QUIZ SUMMARY

A. Python Programming for data analytics III
   a. Why do we need Data Visualization?

➔ To share Findings, To explore data, it easier to detect patterns, trends, and outliers in groups of Data

   b. These are type of plots, except?

➔ Pyplot (true: Boxplot, Lineplot, pyplot)

Why do we need Data Visualization?

Select one:

- a. **To share findings**
- b. **To explore data**
- c. **All above are true**
- d. **It easier to detect patterns, trends, and outliers in groups of data**

The correct answer is: **All above are true**

These are type of plots, except..

Select one:

- a. **Boxplot**
- b. **Swarmplot**
- c. **Lineplot**
- d. **Pyplot**

The correct answer is: **Pyplot**

These are plotting libraries in Python, except..

Select one:

- a. **Pandas**
- b. **Seaborn**
- c. **Matplotlib**
- d. **Pyplot**

The correct answer is: **Pyplot**

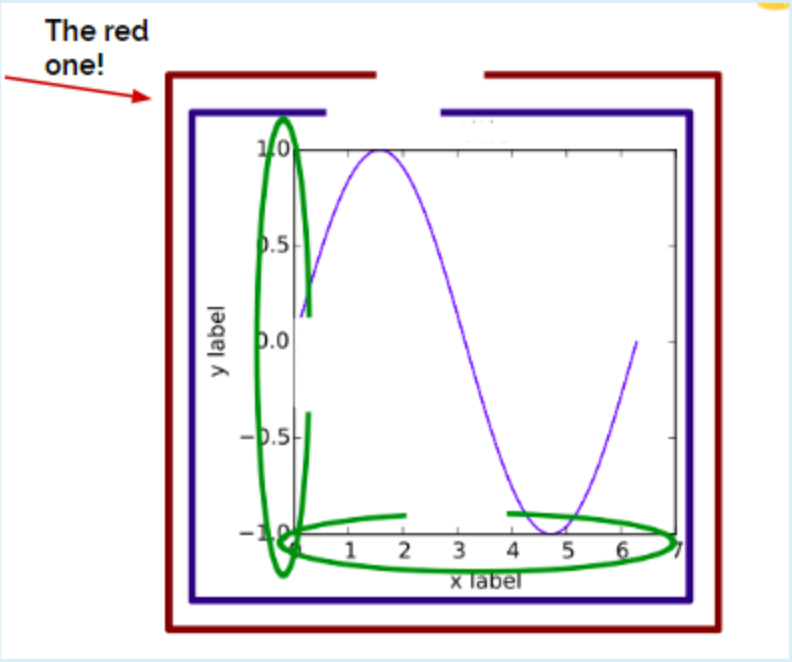
In Matplotlib, if we use matplotlib.plot(), what plot will be created?

Select one:

- a. **Barplot**
- b. **Lineplot**
- c. **Boxplot**
- d. **Histogram**

The correct answer is: **Lineplot**

In Matplotlib,we call this as?



Select one:

- a. **Figure**
- b. **Title**
- c. **Axis**
- d. **Axes**

The correct answer is: **Figure**

Let say we have DataFrame (df) like above, how to select all data in column A?

A	B	C	D
12	23	1	55
22	47	81	12
102	1230	777	712
8	1	21	99

Select one:

- a. **df[1]**
- b. **df.iloc[2:]**
- c. **df[0]**
- d. **df['A']**

The correct answer is: **df['A']**

What is the output of this code?

```
>>> spam = "3"
>>> spam = spam + "0"
>>> eggs = int(spam) + 7
>>> print(float(eggs))
```

Select one:

- a. 37.0
- b. 307.0
- c. 37

The correct answer is: 37.0

What is the output of this code?

```
>>> word = input("Enter a word: ")
Enter a word: chocolate
>>> print(word + ' factory)
```

Select one:

- a. **chocolate factory**
- b. **'chocolatefactory'**
- c. **"chocolate+factory"**

The correct answer is: **chocolate factory**

What does this code do?

```
for i in range(10):
  if not i % 2 == 0:
    print(i+1)
```

Select one:

- a. **Print all the odd numbers between 1 and 9**
- b. **Print all the even numbers between 2 and 10**
- c. **Print all the even numbers between 0 and 8**

The correct answer is: **Print all the even numbers between 2 and 10**

How many lines will this code print?

```
while False:
  print("Looping...")
```

Select one:

- a. 0
- b. **Infinitely many**
- c. 1

The correct answer is: 0

Rearrange the code to define a function that calculates the sum of all numbers from 0 to its argument.

A. def sum(x):

B. for i in range(x):

C. res += i

D. res = 0

E. return res

Select one:

- a. **A-C-D-B-E**
- b. **A-D-B-C-E**
- c. **E-A-C-D-B**
- d. **A-B-C-D-E**

The correct answer is: **A-D-B-C-E**

new_column
John, New York
David, Los Angeles
Chloe, Chicago
Emily, Houston
James, Philadelphia
Andrew, New York
Daniel, New York
Charlotte, Chicago
Samuel, San Diego
Anthony, Los Angeles

Manakah hasil dari query yang menghasilkan hasil seperti gambar di atas:

Select one:

- a. SELECT JOIN(FirstName, ',', city) AS new_column FROM customers;
- b. SELECT CONCAT(FirstName, ',', city) FROM customers;
- c. SELECT CONCAT(FirstName, ',', city) AS new_column FROM customers;
- d. SELECT SUBSTR(LastName, ',', city) AS new_column FROM customers;

SELECT SUBSTR(LastName, ',', city) AS new_column FROM customers;

The correct answer is: SELECT CONCAT(FirstName, ',', city) AS new_column FROM customers;

Table "Costumers"

ID	FirstName	LastName	City
1	John	Smith	New York
2	David	Williams	Los Angeles
3	Chloe	Anderson	Chicago
4	Emily	Adams	Houston
5	James	Roberts	Philadelphia
6	Andrew	Thomas	New York
7	Daniel	Harris	New York
8	Charlotte	Walker	Chicago
9	Samuel	Clark	San Diego
10	Anthony	Young	Los Angeles

Dari table costumres

Manakah dari query dibawah yang menghasilkan records yang berasal dari kota "New York, Los Angeles, San Diego"

Select one:

- a. SELECT * FROM customers WHERE city IN ('New York', 'Los Angeles', 'San Diego')
- b. SELECT * FROM customers WHERE city NOT IN ('Chicago, 'Houston', 'Philadelpia')
- c. SELECT * FROM customers WHERE city = 'New York' OR city='Los Angeles' OR city='San Diego'
- d. Semua benar

The correct answer is: Semua benar

Table "**Costumers**"

ID	FirstName	LastName	City
1	John	Smith	New York
2	David	Williams	Los Angeles
3	Chloe	Anderson	Chicago
4	Emily	Adams	Houston
5	James	Roberts	Philadelphia
6	Andrew	Thomas	New York
7	Daniel	Harris	New York
8	Charlotte	Walker	Chicago
9	Samuel	Clark	San Diego
10	Anthony	Young	Los Angeles

Dari table costumers, select kolom "id" dan "name" dari table "customers", munculkan 12 records, mulai dari records ke-5.

