

Вовед во науката за податоци

Прв Колоквиум

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

- 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.
- E. Нормализација ги праве вредностите меѓу 0 и 1 а стандардизација, со средна вредност 0 а варијанса 1

Одговор: A

2. With the command df.mean() what is the output result?

- A. Only for the categorical columns of the df dataset will the mean be printed
- B. For each of the columns od the df dataset the mean value will be printed.
- C. Only for the numeric columns of the df dataset the mean value will be printed.

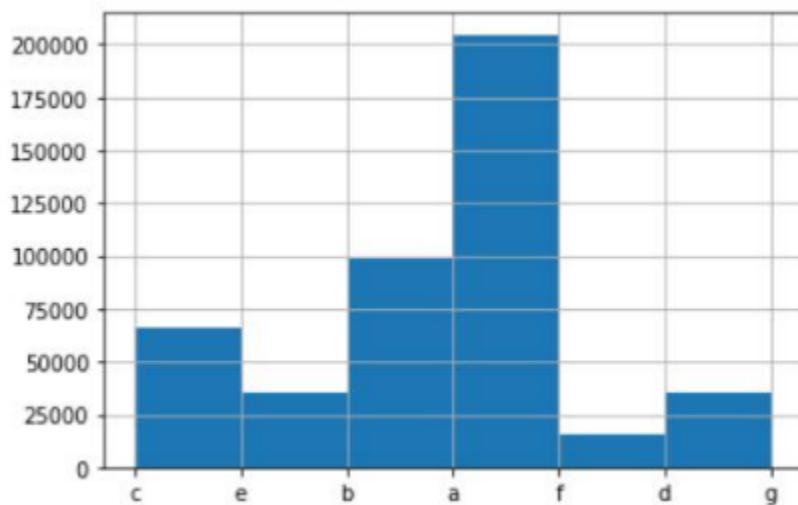
Одговор: C

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

- A. Frequency
- B. Median value
- C. Percent (row, column or total)
- D. Mean value

Одговор: B, D

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



- 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)`

Одговор: D

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.

- A. Ласо регуларизација (LASSO Regularization)
- B. Вкрстена валидација (Cross Validation)
- C. Регуларизација по сртот (Ridge Regularization)
- D. Ентропија

Одговор: B

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

| Class | Class |
|-------|-------|
| 1 | 2 |

$$3 \text{a R1} \rightarrow 1 - ((2/7)^2 + (5/7)^2) = 0.408$$

$$3 \text{a R2} \rightarrow 1 - ((6/10)^2 + (4/10)^2) = 0.48$$

$$\text{R1} \rightarrow 7/17 * 0.408 = 0.168$$

$$\text{R1 и R2} \rightarrow 7/17 * 0.408 + 10/17 * 0.48 = 0.45$$

| | | |
|----|---|---|
| R1 | 2 | 5 |
| R2 | 6 | 4 |

- A. 0.282
- B. 0.45
- C. 0.168
- D. 0.5

Одговор: С

7.

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

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

- A. Visualization
- B. Distance
- C. Prediction of coefficients

Одговор: В

9. Ако треба да се одреди припадноста на даден клиент во една од четирите групи на корисници, за каков вид на машинско учење станува збор?

- A. Класификација (Classification)
- B. Откривање на недостатоците (Anomaly Detection)
- C. Регресија (Regression)
- D. Учење со поттикнување (Reinforcement Learning)

Одговор: А

10. 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?

- A. Средната вредност е поголема од медијаната
- B. Медијаната е поголема од средната вредност
- C. Средната вредност и медијаната се еднакви
- D. Не може да се заклучи од дадениот графика

Одговор: А

11. 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

1. No $\rightarrow \sqrt{(6-2)^2 + (3-2)^2} = 4.12$
 2. No $\rightarrow \sqrt{(5-2)^2 + (4-2)^2} = 3.61$
 3. Yes $\rightarrow \sqrt{(4-2)^2 + (2-2)^2} = 2$
 4. Yes $\rightarrow \sqrt{(3-2)^2 + (3-2)^2} = 1.41$
- $k=3 \rightarrow \#4, \#3, \#2 \rightarrow$ Yes, Yes, No \rightarrow Yes

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 ✓

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.

1. No $\rightarrow \sqrt{(1-4)^2 + (3-2)^2} = 3.16$
 2. No $\rightarrow \sqrt{(0-4)^2 + (4-2)^2} = 4.47$
 3. Yes $\rightarrow \sqrt{(2-4)^2 + (2-2)^2} = 2$
 4. Yes $\rightarrow \sqrt{(3-4)^2 + (5-2)^2} = 3.16$
- $k=3 \rightarrow \#3, \#4, \#1 \rightarrow$ Yes, Yes, No \rightarrow Yes

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

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

```
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')
```

