Вовед во науката за податоци

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

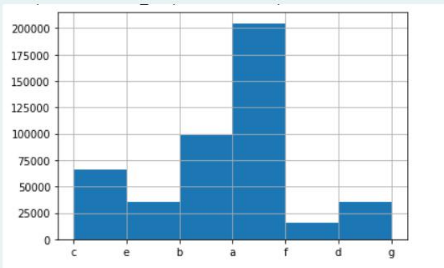
* + - Вредностите ќе бидат во опсегот помеѓу 0 и 1. \*\*\*
    - Средната вредност на податоците е 0 и варијансата е 1
    - Податоците следат нормална дистрибуција
    - Нивниот опсег е помеѓу минималниот и максималниот елемент во датасетот.
    - Нормализација ги праве вредностите меѓу 0 и 1 а стандардизација, со средна вредност 0 а варијанса 1

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

* + - Only for the categorical columns of the df dataset will the mean be printed
    - For each of the columns od the df dataset the mean value will be printed.
    - Only for the numeric columns of the df dataset the mean value will be printed. \*\*\*

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

* + - Frequency
    - Median value \*\*\*
    - Percent (row, column or total)
    - Mean value \*\*\*

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


* + - seaborn.displot(df[‘Bed Grade’], bins=3, kde=True, rug=True)
    - df[‘Hospital\_type\_code’].hist(bins = 3)
    - seaborn.distplot(df[‘Bed Grade’], bins=6, kde=True, rug=True)
    - df[‘Hospital\_type\_code’].hist(bins = 6) \*\*\*

**5. Ако се подели податочното множество на повеќе делови и потоа се остава едно за тестирање, а другите се користат за обука, за која техника на машинското учење станува збор.**

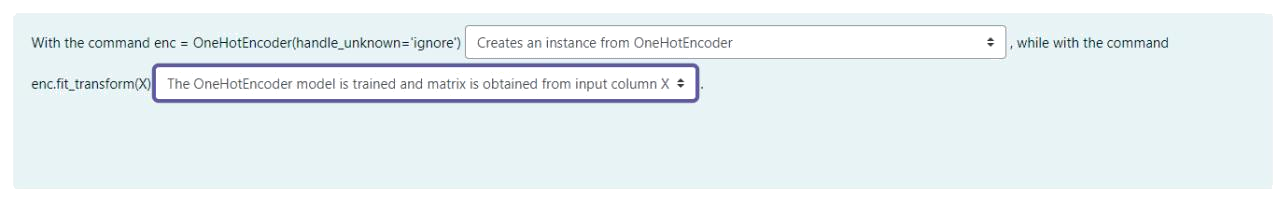
* + - Ласо регуларизација (LASSO Regularization)
    - Врстена валидација (Cross Validation) \*\*\*
    - Регуларизација по сртот (Ridge Regularization)
    - Ентропија

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

|  |  |  |
| --- | --- | --- |
| За R1 → 1 – ((2/7)^2 + (5/7)^2) = 0.408  За R2 → 1 – ((6/10)^2 + (4/10)^2) = 0.48  R1 → 7/17 \* 0.408 = **0.168**  R1 и R2 → 7/17 \* 0.408 + 10/17 \* 0.48 = 0.45 | **Class 1** | **Class 2** |
| R1 | 2 | 5 |
| R2 | 6 | 4 |