Select one:

- a. SELECT id, name FROM customers LIMIT 5 , 12;
- b. SELECT id, name FROM customers LIMIT 12 , 4;
- c. SELECT id, name FROM customers LIMIT 4 , 12;
- d. SELECT id, name FROM customers LIMIT 12 , 5;

The correct answer is: SELECT id, name FROM customers LIMIT 4 , 12;

Hasil dari query SQRT(4) adalah :

Select one:

- a. 16
- b. Semua salah
- c. 8
- d. 2

The correct answer is: 2

What does this code do?

```
from sklearn.preprocessing import OneHotEncoder
scaler = OneHotEncoder()
scaler.fit_transform(df)
```

Answear : Encoding Categorical Feature

Select one:

- True
- False

The correct answer is 'True'.

How to drop duplicate value in Pandas?

Select one:

- a. df.fillna()
- b. df.dropduplicates()
- c. df.duplicates()
- d. df.drop_duplicates()

The correct answer is:
df.drop_duplicates()

How to change integer column to string in Pandas?

Select one:

- a. df['ColumnA'].dtype('str')
- b. df['ColumnA'].type('str')
- c. df['ColumnA'].astype('str')

The correct answer is:
df['ColumnA'].astype('str')

How to change date = '2019-01-01' to datetime format?

Select one:

- a. datetime.date(date,'%Y-%m-%d')
- b. datetime.strptime(date,'%Y-%m-%d')
- c. datetime.strftime(date,'%Y-%m-%d')
- d. datetime.datetime(date,'%Y-%m-%d')

The correct answer is:
datetime.strptime(date,'%Y-%m-%d')

How to apply log function into row in pandas?

Select one:

- a. df.map(log)
- b. df.map(np.log)
- c. df.lapply(np.log)
- d. df.apply(np.log)

The correct answer is:
df.apply(np.log)

Which one is tool for workflow orchestration?

Select one:

- a. **Apache Airflow**
- b. **Apache Hive**
- c. **Apache Spark**

The correct answer is: **Apache Airflow**

Which one of the following is Cron expression for every Sunday at 7.15 PM?

Select one:

- a. **0 * * 7 15**
- b. **15 19 * * 0**
- c. **15 7 * * 0**

The correct answer is: **15 19 * * 0**

How to delete column in Spark DataFrame?

Select one:

- a. **df.remove("awesome_column")**
- b. **df.delete("awesome_column")**
- c. **df.drop("awesome_column")**

The correct answer is: **df.drop("awesome_column")**

What is the responsibilities that suits data engineer the most?

Select one:

- a. **Set up scheduled ingestion of data from the application databases to an analytical database**
- b. **Come up with a database schema for an application.**
- c. **Apply a statistical model to a large dataset to find outliers.**

The correct answer is: **Set up scheduled ingestion of data from the application databases to an analytical database**

What is the correct Spark function to transform table in the left to the right?

Thomas
Jimmy
Tina
Jackeline
Christine
Joseph
James
Cory

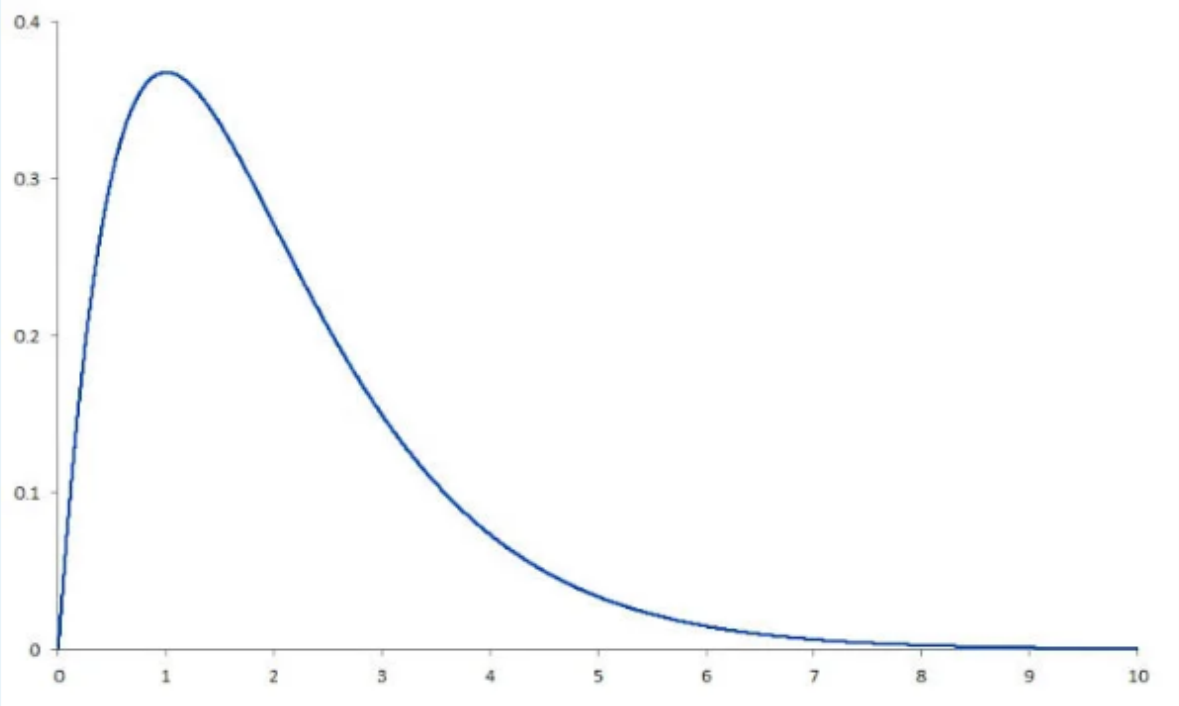
T	Thomas, Tina
J	Jimmy, Jackeline, Joseph, James
C	Christine, Cory

Select one:

- a. **df.groupby()**
- b. **df.extract()**
- c. **df.categorize()**

The correct answer is: **df.groupby()**

what term best to describe this chart?