13. 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

Одговор: A

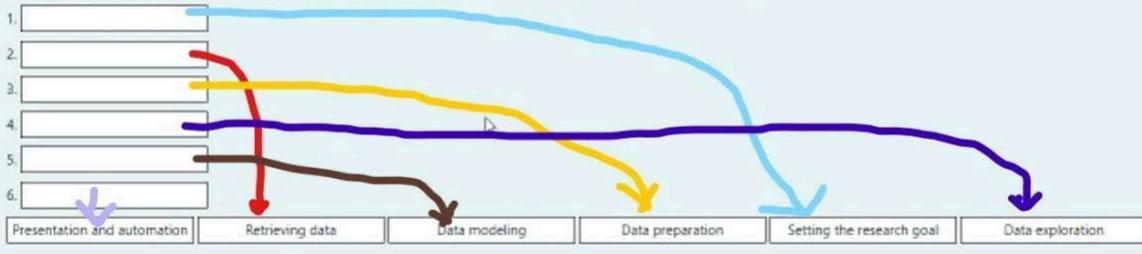
14. Кои од визуелизациите е најдобро да се изберат кога станува збор за датасет од категориски податоци?

- A. Dot plot
- B. Scatter plot
- C. Histogram
- D. Bar charts

Одговор: D

15. Подредете ги по редослед чекорите од кои се состои процесот за развој на Data Science проекти (Data Science Process).

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

16. 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.
- Сличноста помеѓу центроидите на двата кластера.

Одговор: C, D

17. 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.

- A. metric = "entropy"
- B. criterion = "entropy"
- C. spitter = "entropy"

Одговор: B

18. Даден е модел на логистичка рересија (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_

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

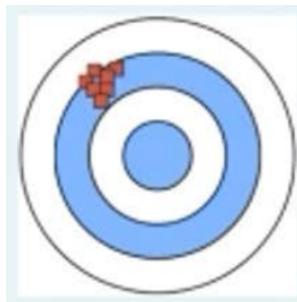
Одговор: A

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

- A. Откривање на недостатоци (Anomaly Detection)
- B. Учење со поттикнување (Reinforcement Learning)
- C. Класификација (Classification)
- D. Регресија (Regression)

Одговор: D

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



- A. Мала наклонетост и мала варијанса
- B. Голема наклонетост и мала варијанса
- C. Мала наклонетост и голема варијанса
- D. Голема наклонетост и голема варијанса

Одговор: B

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

- 1. - температура (десимални вредности)
- 2. - влажност на воздухот (десимални вредности)
- 3. - дали врнело во текот на денот (категорисја вредност → Yes / No)
 - A. Хистограм за секоја од колоните
 - B. Хистограм на целиот датасет
 - C. Хистограм за секоја од нумеричките колони
 - D. Ниту едно од наведените

Одговор: C

22. Што ќе се случи со дадениот код: `df.drop([2,3], axis=0)`

- A. Ќе ги избрише 2 и 3 колона директно во датасетот
- B. Ќе ги избрише 2 и 3 колона од датасетот и ќе го врати новиот датасет како вредност
- C. Ќе ги избрише 2 и 3 редица од датасетот и ќе го врати новиот датасет како вредност
- D. Ќе ги избрише 2 и 3 редица директно во датасетот

Одговор: C

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

- A. Податоците се во стандарден формат за реални броеви
- B. Нивниот ранг сега е помеѓу минималниот и максималниот елемент во датасетот
- C. Средната вредност на податоците е 0, а варијансата е 1

D. Нивниот опсег е помеѓу 0 и 1

Одговор: C

24. Што се случува во дадениот код?

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

html = requests.get(url).text

soup = BeautifulSoup(html, 'html.parser')

- A. Ја враќа и прикажува веб страната
- B. **Ја парсира html содржината од веб страна "sitel.mk"**
- C. Ја симнува веб страната
- D. Ниту едно од наведените

Одговор: B

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

- A. **Фреквенција**
- B. Средна вредност
- C. **Процент (редица, колона или вкупно)**
- D. Медијана

Одговор: A, C

26. Кој е излезот од дадениот код:

d = {'S': 'super', 'G': 'good', 'B': 'bad'}

d['S'] = 'SUPER'

- A. ['S': 'super', 'G': 'good', 'B': 'bad']
- B. **['S': 'SUPER', 'G': 'good', 'B': 'bad']**
- C. ['S': 'super', 'G': 'good', 'B': 'bad', 'S': 'SUPER']

Одговор: B

27. Ако треба да се избере соодветна акција за одреден нов податок, за каков вид на машинско учење станува збор?