* + - 0.282
    - 0.45
    - 0.168
    - 0.5

**7.**



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

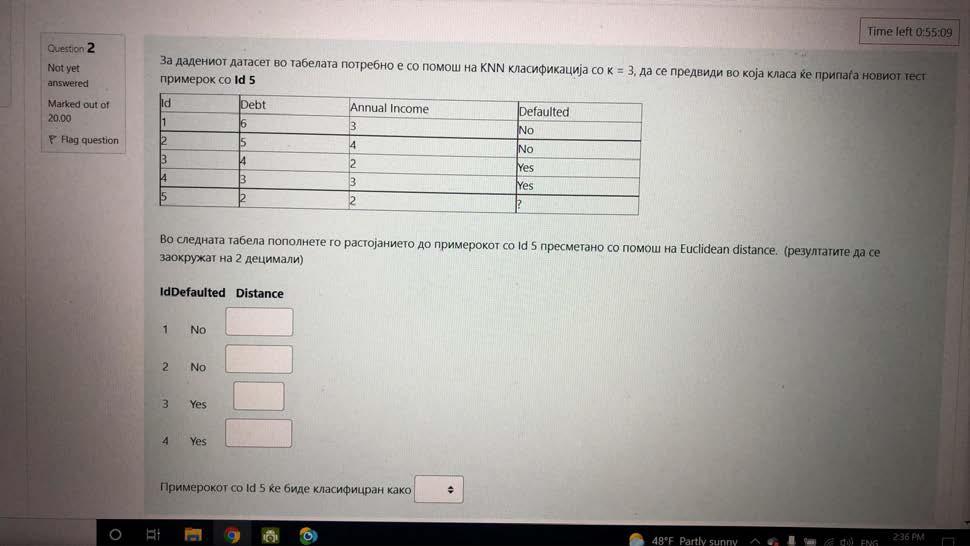
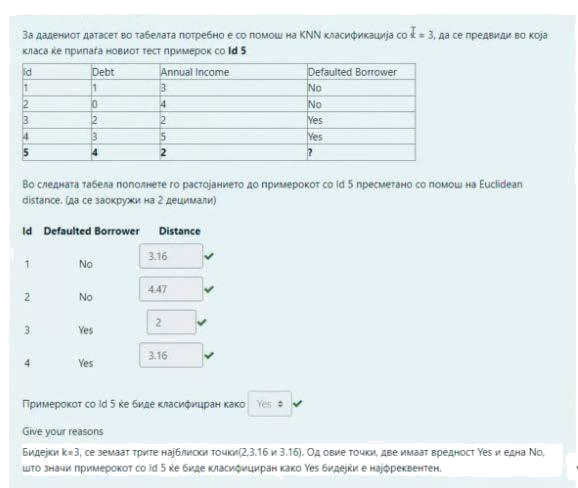
* + Visualization
  + Distance \*\*\*
  + Prediction of coefficients

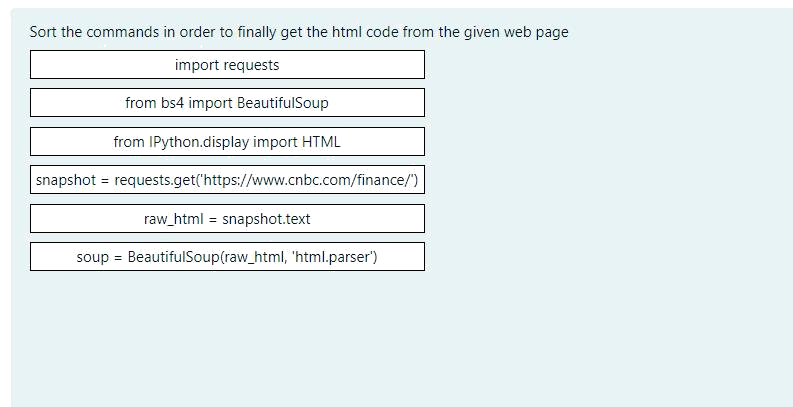
**9. Ако треба да се одреди припадноста на даден клиент во една од четирите групи на корисници, за каков вид на машинско учење станува збор?**

* + Класификација (Classification) \*\*\*
  + Откривање на недостатоците (Anomaly Detection)
  + Регресија (Regression)
  + Учење со поттикнување (Reinforcement Learning)

**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?**

* + - Средната вредност е поголема од медијаната \*\*\*
    - Медијаната е поголема од средната вредност
    - Средната вредност и медијаната се еднакви
    - Не може да се заклучи од дадениот графикa

**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.**  
  
 1. No → sqrt((6-2)^2 + (3-2)^2) = 4.12  
 2. No → sqrt((5-2)^2 + (4-2)^2) = 3.61  
 3. Yes → sqrt((4-2)^2 + (2-2)^2) = 2  
 4. No → sqrt((3-2)^2 + (3-2)^2) = 1.41  
 k=3 → #4, #3, #2 → No, Yes, No → No  
  
 1. No → sqrt((1-4)^2 + (3-2)^2) = 3.16  
 2. No → sqrt((0-4)^2 + (4-2)^2) = 4.47  
 3. Yes → sqrt((2-4)^2 + (2-2)^2) = 2  
 4. No → sqrt((3-4)^2 + (5-2)^2) = 3.16  
 k=3 → #3, #4, #1 → Yes, Yes, No → Yes