Select one:

- a. Right-skewed Distribution
- b. Normal Distribution
- c. Left-skewed Distribution

The correct answer is: Right-skewed Distribution

Number of children in one household is an example of ordinal variable.

Select one:

- a. False
- b. True

The correct answer is: False

What is this formula for?

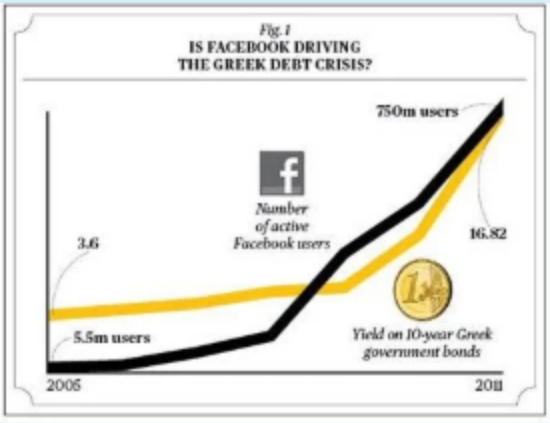
$$\sqrt{\frac{\sum(X-\bar{X})^2}{N}}$$

Select one:

- a. Median
- b. Standard deviation
- c. Mean
- d. Variance

The correct answer is: Standard deviation

Facebook growth caused the greek debt crisis.



Select one:

- a. False
- b. True

The correct answer is: False

The mode is the most appropriate measure of central tendency for which one of the following levels of measurement?

Select one:

- a. Ordinal
- b. Radio
- c. Interval
- d. Nominal

The correct answer is: Nominal

R squared in Regression measures ..

Select one:

- a. the explained sum of squares as a proportion of the Total Sum Of Squares
- b. the amount of variation in Y
- c. the correlation between X and Y
- d. the covariance between X and Y

The correct answer is: the explained sum of squares as a proportion of the Total Sum Of Squares

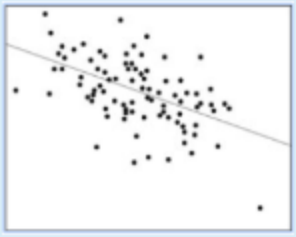
What is the probability of getting number 3 if we toss a fair dice once?

Select one:

- a. 1/6
- b. 3/6
- c. 2/6

The correct answer is: 1/6

Which statement best describe coefficient of correlation of this scatter plot?



Select one:

- a. -1<R<0
- b. 0<R<1
- c. R=-1
- d. R=1
- e. R=0

The correct answer is: **-1<R<0**

Which of the following coefficient of correlation indicates the strongest relationship

Select one:

- a. +0.45
- b. -0.33
- c. -0.67
- d. +0.58

The correct answer is: **-0.67**

Which is an example of Type I Error?

Select one:

- a. **Falsely concluding there is no effect when actually there is an effect**
- b. **Correctly concluding there is an effect**
- c. **Correctly concluding there is no effect**
- d. **Falsely concluding there is an effect when actually there is no effect**

The correct answer is: **Falsely concluding there is an effect when actually there is no effect**

A new automated drink machine is designed to dispense 530ml of liquid on the medium size setting. A restaurant owner suspects that his machine may be dispensing too much in medium drinks. She decide to take a sample of 303,030 medium drinks to see if the average amount is significantly greater than 530ml.

What are appropriate hypotheses for her significance test?

Select one:

- a. $H_0 : \mu = 530 \text{ mL}$, $H_a : \mu < 530 \text{ mL}$ (where μ is the average amount of liquid dispensed on this setting)
- b. $H_0 : \mu = 530 \text{ mL}$, $H_a : \mu > 530 \text{ mL}$ (where μ is the average amount of liquid dispensed on this setting)
- c. $H_0 : p = 530 \text{ mL}$, $H_a : p > 530 \text{ mL}$ (where p is the proportion of liquid dispensed on this setting)
- d. $H_0 : p = 530 \text{ mL}$, $H_a : p < 530 \text{ mL}$ (where p is the proportion of liquid dispensed on this setting)

The correct answer is: $H_0 : \mu = 530 \text{ mL}$, $H_a : \mu > 530 \text{ mL}$ (where μ is the average amount of liquid dispensed on this setting)

Using a significance level of alpha = 0.01, which of these is the most appropriate conclusion for this population of mobile phones?

Regression: Price vs. battery life

Predictor	Coef	SE Coef	T	P
Constant	191.312	94.318	2.028	0.045
Battery life	6.556	3.788	1.731	0.087

Select one:

- a. **This suggests a positive linear relationship between battery life and price because 0.087>0.01**
- b. **We can't conclude a positive linear relationship between battery life and price because 0.087>0.01**
- c. **We can't conclude a positive linear relationship between battery life and price because 0.045>0.01**
- d. **This suggests a positive linear relationship between battery life and price because 0.045>0.01**

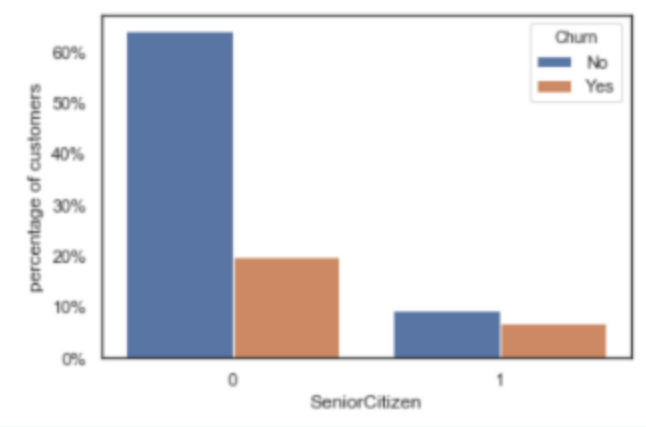
The correct answer is: **We can't conclude a positive linear relationship between battery life and price because 0.087>0.01**

Which one is the output of following syntax?