- A. Откривање на недостатоците (Anomaly Detection)
- B. Учење со поттикнување (Reinforcement Learning)
- C. **Класификација (Classification)**
- D. Регресија (Regression)

Одговор: **B C**

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

| За табелата df: | | |
|-----------------|----------------|--------|
| | house_type | price |
| 0 | Flat | 100000 |
| 1 | House 1 floor | 200000 |
| 2 | House 2 floors | 300000 |

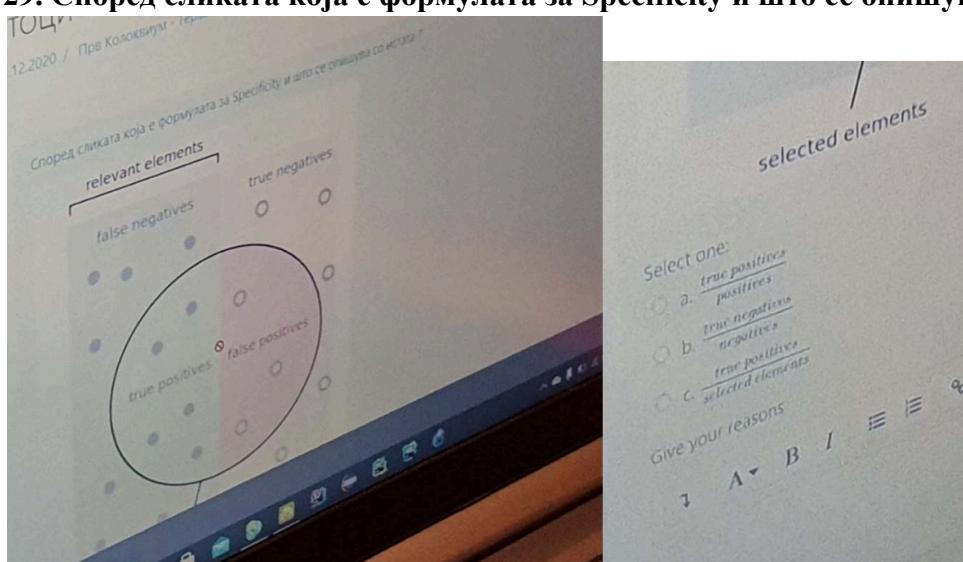
Што резултат ќе изгенерира дадениот код:

```
df.groupby(['house_type'], as_index=False).count()
```

- 36.
- A. Средната вредност на типот на куќи
 - B. Бројот на куќи
 - C. **Бројот на инстанци за секој тип на куќа**
 - D. Ниту едно од наведените

Одговор: C

29. Според сликата која е формулата за Specificity и што се опишува со истата?



- A. true positives / positives (this is sensitivity)
- B. **true negatives / negatives**
- C. true positives / selected elements

Одговор: В

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

| 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 | Yes |
| 8 | Divorced | High | No |

(Заокругли ги десичалните места ако се повеќе на втората десичала)
Classification Error за колоната Marital Status изнесува:
Classification Error за колоната Annual Income изнесува:
За следна поделба на дрвото на одлука се избира колоната

Solution:

| Marital Status | Yes | No | Classification Error |
|----------------|-----|----|------------------------------------|
| Single | 0 | 2 | $1 - \text{Max}\{0/2, 2/2\} = 0$ |
| Married | 1 | 1 | $1 - \text{Max}\{1/2, 1/2\} = 0.5$ |
| Divorced | 2 | 2 | $1 - \text{Max}\{2/4, 2/4\} = 0.5$ |

$$0*2/8 + 0.5*2/8 + 0.5*4/8 = 0.375 \sim 0.38$$

| Annual Income | Yes | No | Classification Error |
|---------------|-----|----|-------------------------------------|
| High | 2 | 1 | $1 - \text{Max}\{2/3, 1/3\} = 0.33$ |
| Medium | 1 | 1 | $1 - \text{Max}\{1/2, 1/2\} = 0.5$ |
| Low | 0 | 3 | $1 - \text{Max}\{0/3, 3/3\} = 0$ |

$$0.33*3/8 + 0.5*2/8 + 0*3/8 = 0.24875 \sim 0.25$$

Се избира **Annual Income**

31. За дадената табела во прилог ако се енкодира колоната Class со помош на One-Hot Encoding како ќе изгледа ново добиената табела?

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

a. (label encoding)

Select one:

a.

| <i>Id</i> | <i>Company_name</i> | <i>Class</i> |
|------------------|----------------------------|---------------------|
| 1 | Pepsi | 1 |
| 2 | Zara | 2 |
| 3 | Coca - Cola | 1 |
| 4 | Apple | 3 |
| 5 | Mcdonald's | 4 |

b. (one-hot encoding)

b.

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

c. (binary encoding)

c.

| <i>Id</i> | <i>Company_name</i> | <i>Class_1</i> | <i>Class_2</i> | <i>Class_3</i> |
|------------------|----------------------------|-----------------------|-----------------------|-----------------------|
| 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 |

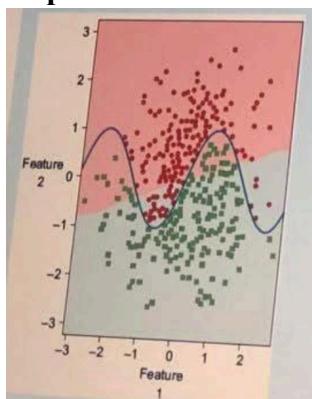
Одговор: В

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

- 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

Одговор: B, D, E

33. Ако и прикажеме податоците како на сликата, со точки во дво-димензионален ... ??? ... поделба во две класи да ја прикажеме со различни бои, за каков модел станува збор?



