Preamble: Load the dataset and examine it.

Notes:

- This exam consists of a Regression problem.
- The target feature is 'actual_productivity'.
- Random state should be taken as 32 wherever applicable.

Q1. [marks : 0] Which dataset are you using for this exam?

Options:

A) v1

B) v2

C) v3

D) v4

E) v5

Answer: v1: A, v2:B, v3: C, v4:D, v5:E

Preprocessing Questions

```
rs = 32

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
from google.colab import files
files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in
Please rerun this cell to enable.

```
{}
```

Q.2 [Marks: 2] How many total number of features (Excluding target variable) are there in the dataset?

(a) 1000

(b) 14

(c) 13

(d) 12

Answer: V1,V2, V3, V4, V5: Option (d)

Q3: [marks: 2] Which of the following column have missing values ?[MCQ]

Options:

A)actual_productivity

B) department

C) incentive

D) wip

E) All of these

Answer: V1,V2, V3, V4, V5 : D

Q.4 [Marks: 2] What are the unique `department` mentioned in the dataset?

A) ['sewing' , 'finishing']

B) ['finishing', 'sweing', 'finishing ']

C) ['sampling', 'sweing']

D) ['purchasing','sampling','finishing']

Answer: V1,V2, V3, V4, V5: **option B**

Q.5 [Marks: 2] Which of the following columns have categorical data: -

A) Team

B) wip

C) department

D) incentive

Answer V1,V2, V3, V4, V5 : **Option C**

Q.6 [Marks: 4] Replace all NaN value in "wip" column by 0 and Check how many outliers are there in "incentive" column (Consider any incentive value >1000 as outlier) ?

A) 3

B) 6

C) 12

D) 5

Answer: V1,V2,V4: **option D** V3,V5: **option A**

Q7 [Marks = 4] Break the dataset into X and y, where the column "actual_productivity" goes to y and rest of the columns go to X. Enter the avg value of "actual_productivity" column? [NAT]

Answer:

V1, : 0.732 Range: [0.72-0.74]

V2 : 0.736 Range: [0.72-0.74]

V3: 0.733 Range: [0.72-0.74]

V4: 0.733 Range: [0.72-0.74]

V5: 0.738 Range: [0.72-0.74]

Q8.[Marks: 4] Plot the heatmap and mark the pair which has highest correlation value.

A) Actual_productivity & Targeted_productivity

B) smv & overtime

C) No of workers & overtime

D) over_time & targeted_productivity

E) No of workers & smv

F) wip & no of workers

Ans: V1, V2,V3,V4,V5 : **Option E**

Common Instructions for Question 9 and 10

Step 1: Drop the date column from X data.

Step 2: We can see in "department" column wrong spelling of "Sewing" is given. Update it to "Sewing" in X dataset.

Step 3: At some places in "department" column "finishing" variable is written with an extra space So, replace "finishing " with "finishing" in X dataset.

Step 4 Use `pd.getdummies` to encode all categorical features of X dataset.

Q9[Mark 3] What is the total number of "finishing" label in "department column of X dataset?

A) 586

B) 414

C) 418

D) 424

E) 429

Ans: V1: B, V2: C, V3:C , V4: D , V5: E

Q10[Mark-3] What is the total number of columns in X dataset now?

A) 12

B) 11

C) 13

D) 14

E) 15

Ans: V1,V2,V3,V4,V5 : A,

Q11 [Marks : 3] Split the dataset into training and test dataset using train test split into `70:30` ratio while keeping `random_state =32.` what is the shape of the `X_train` dataset?

(a) (700, 11)

(b) (700, 13)

(c) (700,12)

(d) (700, 14)

Ans: V1, v2, v3, v4 , v5: C

Q12 [Marks: 3] Apply LinearRegression on the 'x_train' and 'y_train' data. Calculate the score on the 'x_test' and 'y_test'. Which of the following options represent the calculated R^2 score.

(a) 0.272

(b) 0.2743

(c) 0.231

(d) 0.301

(e) 0.250

Ans: V1: (b), V2: (a), V3: (d) , V4: (c) , V5: (e)

Q13 [Marks: 6]Using the Linear regression model, compute the cross-validation scores for 9 splits on training data (`x_train` and `y_train`) using `cross_val_score`.Enter the maximum value of R^2 score obtained upto four decimal places.[NAT]

(**Hint:** By default `cross_val_score` uses `LinearRegression`'s scoring metric, which is R^2 score.)

Ans:

V1: 0.4214 (Range: 0.41-0.43)

V2: 0.3658 (Range: 0.35-0.38)

V3: 0.359 (Range: 0.34-0.37)

v4: 0.445 (Range: 0.42- 0.46)

V5: 0.391 (Range: 0.37- 0.42)

Q14 [Marks 5] Apply `SequentialFeatureSelector` transformer with `LinearRegression()` estimator and select 5 features by fitting to the `x_train` and `y_train.`