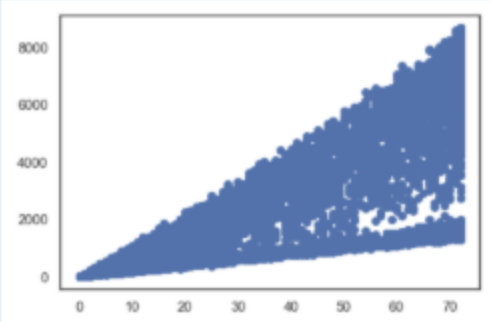
`seaborn.distplot(age)`

Select one:

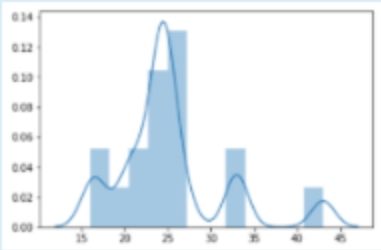
a.



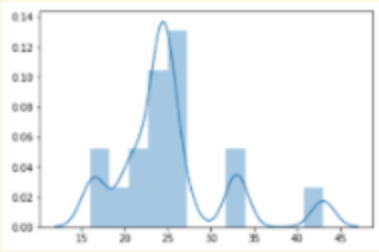
b.



c.



The correct answer is:



What the best way to handle a lot of missing value?

Select one:

- a. **Replace with mean**
- b. **Replace with mode**
- c. **Replace with median**
- d. **Remove**

The correct answer is: **Remove**

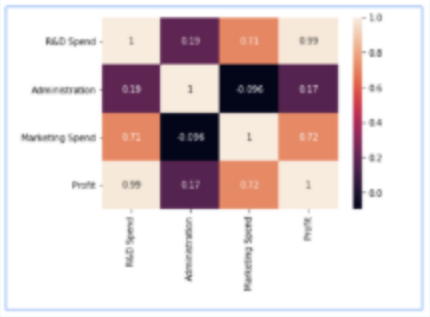
Which statistic is not included in data.describe()

Select one:

- a. **Median**
- b. **Third Quartile**
- c. **Mean**
- d. **Interquartile Range**

The correct answer is: **Interquartile Range**

Based on following chart, which variable correlate the most with Profit?



Select one:

- a. **Administration**
- b. **Marketing Spend**
- c. **Location of the company**
- d. **R&D Spend**

The correct answer is: **R&D Spend**

Based on following result, which one is our linear model equation?

```
print('coefficient of determination:', model.score)
```

```
coefficient of determination: 0.6825707454050335
```

```
print('intercept:', model.intercept_)
```

```
intercept: -188.71893615394083
```

```
print('slope:', model.coef_)
```

```
slope: [76.25473122]
```

Select one:

- a. y = 76.25 + 0.68x
- b. y = 0.68x - 188.72
- c. y = 0.68 - 188.72x
- d. y = -188.72 + 76.25x

The correct answer is: y = -188.72 + 76.25x

Which of the following sentence is FALSE regarding regression?

Select one:

- a. **It relates inputs to outputs**
- b. **It may be used for interpretation.**
- c. **It is used for prediction.**
- d. **It discovers causal relationships.**

The correct answer is: **It discovers causal relationships.**

What is Machine learning?

Select one:

- a. **The selective acquisition of knowledge through the use of computer programs**
- b. **The autonomous acquisition of knowledge through the use of computer programs**
- c. **The selective acquisition of knowledge through the use of manual programs**
- d. **The autonomous acquisition of knowledge through the use of manual programs**

The correct answer is: **The autonomous acquisition of knowledge through the use of computer programs**

You can solve this problem using supervised learning, except ..

Select one:

- a. **None of all**
- b. **Predict Stock Prices**
- c. **Classify Viruses**
- d. **Cluster behaviour of e-commerce user**

The correct answer is: **Cluster behaviour of e-commerce user**

Below are machine learning type, except ...

Select one:

- a. **Unsupervised Learning**
- b. **Natural Learning**
- c. **Supervised Learning**
- d. **Reinforcement Learning**

The correct answer is: **Natural Learning**

For applying machine learning, we need all of this, except

Select one:

- a. **Pattern**
- b. **Data**
- c. **Recursive Algorithm**
- d. **Can't derived mathematically**

The correct answer is: **Recursive Algorithm**

Berikut merupakan metode klasifikasi, kecuali :

Select one:

- a. **Linear Regression**
- b. **Logistic Regression**
- c. **K-Nearest Neighbor**
- d. **Naive Bayes**

The correct answer is: **Linear Regression**

Berikut merupakan metrik untuk regresi :

Select one:

- a. **Manhattan Distance**
- b. **Confusion Matrix**
- c. **Accuracy Score**
- d. **Mean Squared Error**

The correct answer is: **Mean Squared Error**

Berapa jumlah hyper-parameter dari metode decision tree ?

Select one:

- a. 2
- b. 1
- c. **Tidak ada**
- d. 3

The correct answer is: **Tidak ada**

Dibawah ini termasuk algoritma lazy learning adalah ..

Select one:

- a. **Decision Tree**
- b. **Random Forest**
- c. **Support Vector Machine**
- d. **k-Nearest Neighbors**

The correct answer is: **k-Nearest Neighbors**

Misalkan sebuah provider selular mengadakan promosi cashback pulsa 50% untuk pelanggan yang pertama kali isi pulsa.

Saya adalah pengguna provider tersebut lebih dari 2 tahun dan sudah beberapa kali isi pulsa, namun karena kesalahan klasifikasi sistem, saya mendapat cashback 50%, kesalahan tersebut merupakan ..

Select one:

- a. **False-Positive**
- b. **True-Positive**
- c. **True-Negative**
- d. **False-Negative**

The correct answer is: **False-Positive**

Berikut merupakan metode unsupervised learning :

Select one:

- a. **Dimensional Reduction**
- b. **Regression**
- c. **Classification**
- d. **Actor-Critic Algorithm**

The correct answer is: **Dimensional Reduction**

Berikut merupakan metode reduksi dimensi, kecuali :

Select one:

- a. **Auto-Encoder**
- b. **Deep Belief Network**
- c. **Principal Component Analysis**
- d. **Non-negative Matrix Factorization**

The correct answer is: **Deep Belief Network**

Berikut merupakan algoritma clustering, kecuali :

Select one:

- a. **K-Means**
- b. **Hierarchical Clustering**
- c. **DBScan**
- d. **Support Vector Machine**

The correct answer is: **Support Vector Machine**

Dibawah ini termasuk algoritma flat clustering, kecuali ..