- A. Модел од машинското учење
- B. Линеарен модел
- C. Детерминантен модел
- D. Колинеарен модел
- E. Нелинеарен модел

Одговор: B

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

| Клуч предност | |
|---------------|-------|
| S | super |
| G | good |
| B | bad |

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

Ги нема останатите одговори, ама точните се:

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

35. 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_type – 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 |
| Raspberry | Medium | Medium | Sour |
| Cherry | Medium | Medium | Sweet |
| Banana | High | Low | Sweet |
| Watermelon | High | Low | Sweet |
| Mandarin | Extremely Low | Medium | None |

(Задржуј ги десималните места ако се повеќе на втората десимала)

Classification Error за колоната Sweetness изнесува: 0.33

Classification Error за колоната Sourness изнесува: 0.21

За следна поддршка на одлуката се избира колоната – Sweetness

Solution:

| Sweetness | Sou r | Sweet | Non e | Classification Error |
|---------------|----------|-------|----------|---|
| Extremely low | 1 | 0 | 1 | $1 - \text{Max}\{1/2, 0/2, 1/2\} = 0.5$ |
| Low | 2 | 0 | 0 | $1 - \text{Max}\{2/2, 0/2, 0/2\} = 0$ |
| Medium | 1 | 1 | 0 | $1 - \text{Max}\{1/2, 1/2, 0/2\} = 0.5$ |
| High | 0 | 2 | 0 | $1 - \text{Max}\{0/2, 2/2, 0/2\} = 0$ |

$$0.5*2/8 + 0*2/8 + 0.5*2/8 + 0*2/8 = \underline{\underline{0.25}}$$

| Sourness | Sou r | Sweet | Non e | Classification Error |
|----------|----------|-------|----------|---|
| Low | 0 | 2 | 0 | $1 - \text{Max}\{0/2, 2/2, 0/2\} = 0$ |
| Medium | 3 | 1 | 1 | $1 - \text{Max}\{3/5, 1/5, 1/5\} = 0.4$ |
| High | 1 | 0 | 0 | $1 - \text{Max}\{1/1, 0/1, 0/1\} = 0$ |

$$0*2/8 + 0.4*5/8 + 0*1/8 = \underline{\underline{0.25}}$$

Се избира ???

36.

question 2
not yet
answered
marked out of
5.00
Flag question

За дадено податочно множество во табелата потребно е, со помош на индексот Цини, да се одреди следната колона по која ќе се врши разграничување на дрвото на одлука. **(Вид_на_овошје** е целна колона т.е според неа се врши класификацијата

| Овошје | Слаткост | Киселост | Вид_на_овошје |
|----------|-------------|----------|---------------|
| Лимон | Многу ниска | Висока | Кисело |
| Цитрон | Многу ниска | Висока | Кисело |
| Портокал | Ниска | Висока | Кисело |
| Малина | Ниска | Средна | Кисело |
| Цреша | Ниска | Средна | Благо |
| Банана | Висока | Ниска | Благо |
| Лубеница | Висока | Ниска | Благо |

(Заокружете ги децималните места, ако се повеќе, на втората децимала)

Просечниот индекс Цини (со тежински фактор) за колоната Слаткост изнесува:

Просечниот индекс Цини (со тежински фактор) за колоната Киселост изнесува:

За следна поделба на дрвото на одлука се избира колоната

Give your reasons

Solution:

| Слаткост | Кисело | Благо | Gini Index |
|-------------|--------|-------|------------------------------------|
| Многу ниска | 2 | 0 | $1 - \{(2/2)^2 + (0/2)^2\} = 0$ |
| Ниска | 2 | 1 | $1 - \{(2/3)^2 + (1/3)^2\} = 0.44$ |
| Висока | 0 | 2 | $1 - \{(0/2)^2 + (2/2)^2\} = 0$ |

$$0*2/7 + 0.44*3/7 + 0*2/7 = 0.19048... \sim 0.19$$

| Киселост | Кисело | Благо | Gini Index |
|----------|--------|-------|-----------------------------------|
| Ниска | 0 | 2 | $1 - \{(0/2)^2 + (2/2)^2\} = 0$ |
| Средна | 1 | 1 | $1 - \{(1/2)^2 + (1/2)^2\} = 0.5$ |
| Висока | 3 | 0 | $1 - \{(3/3)^2 + (0/3)^2\} = 0$ |

$$0*2/7 + 0.5*2/7 + 0*3/7 = 0.14285... \sim 0.14$$

Се избира **Киселост**

37. Накратко објаснете ја мерката R-squared, кога и како би ја користеле.

Ecejско.

38.

