Modules

Grades

Inbox

Calc

○ CoC & Instructions

○ Pre_Processing

● Pre_Processing Assessment
Assignment

○ Model_Building

Pre_Processing Assessment

The due date for submitting this assignment has passed.
Due on 2024-04-13, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2024-04-07, 18:54 IST

Prepare the dataset

Dataset description

Gender: Gender of the patient.
HasTension: If the patient is known to have hyper tension.
Age: Age of the patient in years.
AnyHeartDisease: If the patient is known to have any cardio pathy.
NeverMarried: Status of the patient if he/she was never married.
Occupation: Occupation or job type of the patient.
LivesIn: If the patient lives in a city or village.
GlucoseLevel: Randomly sampled glucose level of a patient.
BMI: Body mass index.
SmokingStatus: How frequent a patient smokes or if he has smoked before.
HeartAttack: If the patient has a cardiac arrest before.

Note: For numerical type questions, always enter the answer correct upto 3 decimal places without rounding off, unless otherwise stated.

Preamble: Load the dataset and examine it.

Click here to view the sklearn library reference

Click here to view the Colab File

Click here to view the Questions.MD file

1) Which dataset are you using for this exam? Write the last two letters of the dataset file name. ***0 points***
Hint: the version number will be last two letters of the dataset file name.

Click here to view the Dataset

○ V1

● V2

○ V3

Yes, the answer is correct.
Score: 0

Accepted Answers:

V2

2) What is the total number of missing or unknown values in the column *Gender*?

(Hint: carefully look at the values the feature takes and find out implausible value(s).)

5

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Numeric) 5

2 points

3) What is the total number of missing or unknown values in the column *Age*?

(Hint: carefully look at the values the feature takes and find out implausible value(s).)

9

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Numeric) 9

2 points

4) What is the total number of missing or unknown values in the column GlucoseLevel?

(Hint: carefully look at the values the feature takes and find out implausible value(s).)

6

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Numeric) 6

2 points

5) What is the total number of missing or unknown values in the column *LivesIn*?

(Hint: carefully look at the values the feature takes and find out implausible value(s).)

8

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Numeric) 8

2 points

6) What is the total number of missing or unknown values in the column *BMI*?

(Hint: carefully look at the values the feature takes and find out implausible value(s).)

155

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Numeric) 155

2 points

7) What is the total number of missing or unknown values in the column *SmokingStatus?*

(Hint: carefully look at the values the feature takes and find out implausible value(s).)

1234

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Numeric) 1234

2 points

8) What is the mean value of the *BMI* in the dataset? Ignore the missing values if any

28.937

Yes, the answer is correct.
Score: 3

Accepted Answers:

(Type: Range) 28.932,28.942

3 points

9) How many people live in city, smoked at least once in life and had a heartattack? Ignore records/rows with any missing values. ***3 points***

○ 52

● 53

○ 54

○ 55

○ 56

○ None of these

Yes, the answer is correct.
Score: 3

Accepted Answers:

53

10) Which of the following categories have highest frequency? Ignore rows with missing values. ***4 points***

● female patients without tension, without any heart disease and never married

○ female patients without tension, without any heart disease and either currently married or married before

○ male patients without tension, without any heart disease and never married

○ male patients with tension, with a heart disease and never married

○ There is a tie between 2 or more options.

Yes, the answer is correct.
Score: 4

Accepted Answers:

female patients without tension, without any heart disease and never married

11) Select columns with categorical values : ***2 points***

☑ Gender

□ BMI

☑ NeverMarried

☑ SmokingStatus

□ GlucoseLevel

☑ HeartAttack

□ None of these

Yes, the answer is correct.
Score: 2

Accepted Answers:

Gender
NeverMarried
SmokingStatus
HeartAttack

12) 'HeartAttack' is the target column. What is the distribution count of "No" and "Yes" classes? ***2 points***

○ 3025, 455

○ 3797, 203

○ 3806, 194

● 3804, 196

○ None of these

Yes, the answer is correct.
Score: 2

Accepted Answers:

3804, 196

13) Divide the data into training and test sets
Keep 30% of the data as test set.
Use random_state as 0
HeartAttack is the target, rest of the columns are the features.
For the label/target vector, replace "Yes" with 1 and "No" with 0.
Divide the dataset into training and test sets keeping target(y) in stratified manner.

2 points

Hint: look for the documentation of the usual function that divides the data into training and test datasets.

Prepare a data preprocessing pipeline to process features in following order:

Gender: Impute with most frequent then ordinally encode.
Age: Impute with mean then standard scale.
HasTension: Ordinally encode.
AnyHeartDisease:Ordinally encode.
NeverMarried:Ordinally encode.
Occupation: One hot encode.
LivesIn: Impute with most frequent then ordinally encode.
GlucoseLevel: Impute with mean, then min-max scaling.
BMI: Impute with mean, then standard scale.
SmokingStatus: Impute with most frequent, then one hot encode

Hint: After transformation, your feature matrix must have columns in following order:

0. Gender
1. Age
2. HasTension
3. AnyHeartDisease
4. NeverMarried
5. Occupation_Govt_job
6. Occupation_Never_worked
7. Occupation_Private
8. Occupation_Self-employed
9. Occupation_children
10. LivesIn
11. GlucoseLevel
12. BMI
13. SmokingStatus_formerly smoked
14. SmokingStatus_never smoked
15. SmokingStatus_smokes

NOTE:

Make sure to preprocess the features in the above order exactly. Answer(s) of later question(s) depend(s) upon correct order of featuring processing.
You may have to use multiple instances of a trasnformer for this question.

Calculate the shape of the feature matrix of training dataset.

● (2800, 16)

○ (2000, 16)

○ (3000, 16)

○ (3200, 16)

Yes, the answer is correct.
Score: 2

Accepted Answers:

(2800, 16)

14) What is the mean of the transformed test data (features only)?

Note : Compute the mean of the whole feature matrix i.e. mean of all values in the transformed test feature matrix

0.2600

No, the answer is incorrect.

Score: 0

Accepted Answers:

(Type: Range) 0.245,0.255

6 points

15) If you eliminate 1 feature with recursive feature elimination, which feature will be eliminated?

Type the index of the eliminated feature (index starts from 0).
Use LogisticRegression model with random state as 1729 and rest of the parameters with default values, as an estimator.
Use processed training data.

10

No, the answer is incorrect.

Score: 0

Accepted Answers:

(Type: Numeric) 6

6 points