Select one:

- a. **DBScan**
- b. **K-Means**
- c. **Gaussian Mixture Model**
- d. **Agglomerative Clustering**

The correct answer is: **Agglomerative Clustering**

Pada algoritma berikut, kita perlu menentukan jumlah clusternya

Select one:

- a. **Agglomerative Clustering**
- b. **DBScan**
- c. **Gaussian Mixture**
- d. **Meanshift**

The correct answer is: **Gaussian Mixture**

Below are example of business goals, expect...

Select one:

- a. **Market Basket Analysis**
- b. **Fraud Prevention**
- c. **Price Optimization**
- d. **Customer Satisfaction Improvement**

The correct answer is: **Market Basket Analysis**

Which one is defined as Marketing Endgame?

Select one:

- a. **AWARENESS**
- b. **ENGAGEMENT**
- c. **CONVERSION**
- d. **TRANSACTION**

The correct answer is: **TRANSACTION**

Which one IS NOT included in data solution proposal design?

Select one:

- a. **None of the above**
- b. **Determine the approach**
- c. **Data preprocessing**
- d. **Define business background**

The correct answer is: **Data preprocessing**

Below are list of processes to achieve goals, except ...

Select one:

- a. **Customer Segmentation**
- b. **Dynamic Pricing**
- c. **Productivity Improvement**
- d. **Product Recommendation**

The correct answer is: **Productivity Improvement**

Which one is less recommended to be included in your team when designing business solution?

Select one:

- a. **Data Project Manager**
- b. **Data Engineer**
- c. **Pricing Analyst**
- d. **Business Analyst**

The correct answer is: **Pricing Analyst**

Which one define the relationship between data governance and data management?

Select one:

a.
Both are unrelated subjects

b.
Both are related but has to implemented exclusively

c.
None of them are right

d.
Both are related and inseparable

The correct answer is:
Both are related and inseparable

Which one IS NOT included in elements of Data Governance?

Select one:

a.
None of the above

b.
Data quality

c.
Data strategy

d.
Data risk management

The correct answer is:
None of the above

What is the goal of Data Warehousing?

Select one:

- a. Achieve company's endgame
- b. Store all of the company's data from across services
- c. Generating business insights
- d. Store specific data for each business functions

The correct answer is:
Store all of the company's data from across services

Which one IS NOT included in BI Architecture Framework

Select one:

- a. Distribute
- b. Integrate
- c. Store
- d. Create

The correct answer is:
Create

What is the goal of Data Pipeline?

Select one:

a.
Create journey from raw data to the data warehouse

b.
Standardized data from various sources

c.
Extract data from various sources

d.
Load data to data warehouse

The correct answer is:
Create journey from raw data to the data warehouse

These are supervised learning, except?

Select one:

- a. **Regression**
- b. **Classification**
- c. **Clustering**

The correct answer is: **Clustering**

What should we do if we have a lot of feature?

Select one:

- a. **Tuning Hyperparameter**
- b. **Feature Engineering**
- c. **Feature Selection**
- d. **Gain model complexity**

The correct answer is: **Feature Selection**

Which is the evaluation metric for regression is?

Select one:

- a. **Accuracy**
- b. **Logloss**
- c. **F1-Score**
- d. **R-squared**

The correct answer is: **R-squared**

These are application for classification, except?

Select one:

- a. **Spam detection**
- b. **Face Recognition**
- c. **Voice Recognition**
- d. **Stock Prediction**

The correct answer is: **Stock Prediction**

If we have high bias, so?

Select one:

- a. **Sensitif terhadap data baru**
- b. **A dan C benar**
- c. **Model terlalu kompleks**
- d. **Model terlalu sederhana**

The correct answer is: **Model terlalu sederhana**

Which is the method of validations?

Select one:

- a. A,B, and C is correct
- b. **Leave one out cross validation**
- c. **Validation Set Approach**
- d. **k-Fold Cross Validation**

The correct answer is: A,B, and C is correct

How to make interactions for feature engineering, except?

Select one:

- a. **Multiplications**
- b. **Subset Selection**
- c. **Addition**
- d. **Substractions**

The correct answer is: **Subset Selection**

What is the evaluation metric for classification?

Select one:

- a. **Precision**
- b. **Mean squared error**
- c. **MAPE**
- d. **R-squared**

The correct answer is: **Precision**

What is the method that have the highest computational cost for feature selection?

Select one:

- a. **Backward Stepwise**
- b. **Forward Stepwise**
- c. **Hybrid approach (Backward & Forward)**
- d. **Best Subset Selection**

The correct answer is: **Best Subset Selection**

How to measure %Event in WOE

Select one:

- a. **failed event in specific category / all event in specific category**
- b. **success event in specific category / success event in all category**
- c. **success event in specific category / all event in specific category**

The correct answer is: **success event in specific category / all event in specific category**

Things should be considered in A/B Test?

Select one:

- a. Hypotesis
- b. All true
- c. Measurement
- d. Randomization

The correct answer is: All true

Which method can you use for regression ?

Select one:

- a. Decision Tree
- b. Random Forest
- c. XGBoost
- d. All true

The correct answer is: All true

In linear regression, what transformation should you use when your target variable is higly right skewed ?

Select one:

- a. Exponential transform(exp(y))
- b. All true
- c. Quadratic transform(y2)
- d. Log transform(log y)

The correct answer is: Log transform(log y)

Metrics used in regression is ?

Select one:

- a. f1-score
- b. Precision
- c. Mean squared error
- d. Recall

The correct answer is: Mean squared error

What is the purpose of randomization ?

Select one:

- a. All true
- b. Distributing co-variates evenly
- c. Remove selection bias

The correct answer is: All true

Berikut ini yang tidak termasuk komponen penting untuk dipahami untuk menyelesaikan business terkait dengan data warehouse modernization

Select one:

- a. Architecture
- b. Flow
- c. Operating System
- d. Data Tech Component

The correct answer is: Operating System

Data flow terdiri atas

Select one:

- a. OLAP & OLDP
- b. OLTP & OLDP
- c. OLAP & OLTP
- d. OLAP & BI

The correct answer is: OLAP & OLTP

Operational Database umumnya menggunakan flow..