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). Ваши задача е да ја предвидите fraudulent target 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 испити 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

```
1 import pandas as pd
2 from sklearn.metrics import f1_score
3
4
5
6 def build_model(train,test):
7     #define X_train,Y_train,X_train,X_test by selecting the columns from the datasets
8
9     #preprocessing of 'description' column
10    #just fit it on training data
11
12    #implement knn model
13
14    #train and predict the values
15    y_pred =
16
17    return f1_score(Y_test,y_pred)
```

```
#define ... by selecting the columns from the datasets
x_train, x_test = train.iloc[ :, :-1], test.iloc[ :, :-1]
y_train, y_test = train.iloc[ :, -1], test.iloc[ :, -1]
#preprocessing of 'description' column
#just fit it on training data
scaler = StandardScaler()
x_train['description'] = scaler.fit_transform(x_train['description'])
x_test['description'] = scaler.fit_transform(x_test['description'])
#implement knn model
model = KNeighborsClassifier()
```

```
#train and predict the values
model.fit(x_train,y_train)
y_pred = model.predict(x_test)
```

39.

Question 4
Not complete
Marked out of 25.00
[Flag question](#)

Time left 0:52:24

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 испити 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](#)

```

1 import pandas as pd
2 from sklearn.metrics import roc_auc_score
3
4
5 def split_data(x,y,test_size):
6     X_train = x[:int(len(x)*(1-test_size))]
7     X_test = x[int(len(x)*(1-test_size)):]
8     Y_train = y[:int(len(y)*(1-test_size))]
9     Y_test = y[int(len(y)*(1-test_size)):]
10    return X_train, X_test, Y_train, Y_test
11
12 def build_model(df,test_size):
13     #Handle missing values
14
15
16     #Normalize the values on whole dataset input
17
18     #Split the data with the use of train_split function
19
20
21     # Create Decision Tree model
22     y_pred =

```

Answer: (penalty regime: 0 %)

[Reset answer](#)

```

1 import pandas as pd
2 from sklearn.metrics import roc_auc_score
3
4
5 def split_data(x,y,test_size):
6     X_train = x[:int(len(x)*(1-test_size))]
7     X_test = x[int(len(x)*(1-test_size)):]
8     Y_train = y[:int(len(y)*(1-test_size))]
9     Y_test = y[int(len(y)*(1-test_size)):]
10    return X_train, X_test, Y_train, Y_test
11
12 def build_model(df,test_size):
13     #Handle missing values
14
15
16     #Normalize the values on whole dataset input
17
18     #Split the data with the use of train_split function
19
20
21     # Create Decision Tree model
22     y_pred =

```

[Precheck](#) [Check](#)

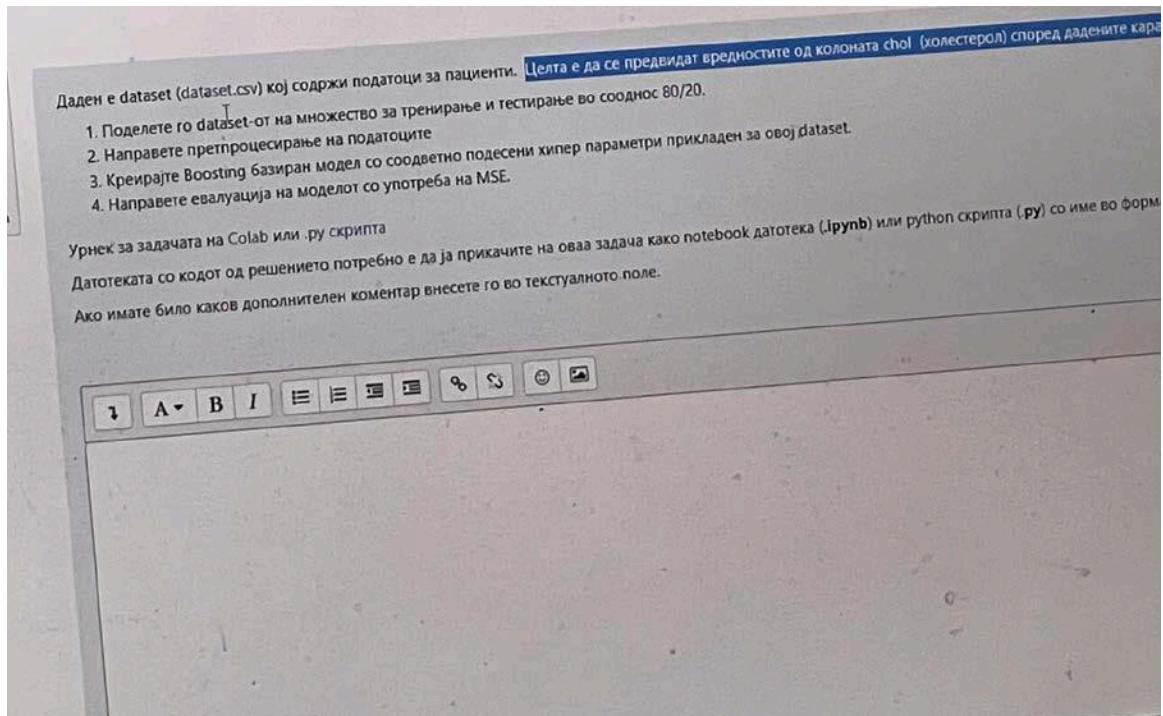
```
#Handle missing values
df.fillna(method="bfill")
#Normalize the values on whole dataset input
scaler = StandardScaler() / scaler = MinMaxScaler()
```

```

df = scaler.fit_transform(df)
#Split the data with the use of train_split function
x_train, x_test, y_train, y_test = train_split(df.iloc[:, :-1], df.iloc[:, -1], test_size)
# Create Decition Tree model
model = DecisionTreeClassifier(max_depth=15)
model.fit(x_train, y_train)
y_pred = model.predict(x_test)