Which of the following option represents the correct integer index,of selected features list?

(a)[0 1 2 8 9]

(b) [0 1 5 7 8]

(c) [0 1 6 7 8]

(d) [0 1 2 8 9]

(e) [0 1 3 7 8]

Ans: v1: (b) v2:(d) v3: (a) v4: (c) v5: (e)

Q 15:[Marks 5] Take `SelectKBest` feature selector with `k=5` and `mutual_info_regression` as scoring function and fit it to training data(`x_train` and `y_train`), then transform it.Which of the following options represent the 5 selected features using above instructions?

(a) 'targeted_productivity', 'over_time', 'incentive', 'no_of_workers'

(b) 'over_time', 'incentive', 'no_of_workers

(c) 'targeted_productivity', 'smv', 'over_time', 'incentive'

(d) 'targeted_productivity', 'smv', 'over_time', 'incentive', 'no_of_workers'

v1, v2,v3, v4, v5: Option(d)

Q16 [Marks :4]Apply Ridge regression `Ridge()` with default penalty value on `x_train` and `y_train` and calculate the R^2 score on

`x_train` and `y_train`. Which of the following option represents the correct score (Upto 4 digits after decimal points)?

(a) 0.2907

(b) 0.2938

(c) 0.2675

(d)0.2555

(e) 0.2708

Ans: V1: Option (b), v2: Option (d), v3: Option (c), v4: Option (a), v5: Option (e)

Q 17: [Marks 5] Apply `Lasso` regression with `alpha=0.1` on the training data. Enter the value of the intercept you got correctly upto 4 digits after decimal points .

V1:0.7448 Range: 0.74-0.75

V2: 0.7556 Range: 0.75-0.76

v3: 0.7641 Range: 0.76-0.77

v4: 0.7472 Range: 0.74-0.75

v5: 0.7543 Range: 0.75-0.76

Q18 [Marks 5] Fit `SGDRegressor()` estimator with `random_state=32` on the training data and predict `actual_productivity` (`y_test`) for test data. The parameters are initialized with default values. Calculate and mark the correct `mean_squared_error` value between `y_test` and predicted `y_test` from the given options.

(a) 6.9888e+32

(b) 2.6858e+32

(c) 9.9176e+31

(d) 1.3689e+33

(e)1.4773e+31

Ans: V1: option (e) v2: option (c) v3: option (d) v4: Option (b) v5: Option (a)

Q19: [Marks 6] Apply cross validation strategy on training data using `SGDRegressor(random_state=32)` as an estimator and cv= `ShuffleSplit`. Set paramters for `ShuffleSplit` as following:

- `n_split` to be taken as 8.
- `test_size` to be taken as 40%.
- `random_state` value to be taken as 32

Mark the correct mean value of `cross_val_score` obtained upto four decimal places from the given options.

(a) -2.5183e+34

(b) -1.7301e+34

(c) -1.6280e+34

(d) -2.3437e+34

(e) -5.1595e+34

Ans:

v1: Option (e) v2: Option (d) v3: option (c), v4: Option (b) v5: Option (a)

(Common Instructions for Question 20 and 21)

Create a pipeline Using PolynomialFeatures as transformer and Lasso as estimator. Use GridSearchCV with this created pipeline and following hyperparameter values on training data(x_train, y_train) to fit the model .

i) Keep polynomial degree as : [1, 2, 3]

ii) alpha value to be taken as : np.logspace(-3, 0, num=8)

iii) scoring : neg_mean_absolute_error .

(**Note:** Kindly ignore the warning.)

Q20 [6 Marks] Enter the best alpha value you got using above instructions. (Mark correctly upto 4 decimal places)[NAT]

Ans:

v1: 1.0 Range: 0.99-1.01

v2: 0.019 Range: 0.015- 0.025

v3: 0.0026 Range: 0.002- 0.003

v4: 0.0026 Range: 0.002- 0.003

Q21 [5 Marks] Enter the best polynomial degree value you got using above instructions.[NAT]

Ans:

v1: 2

v2: 2

v3: 2

v4 2

v5: 1

(Common Instructions for Question 22 and 23)

Fit the PCA model using following parameter values on training data and apply the dimensionality reduction on it(training data).

- n_components=5
- svd_solver='full'
- whiten=True
- random_state=32

Q22: [Marks 4] What is the sum of explained variance ratio corresponding to each of the selected components.

(a) 0.99

(b) 0.87

(c) 0.75

(d) 0.64

(e) 0.27

Ans: v1: (A), v2: (A) v3: (A) , v4: (A) v5: (A)

Q23: [Marks 6] Use this transformed training data and `y_train` to fit the RidgeCV estimator model having alpha value as [0.1,0.01,1,0.005] .Enter the `best_score_` value you got for this model.[NAT]

V1:-0.0519 Range: -0.07,-0.04

V2: -0.029 Range: -0.04,-0.02

v3: -0.0309 Range: -0.04,-0.02

v4: -0.032 Range: -0.04, -0.02

v5: -0.0308 Range: -0.04,-0.02