Select one:

- a. OLAP
- b. OLTP
- c. OLL
- d. OLDP

The correct answer is: OLTP

Apache Kafka adalah aplikasi yang mendukung

Select one:

- a. Data point expansion
- b. Batch process
- c. Real-time streaming process
- d. Descriptive analytics

The correct answer is: Real-time streaming process

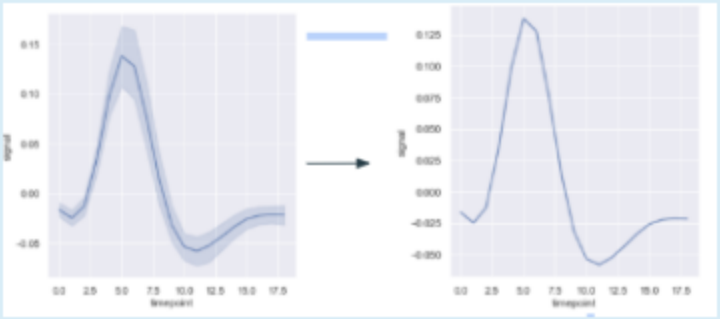
Data delivery guarantee dengan konsep "minimal satu kali" pengiriman adalah :

Select one:

- a. Best effort
- b. At least once
- c. Best Practice
- d. Exacity once

The correct answer is: At least once

What command to turn off the shades:



Select one:

- a. **shades = None**
- b. **shades = False**
- c. **ci = False**
- d. **ci = None**
- e. **conf.int = None**

The correct answer is: **ci = None**

Which statement is correct:

Select one:

- a. **Seaborn package requires Matplotlib package to run**
- b. **Matplotlib package requires Seaborn package to run**

The correct answer is: **Seaborn package requires Matplotlib package to run**

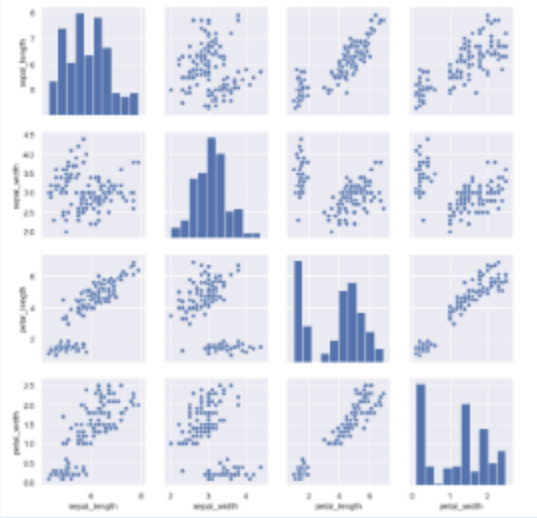
Which statement is wrong:

Select one:

- a. **Violin plot allows the user to identify bi-modal distribution**
- b. **Joint plot can show the distribution relationship up to 3 distributions**
- c. **Box plot helps the user to spot outliers**

The correct answer is: **Joint plot can show the distribution relationship up to 3 distributions**

What Seaborn function to produce plot like this?

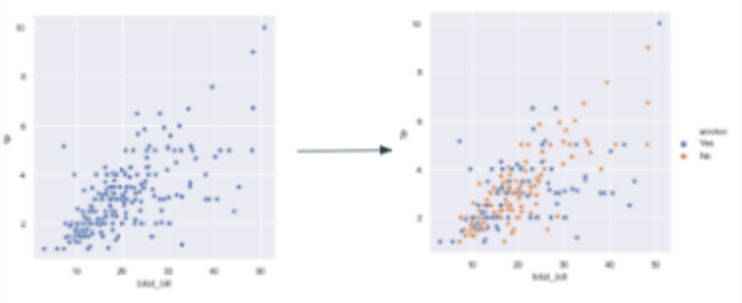


Select one:

- a. sns.facet_wrap()
- b. sns.jointplot()
- c. sns.pairplot()

The correct answer is: sns.pairplot()

What Seaborn command to color-code different groups?



Select one:

- a. hue
- b. kind
- c. col

The correct answer is: hue

sqoop import-all-tables -m 1 --connect jdbc:postgresql://host:port /postgres --username iykra_trainee --password passwordiykra --warehouse-dir user/hive/app_loan2 -- --schema training

Whats the directory of the output file from above syntax?

Select one:

- a. //host:port
- b. user/hive/app_loan2
- c. host:port/postgres

The correct answer is: user/hive/app_loan2

Why we need disable SElinux before we install cloudera?

Select one:

- a. SElinux will thread cloudera as virus
- b. No need to do that
- c. We need to disable the SElinux to enhance node performance while installing Hadoop

The correct answer is: SElinux will thread cloudera as virus

What is the diffrence between external and internal table?

Select one:

- a. The data from external table was not disappear if we drop the table, while the internal table was
- b. The data from internal table was not disappear if we drop the table, while the external table was
- c. Nothing, exactly the same

The correct answer is: The data from external table was not disappear if we drop the table, while the internal table was

Define sqoop?

Select one:

- a. Application for distributing file in hdfs
- b. Application for transfering bulk data in hadoop ecosystem
- c. Application for transfering bulk data between hadoop and non hadoop datastores

The correct answer is: Application for transfering bulk data between hadoop and non hadoop datastores

Whats the different between sqoop import and sqoop import-all-table?

Select one:

- a. Transfering all table from specific database on hadoop
- b. Transfering all databases on hadoop
- c. Transfering all table value from specific database non hadoop

The correct answer is: Transfering all table value from specific database non hadoop

Whats --hive-import argument really do?

Select one:

- a. Import tables into hive with default delimiters
- b. Import tables into hive with default delimiters if none are set
- c. Import tables into hive with specific delimiters

The correct answer is: Import tables into hive with default delimiters if none are set

sqoop import-all-tables -m 1 --connect jdbc:postgresql://host:port /postgres --username iykra_trainee --password passwordiykra --warehouse-dir user/hive/app_loan2 -- --schema training

Whats the database are imported?

Select one:

- a. app_loan2
- b. PostgresSQL
- c. postgres

The correct answer is: postgres

sqoop import -m 1 --connect jdbc:postgresql://host:port /postgres --username iykra_trainee --password passwordiykra --target-dir user/hive/app_loan2

Whats the output of above syntax?

Select one:

- a. Table in HIVE
- b. Error
- c. File that include all values in specific table
- d. Empty file

The correct answer is: Error

sqoop import-all-tables -m 8 --connect jdbc:postgresql://host:port /postgres --username iykra_trainee --password passwordiykra --warehouse-dir user/hive/app_loan2 -- --schema training

How much map task are used from above syntax?