```

41.



```

import pandas as pd
import xgboost as xgb
from sklearn.metrics import mean_squared_error
df = pd.read_csv('dataset.csv')
train = df[:int(0.8 * len(df))]
test = df[int(0.8 * len(df)):]
x_train, x_test = train.iloc[:, :-1], test.iloc[:, :-1]
y_train, y_test = train.iloc[:, -1], test.iloc[:, -1]
scaler = StandardScaler()
df = scaler.fit_transform(df)
model = xgb.XGBClassifier()
model.fit(x_train, y_train)
y_pred = model.predict(y_test)
mse = mean_squared_error(y_test, y_pred)
print(mse)

```

40.

Во оваа задача е потребно да напишете три функции:

- **encoding(data, columns):** **data** е од тип pandas DataFrame, **columns** е листа од имена на колоните кои се лабелирани и е потребно да се јнкодираат. Целта на оваа функција е да се јнкодираат 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 имената 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 ispti со име "Датасет - Термин 2" и истиот можете од таму да го симнете!

For example:

Втор Колоквиум

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

- A. Обично се подобри од еднонацочните 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

Одговор: A, C

2. Каква дводимензионалност треба да е влезно тренирачко множество кај LSTM невронска мрежа?

- a. 2D – матрица
- b. 1D
- c. 3D

Одговор: C

3. Каков вид на учење се реализира кај Автоенкодерите ?

- a. Нагледувано (supervised)
- b. Полу-нагледувано(semi-supervised)
- c. Само-нагледувано(self-supervised)
- d. Со поттикување(reinforcement)

Одговор: B, C

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

- a. Метод со кој оптимизацијата на тежините обезбедува глобален оптиум.
- 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

Одговор: A

5. Кај Обработката на природни јазици се среќаваат следниве задачи:

- 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

Одговор: A, B, C, D

6. На кој начин се добиваат embedding на зборовите при тренирање на BERT модел?

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

Одговор: C

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

- a. Кога влезовите во мрежата се дискретни вредности
- b. Кога како мрежа за пресметка на загуба во мрежата се користи MSE (Mean Squared Error)
- c. Кога бројот на влезови е поголем од бројот на излези во нервонската мрежа
- d. Кога сакаме да добиеме побрзо процесирање на резултатите на GPU
- e. Кога имаме бинарна класификација

Одговор: Е

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

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

Одговор: А

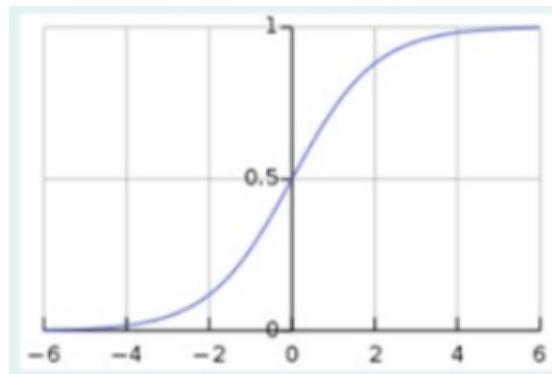
9. Нека е дадена реченицата: “It was a bright cold day in April, and the clocks were striking” Skip-gram со големина на прозорец три за зборот day е:

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

Одговор: D

10. Која активацијска функција е претставена на графикот?

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



Одговор: В

11. Што е точно за моделот seq2seq?

- 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. ??? (во лекциите пишува: "the entire "meaning" of the 1st sequence is expected to be packed into this one embedding", одлучи си дали тоа е исто со понуденото)
- d. При тестирањето се генерираат збор по збор, се додека не се добие на излез знак за крај на реченицата.

Одговор: A, **B**, D

12. Кои карактеристики треба да ги има активиската функција кај невронските мрежи?

- 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

Одговор: A, B

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

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

Одговор: A, B

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

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

Одговор: C

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

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

Одговор: C

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

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

Одговор: B

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

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

Одговор: A, B, D

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

- 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

Одговор: B, D, E

19. Кај Наивните Баесови класификатори, за атрибути A_i за дадена класа C може да претпоставиме:

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

Одговор: B, C

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

- a. Можноста за учење на долги низи
- b. Потреба од мала меморија
- c. Краткотрајно бришење од меморијата
- d. Брзо учење при обука

Одговор: A

Одговор: B

21. Што претставува хиперпараметарот n_estimators = 5 во XGBoost модел?