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


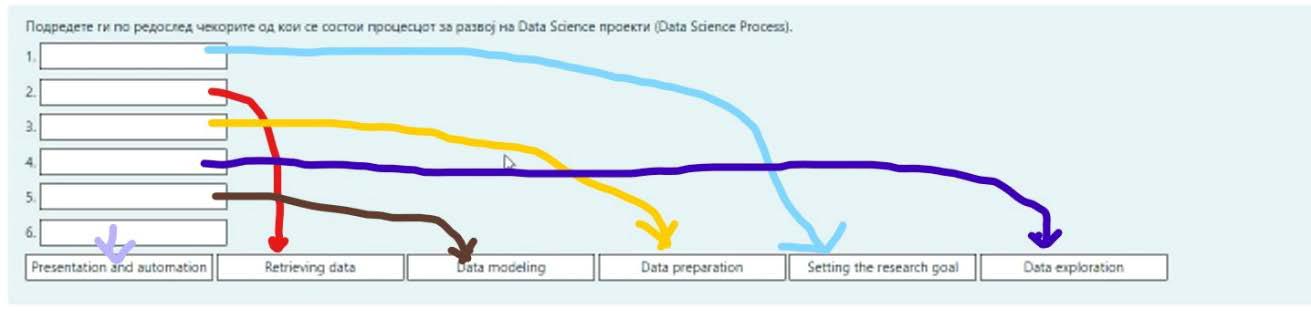
**13. Што резултат враќа дадениот код: df.isnull()**

* + - Целата табела (df) со True/False вредности во зависност дали на дадената позиција има/нема NAN вредност \*\*\*
    - Целата табела (df) само со позициите каде има NAN вредност
    - Целата табела (df) само со позициите каде нема NAN вредност
    - Број на NAN вредности по колона

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

* + - Dot plot
    - Scatter plot
    - Histogram
    - Bar chatts \*\*\*

**15. 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. \*\*\*

**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.**

* + - metric = “entropy”
    - criterion = “entropy”
    - spitter = “entropy”

**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\_**

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

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

* + - Откривање на недостатоци (Anomaly Detection)
    - Учење со поттикнување (Reinforcement Learning)
    - Класиикација (Classification)
    - Регресија (Regression) \*\*\*

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

* + - Мала наклонетост и мала варијанса
    - Голема наклонетост и мала варијанса \*\*\*
    - Мала наклонетост и голема варијанса
    - Голема наклонетост и голема варијанса

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

1. **- температура (децимални вредности)**
2. **- влажност на воздухот (децимални вредности)**
3. **- дали врнело во текот на денот (категирусја вредност → Yes / No)**
   * + Хистограм за секоја од колоните
     + Хистограм на целиот датасет
     + Хистограм за секоја од нумеричките колони \*\*\*
     + Ниту едно од наведените

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

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

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

* + - Податоците се во стандарден формат за реални броеви
    - Нивниот ранг сега е помеѓу минималниот и максималниот елемент во датасетот
    - Средната вредност на податоците е 0, а варијансата е 1 \*\*\*
    - Нивниот опсег е помеѓу 0 и 1

**24. Што се случува во дадениот код?   
url =** [**https://sitel.com.mk/**](https://sitel.com.mk/) **html = requests.get(url).text**

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

* + Ја враќа и прикажува веб страната
  + Ја парсира html содржината од веб страна “sitel.mk” \*\*\*
  + Ја симнува веб страната
  + Ниту едно од наведените

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

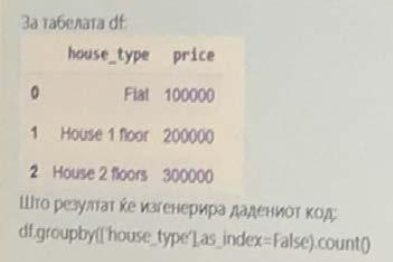
* + - Фреквенција \*\*\*
    - Средна вредност
    - Процент (редица, колона или вкупно) \*\*\*
    - Медијана