Select one:

- a. 3
- b. 4
- c. 1
- d. 8

The correct answer is: 8

sqoop import- -m 8 --connect jdbc:postgresql://host:port /postgres --username iykra_trainee --password passwordiykra --target -dir user/hive/app_loan2 -table application_loan

If chosen schema have 4 tables on postgreSQL, how much the file part will be on the output directory?

Select one:

- a. 32
- b. 1
- c. 4
- d. 12

The correct answer is: 1

The following entity is included in tree, exclude:

Select one:

- a. Branch
- b. Node
- c. Leaf
- d. Probability

The correct answer is: Probability

Whats the meaning of 'pure dataset' in decision tree?

Select one:

- a. All answer are true
- b. Perfectly split based on "rules"
- c. Clean data
- d. Dynamic dataset

The correct answer is: Clean data

The following method is used to Attribute Selection Measures

Select one:

- a. Confidence Interval
- b. All answer are true
- c. P value
- d. Entropy

The correct answer is: Entropy

What is the parent node means?

Select one:

- a. All answer are true
- b. The train dataset
- c. The feaeture variable that we have to split
- d. The target variable that we want to predict

The correct answer is: The target variable that we want to predict

The advantage for using Decision Tree:

Select one:

- a. Need to use normalization
- b. Black box, hard to interpret model
- c. Can model non linear relationship

The correct answer is: Can model non linear relationship

Contoh hasil case study predictive analytics dibawah ini adalah..

Select one:

- a. Jumlah Penduduk Indonesia menurut Provinsi Tahun 2010
- b. Jumlah Penggunaan Obat Generik di Surabaya Tahun 2012
- c. Persebaran Order Ojek Online di Jakarta Tahun 2018
- d. Credit Risk Scoring Bank X untuk Produk Unsecured Loan

The correct answer is: Credit Risk Scoring Bank X untuk Produk Unsecured Loan

CRISP-DM adalah

Select one:

- a. Cross Industry Standard Process for Data Mining
- b. Cross Section Process for Data Mining
- c. Cross Industry Sectional Process for Data Mining
- d. Cross Tabulation Standard Process for Data Mining

The correct answer is: Cross Industry Standard Process for Data Mining

Step pertama dalam Business Analytics Workflow adalah

Select one:

- a. Data Presentation
- b. Deliver The Results
- c. Machine Learning
- d. Frame The Problem

The correct answer is: Frame The Problem

Berapa jumlah steps yang ada di dalam CRISP-DM ?

Select one:

- a. 4
- b. 6
- c. 5
- d. 3

The correct answer is: 6

"Ada sebuah studi kasus tentang forecasting demand dari suatu barang di perusahaan e-commerce," ini termasuk tipe analytics..

Select one:

- a. explorative
- b. predictive
- c. descriptive
- d. initiative

The correct answer is: predictive

Jika model yang sudah kita train dan test tidak memenuhi standar akurasi yang ditetapkan, maka harus melakukan step :

Select one:

- a. Machine Learning Model and Tuning
- b. Experimentation
- c. Gather Raw Data
- d. Deploying Model

The correct answer is: Machine Learning Model and Tuning

ekstensi .pkl adalah python object yang sudah melalui tahap

Select one:

- a. Debugging
- b. Unit Testing
- c. Serialization
- d. Error Check

The correct answer is: Serialization

Berikut adalah consideration dalam model deployment dan pembuatan analytics architecture, kecuali :

Select one:

- a. Extensibility
- b. Visibility
- c. Reproducibility
- d. Scalability

The correct answer is: Visibility

Dibawah ini mana yang merupakan programming language yang disupport oleh Spark?

Select one:

- a. Scala
- b. Visual Basic
- c. C++
- d. C#

The correct answer is: Scala

Spark adalah

Select one:

- a. High performance file storage
- b. High performance single computation system
- c. High performance in-memory cluster computing
- d. High performance hardware

The correct answer is: High performance in-memory cluster computing

Berikut adalah components of scaling, kecuali

Select one:

- a. Model training
- b. Evaluation and deployment
- c. Data handling
- d. Preparing and fitting

The correct answer is: Preparing and fitting

Feature storage berguna untuk menyimpan

Select one:

- a. variabel-variabel yang akan digunakan untuk modelling
- b. variabel-variabel yang ada pada aplikasi mobile
- c. variabel-variabel yang ada pada entry form
- d. variabel-variabel yang akan dievaluasi

The correct answer is: variabel-variabel yang akan digunakan untuk modelling

Berikut adalah schema data processing

Select one:

- a. Real-time and out-time processing
- b. Real-time and at least once processing
- c. Batch and real-time processing
- d. Batch and out-time processing

The correct answer is: Batch and real-time processing

Serialisasi objek dalam bahasa pemrograman python akan berbentuk

Select one:

- a. Pickle
- b. Tuple
- c. PMML
- d. Series

The correct answer is: Pickle

Proses yang dilakukan terhadap feature sebelum masuk ke feature store adalah

Select one:

- a. feature mining
- b. feature blocking
- c. feature piping
- d. feature engineering

The correct answer is: feature engineering

Berikut adalah contoh messaging service

Select one:

- a. Oracle
- b. ActiveMQ
- c. Hadoop
- d. Zookeeper

The correct answer is: ActiveMQ

Dua pihak yang berinteraksi di dalam kafka disebut

Select one:

- a. Producer - Consumer
- b. Consumer - Collaborator
- c. Distributor - Consumer
- d. Producer - Distributor

The correct answer is: Producer - Consumer

Producer bertugas untuk .... sebuah message

Select one:

- a. Write-off
- b. Channel
- c. Publish
- d. Subscribe

The correct answer is: Publish

Zookeeper bertugas untuk

Select one:

- a. menyimpan metadata dari kafka
- b. memberihkan data dari kafka
- c. membuat metadata dari kafka
- d. menganalisis data dari kafka

The correct answer is: menyimpan metadata dari kafka

Manakah perusahaan yang menggunakan kafka pertama kali

Select one:

- a. Spotify
- b. LinkedIn
- c. Grab
- d. Adidas

The correct answer is: LinkedIn

Jenis data apa yang dapat digunakan untuk membuat visualisasi?

