Screenshot (268).png 79%**

**Quiz navigation:** 1, 2, 3, 4, 5, 6, 7, 8

**Question 6:** Not yet answered. Marked out of 1.00. Flag question.

За дадената визуелизација определи која од наведените команди одговара.

**Figure Description:** A histogram showing the distribution of 'Bed Grade' values. The x-axis categories are labeled c, e, b, a, f, d, g. The y-axis ranges from 0 to 200,000. The distribution is highly right-skewed, with the highest frequency in category 'a' (approx. 200,000) and lower frequencies in other categories.

**Select one:**

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

[Clear my choice](#)

System tray icons: Windows, Task View, File Explorer, Mozilla Firefox, Google Chrome, Mail, File Explorer, Task View.

System status bar: ENG, 8:29 PM, 15-Nov-21.

Odgovor tocen pod a