**26. Кој е излезот од дадениот код:  
d = [‘S’: ‘super’, ‘G’: ‘good’, ‘B’: ‘bad’]  
d[‘S’] = ‘SUPER’**

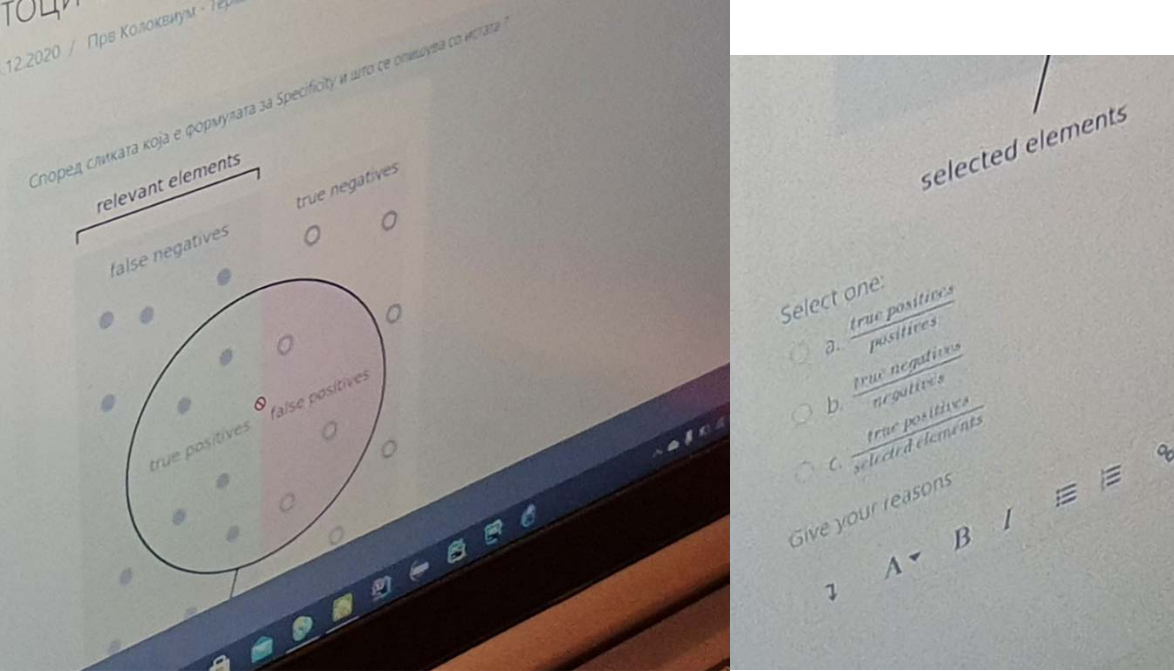
* + [‘S’: ‘super’, ‘G’: ‘good’, ‘B’: ‘bad’]
  + [‘S’: ‘SUPER, ‘G’: ‘good’, ‘B’: ‘bad’]
  + [‘S’: ‘super’, ‘G’: ‘good’, ‘B’: ‘bad’, ‘S’: ‘SUPER]

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

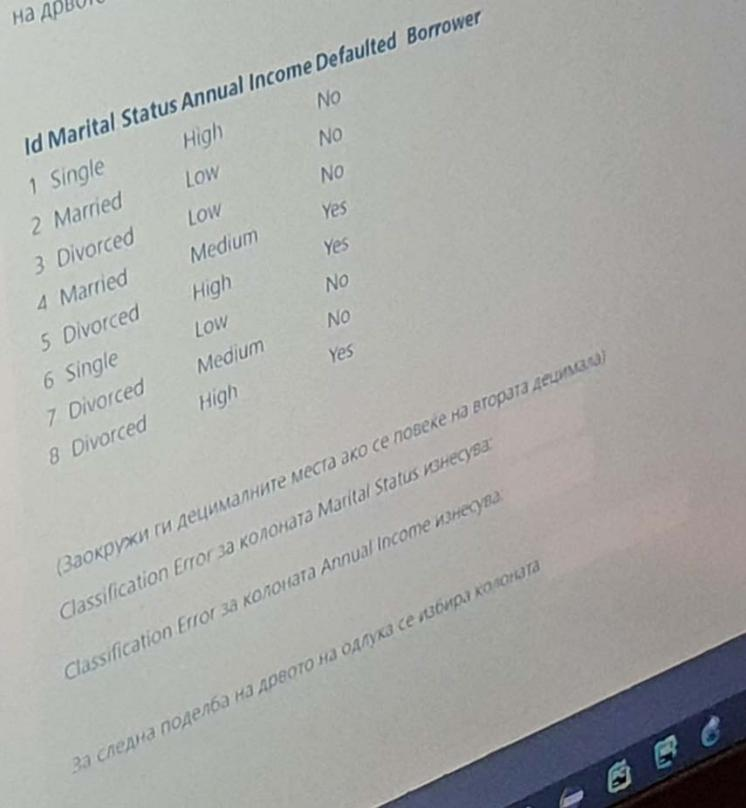
* + Откривање на недостатоците (Anomaly Detection)
  + Учење со поттикнување (Reinforcement Learning) \*\*\*
  + Класификација (Classification)
  + Регресија (Regression)