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

Одговор: C

22. Word2vec како основа за креирање на Embeddings користи:

- a. n-grams
- b. part of speech tagging
- c. skip-grams
- d. one-hot embeddings

Одговор: C

23. За што се користи Latent Dirichlet Allocation(LDA) алгоритмот?

- a. Topic Modeling
- b. Part-of-Speech (POS) tagging
- c. Named Entity Recognition
- d. Open Information Extraction

Одговор: А

24. Колку често можат да се ажурираат тежините кај невронските мрежи?

- a. Ажурирање во серии (batch)
- b. Ажурирање во случајно расфрлани мини-серии (mini-batches)
- c. Ажурирање после секој примерок во множество за обука
- d. Ажурирање во моменти
- e. Ажурирање во конволуции

Одговор: А, В, С

25. Кои особености ги има Преносното учење (Transfer Learning)?

- a. Врши пренос на моментите во друга невронска мрежа
- b. Овозможува подобрување на перформансите
- c. Може да научи преносно значење на зборовите
- d. Врши пренос на испуштените јазли (drop-out) во друга невронска мрежа
- e. Користи означенни податоци од други или сродни области

Одговор: В, С, Е

26. Кои мерки ги користиме за сличност помеѓу два кластера?

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

Одговор: С, D

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

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

Одговор: А, С, D

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

- a. K-Means Clustering

- b. AffinityPropagation Clustering
- c. DBCAN Clustering
- d. Agglomerative Clustering

Одговор: A

29. Што се Skip-grams?

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

Одговор: B

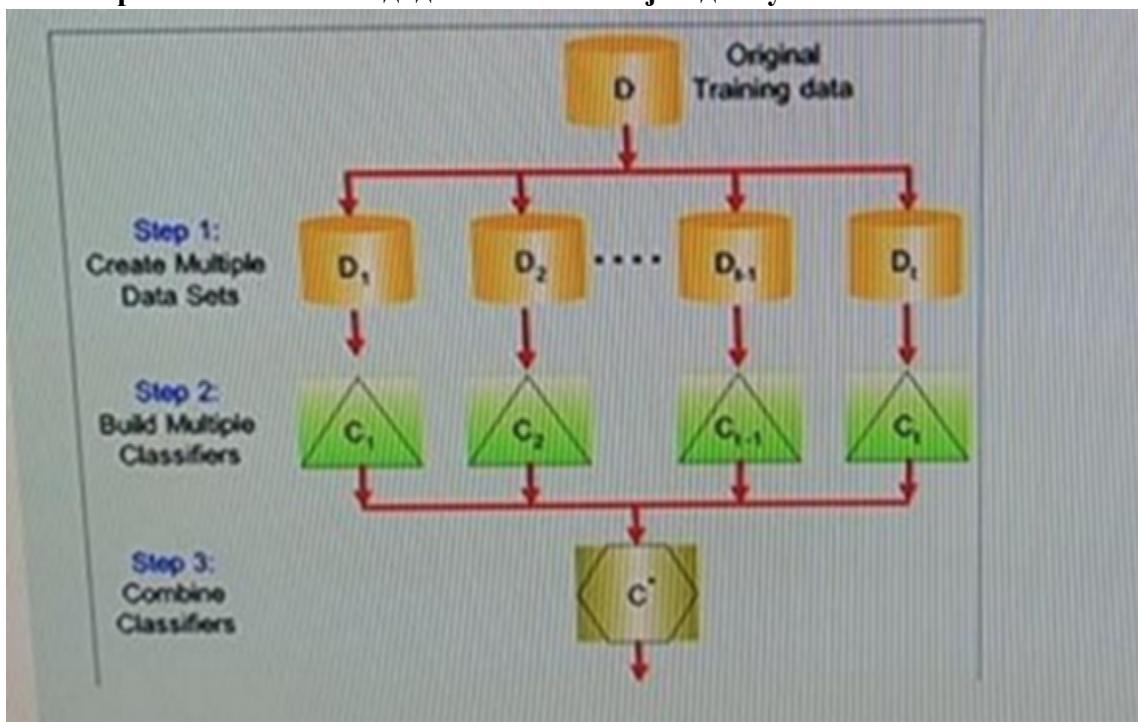
30. Кои од следниве се називи на алгоритми за оптимизација кај невронските мрежи?

- Adam , Adagrad (some others are: Momentum, Adadelta, RMSProp)

31. Што претставува Parts of Speech Tagging?