**28. За табелата df, што резултат ќе изгенерира дадениот код:   
df.groupby([‘house\_type’], as\_index=False).count()**

* + Средната вредност на типот на куќи
  + Бројот на куќи
  + Бројот на инстанци за секој тип на куќа \*\*\*
  + Ниту едно од наведените

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


* + - true positives / positives (this is sensitivity)
    - true negatives / negatives \*\*\*
    - true positives / selected elements

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

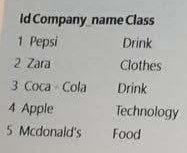
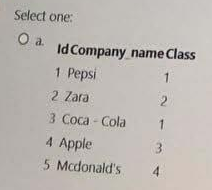
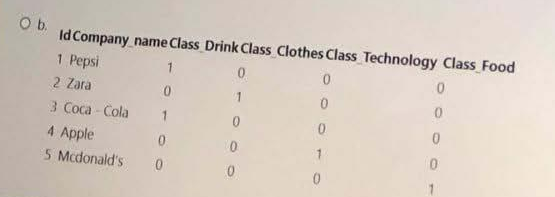
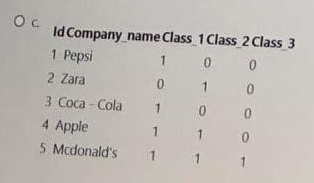
|  |  |  |  |
| --- | --- | --- | --- |
| **Marital Status** | **Yes** | **No** | **Classification Error** |
| Single | 0 | 2 | 1 - Max{0/2,2/2} = 0 |
| Married | 1 | 1 | 1 - Max{1/2,1/2} = 0.5 |
| Divorced | 2 | 2 | 1 - Max{2/4,2/4} = 0.5 |

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

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

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

Се избира **Annual Income**

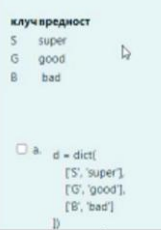
**31. За дадената табела во прилог ако се енкодира колоната Class со помош на One-Hot Encoding како ќе изгледа ново добиената табела?**  
 a. (label encoding)   
   
 b. (one-hot encoding)   
   
 c. (binary encoding)  
 

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

* + - Ако бројот на примероци што припаѓаат на дадена класа го надмине дозволениот број
    - Ако сите примероци во јазелот припаѓаат на истата класа \*\*\*
    - Ако бројот на циклуси надмине даден праг
    - Ако бројот на примероци во под-јазлите спадне под даден праг (min\_samples\_leaf) \*\*\*
    - Ако бројот на јазлите во дрвото надмине даден праг \*\*\*

**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?**  

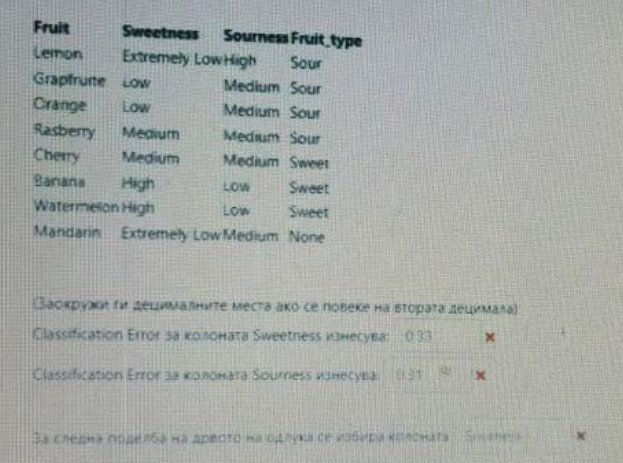

* + - Модел од машинското учење
    - Линеарен модел \*\*\*
    - Детерминантен модел
    - Колинеарен модел
    - Нелинеарен модел

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

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\_typeе таргет колона т.е.**

**по неа се врши класификацијата).**  
  
Solution:

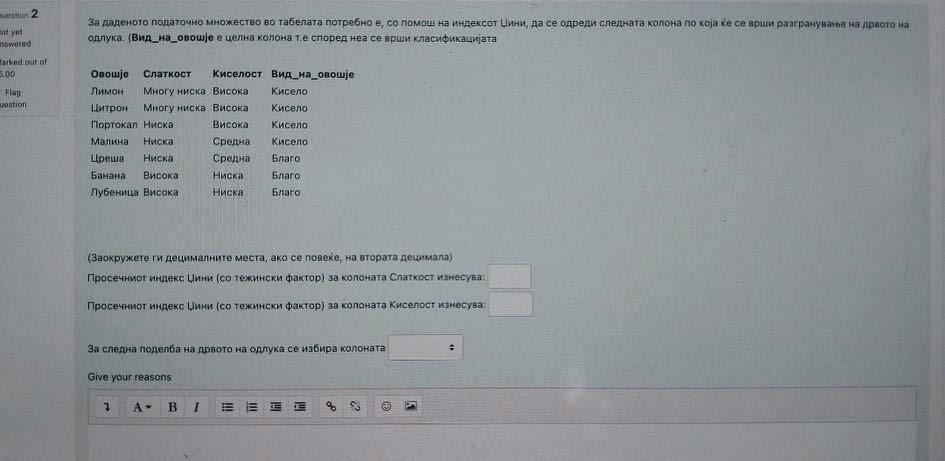
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sweetness** | **Sour** | **Sweet** | **None** | **Classification Error** |
| Extremely low | 1 | 0 | 1 | 1 – Max{1/2,0/2,1/2} = 0.5 |
| Low | 2 | 0 | 0 | 1 – Max{2/2,0/2,0/2} = 0 |
| Medium | 1 | 1 | 0 | 1 – Max{1/2,1/2,0/2} = 0.5 |
| High | 0 | 2 | 0 | 1 – Max{0/2,2/2,0/2} = 0 |

0.5\*2/8 + 0\*2/8 + 0.5\*2/8 + 0\*2/8 = **0.25**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sourness** | **Sour** | **Sweet** | **None** | **Classification Error** |
| Low | 0 | 2 | 0 | 1 – Max{0/2,2/2,0/2} = 0 |
| Medium | 3 | 1 | 1 | 1 – Max{3/5,1/5,1/5} = 0.4 |
| High | 1 | 0 | 0 | 1 – Max{1/1,0/1,0/1} = 0 |

0\*2/8 + 0.4\*5/8 + 0\*1/8 = **0.25**

Се избира **???**

**36.**Solution:

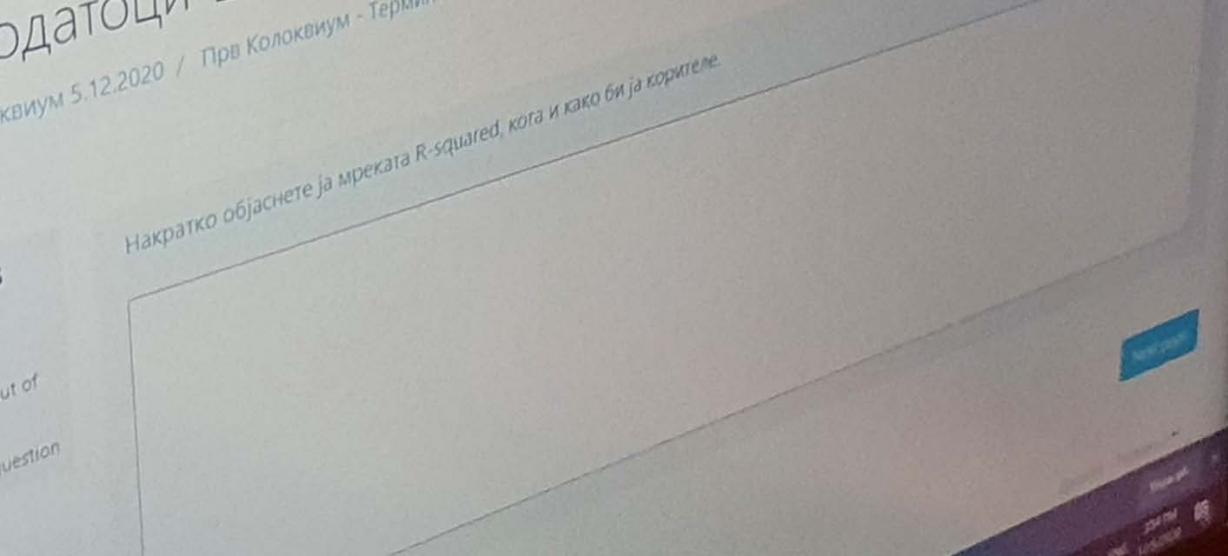
|  |  |  |  |
| --- | --- | --- | --- |
| **Слаткост** | **Кисело** | **Благо** | **Gini Index** |
| Многу ниска | 2 | 0 | 1 - {(2/2)^2+(0/2)^2} = 0 |
| Ниска | 2 | 1 | 1 - {(2/3)^2+(1/3)^2} = 0.56 |
| Висока | 0 | 2 | 1 - {(0/2)^2+(2/2)^2} = 0 |

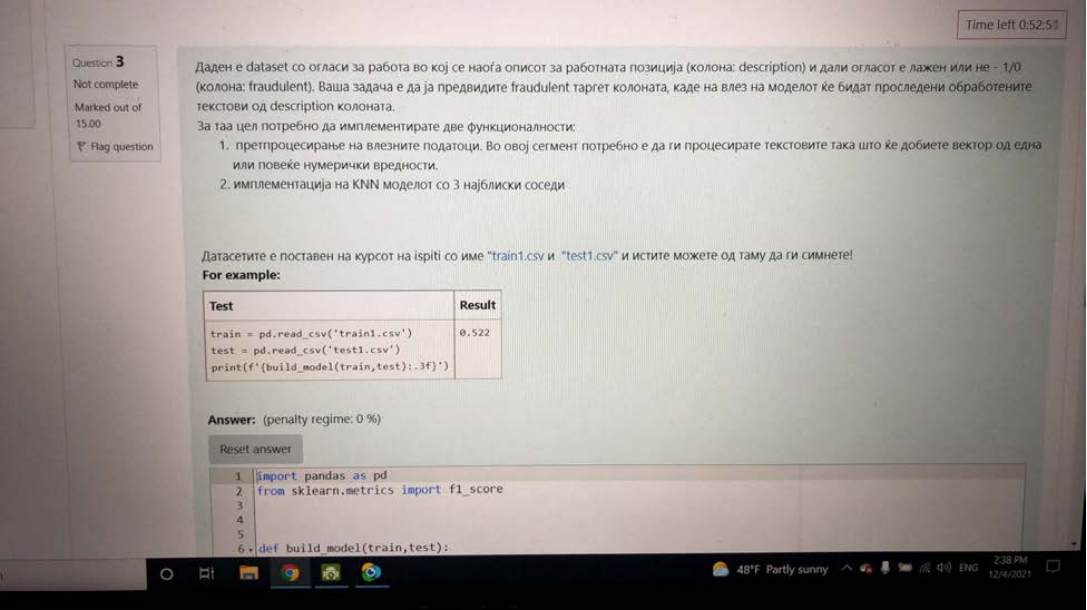
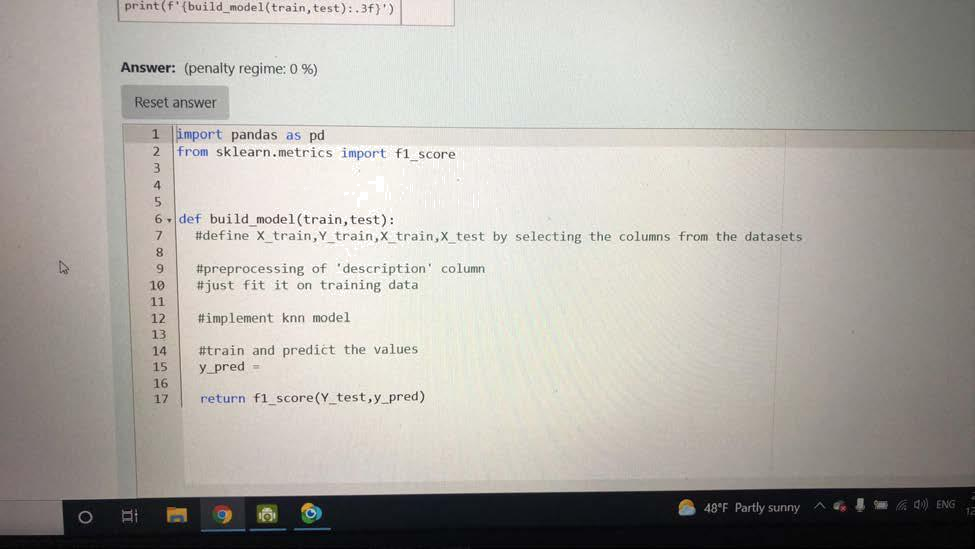
0\*2/7 + 0.56\*3/7 + 0\*2/7 = 0.23809... ~ **0.24**