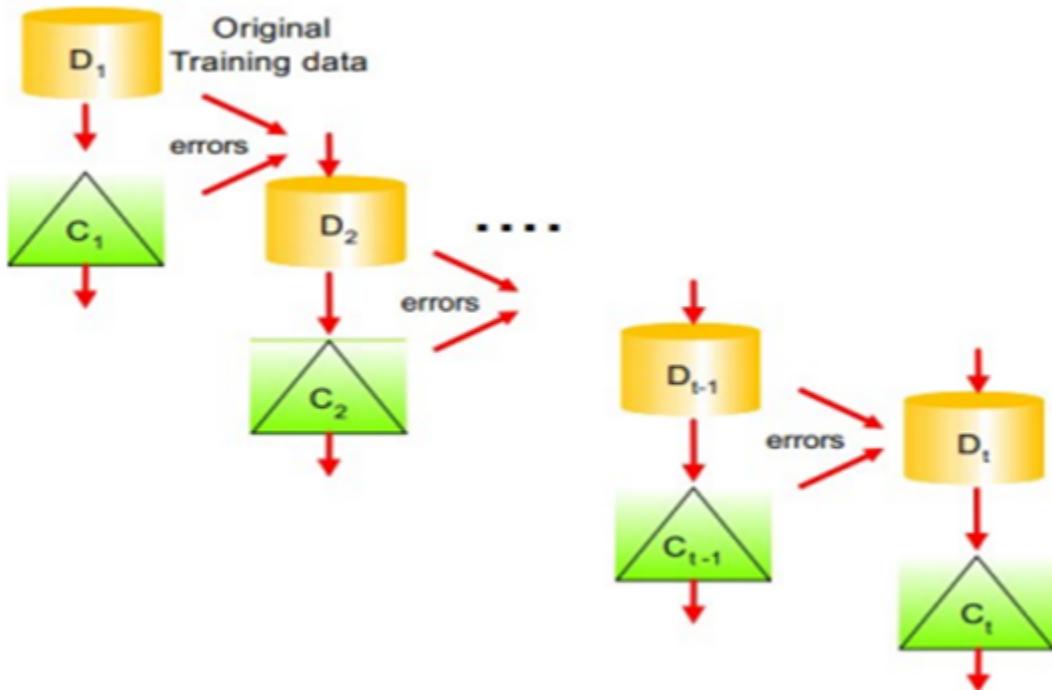
- Natural language processing (NLP) task that involves assigning a specific grammatical category (such as noun, verb, adjective, etc.) to each word in a sentence.

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



- bagging

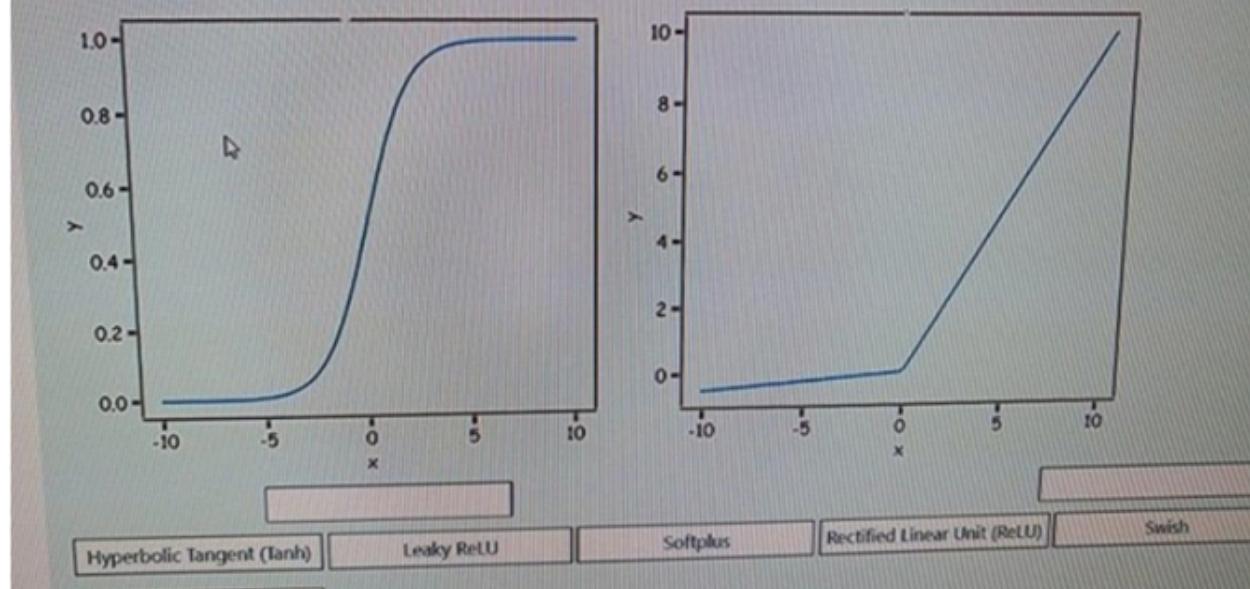
33. На прикажана слика е дадена шема за кој вид на учење со ансамбли?



- boosting

34. На слика се прикажани кои активацииски функции?

На слика се прикажани кои активацииски функции?



- 1) Sigmoid 2) leaky ReLU