|  |  |  |  |
| --- | --- | --- | --- |
| **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... ~ **0.14**

Се избира **Киселост**

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

**38.** **  
**

#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

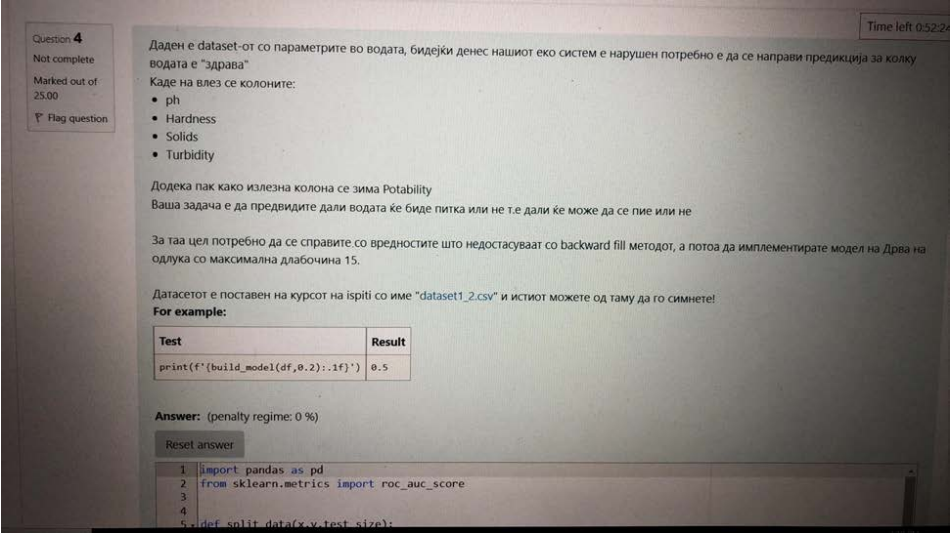
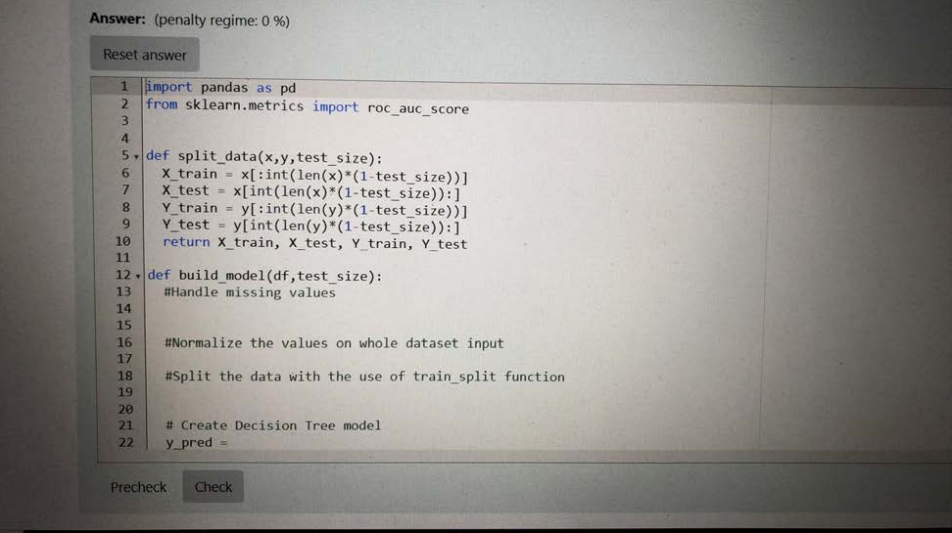
scaler = StandardScaler()  
x\_train['description'] = scaler.fit\_transform(x\_train['description'])

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.**  
  


#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.**  