Select one:

- a. Data diskrit dan kontinyu
- b. Semua benar
- c. Data internal dan eksternal
- d. Data kualitatif dan kuantitatif

The correct answer is: Semua benar

Contoh dari data Internal adalah :

Select one:

- a. Daya beli masyarakat
- b. Jumlah karyawan
- c. Tingkat inflasi
- d. Jumlah penduduk kota

The correct answer is: Jumlah karyawan

Contoh data berkala (time series data) adalah

Select one:

- a. Perkembangan nilai tukar Rupiah terhadap US Dollar
- b. Laporan keuangan perusahaan posisi 31 Desember 2016 dan 31 Desember 2017
- c. Laporan keuangan penjualan tahun 2016-2018
- d. Semua benar

The correct answer is: Semua benar

Software-software dibawah ini dapat digunakan untuk membuat visualisasi data, kecuali ...

Select one:

- a. Microsoft Power Point
- b. Microsoft Power BI
- c. Microsoft Excel
- d. Microsoft Word

The correct answer is: Microsoft Word

Tujuan utama dari visualisasi data adalah :.

Select one:

- a. membuat data yang kompleks menjadi mudah dipahami dan berguna
- b. untuk mengkomunikasikan informasi secara jelas dan efisien kepada pengguna lewat grafik
- c. Semua benar
- d. membantu pengguna dalam menganalisa dan penalaran tentang data

The correct answer is: Semua benar

Common definition of big data?

Select one:

- a. Data that is too large or too complex to be managed using traditional data processing, analysis, and storage techniques
- b. Set of the data
- c. Data that is simple to be managed using traditional data processing, analysis, and storage techniques
- d. Data that is easy to be managed using traditional data processing, analysis, and storage techniques

The correct answer is: Data that is too large or too complex to be managed using traditional data processing, analysis, and storage techniques

What's the function of Map/Reduce?

Select one:

- a. Provides parallel processing of data in HDFS.
- b. Data warehousing tools.
- c. NoSQL database.
- d. To analyze data sets.

The correct answer is: Provides parallel processing of data in HDFS.

Whats volume of data really say in terms of Big Data?

Select one:

- a. Not only the volume of data but also the correction of the data.
- b. Volume of data.
- c. All answers are true
- d. How data correction works in Big Data.

The correct answer is: Volume of data.

Different types and forms of data dikenal dengan istilah?

Select one:

- a. Veracity
- b. Variety
- c. Velocity
- d. Volume

The correct answer is: Variety

Data Scientist menangani Data, Information, Knowledge dan Understanding yang bersifat?

Select one:

- a. Known Knowns
- b. Known Unknowns
- c. Known Unknowns dan Unknown Unknowns
- d. Unknown Unknowns

The correct answer is: Known Unknowns dan Unknown Unknowns

metodologi project yang digunakan untuk sebuah data science project adalah

Select one:

- a. agile atau MVP
- b. agile atau watermark
- c. waterfall atau MVP
- d. agile atau waterfall

The correct answer is: agile atau waterfall

PMML adalah

Select one:

- a. Prescriptive Model Management Language
- b. Prescriptive Model Markup Language
- c. Predictive Model Markup Language
- d. Predictive Model Management Language

The correct answer is: Predictive Model Markup Language

Jika kita melakukan training model di Jupyter Notebook, lalu akan melakukan deployment di Application dengan Programming Language Java, maka disarankan kita mengeluarkan output model dalam format

Select one:

- a. PMML
- b. Pickle
- c. HTML
- d. JSON

The correct answer is: PMML

Berikut ini adalah contoh output dari sebuah data science project kecuali

Select one:

- a. Archive File
- b. Aplikasi
- c. Model
- d. Dashboard

The correct answer is: Archive File

Seseorang yang bertugas mengatur workload dari sebuah Data Science Team disebut

Select one:

- a. Project Manager
- b. Project Assistant
- c. Data Engineer
- d. Product Manager

The correct answer is: Project Manager

"process of using domain knowledge of the data to create features that make machine learning algorithms work" adalah definisi dari

Select one:

- a. feature engineering
- b. feature filtering
- c. feature extraction
- d. feature system

The correct answer is: feature engineering

Contoh visualisasi data yang penting dalam sebuah Data Science Project di bawah ini, kecuali

Select one:

- a. Lorentz Curve
- b. ROC Curve
- c. Correlation Matrix
- d. Distribution Plot

The correct answer is: Lorentz Curve

Dalam proses model deployment, tujuan dari pickling model adalah

Select one:

- a. Meminimalisir size dari suatu model
- b. Memindahkan model ke suatu aplikasi java
- c. Serialisasi python object agar dapat di load ke python program
- d. Membuat model menjadi readable

The correct answer is: Serialisasi python object agar dapat di load ke python program

Jika kita menggunakan One-Hot Encoding sebelum melakukan Data Modeling, serialization lebih baik dilakukan di

Select one:

- a. Jupyter Notebook (training environment)
- b. Postman
- c. Flask
- d. Application Environment (ex : VSCode)

The correct answer is: Application Environment (ex : VSCode)

Dalam proses model deployment, tujuan dari pickling model adalah

Select one:

- a. Meminimalisir size dari suatu model
- b. Membuat model menjadi readable
- c. Memindahkan model ke suatu aplikasi java
- d. Serialisasi python object agar dapat di load ke python program

The correct answer is: Serialisasi python object agar dapat di load ke python program

Library python untuk mengkonversi python object menjadi json disebut

Select one:

- a. Conda
- b. Jsonify
- c. XGBoost
- d. Spyder

The correct answer is: Jsonify

Dibawah ini adalah library yang dilakukan untuk menghandle imbalance dataset

Select one:

- a. Feature engineering
- b. Feature extraction
- c. SMOTE
- d. Hyperparameter tuning

The correct answer is: SMOTE

Berikut adalah syntax yang benar untuk write model ke dalam bentuk pickle

Select one:

- a. pickle.loads(f.read())
- b. pickle.dump(Model, open('Models.pkl', 'wb'))
- c. joblib.load('Models.pkl')
- d. pd.write_models('models.pkl')

The correct answer is: pickle.dump(Model, open('Models.pkl', 'wb'))

Berikut adalah syntax yang benar untuk write model ke dalam bentuk pickle

Select one:

- a. pickle.loads(f.read())
- b. pickle.dump(Model, open('Models.pkl', 'wb'))
- c. joblib.load('Models.pkl')
- d. pd.write_models('models.pkl')

The correct answer is: pickle.dump(Model, open('Models.pkl', 'wb'))

Berikut adalah cara penghitungan F1 Score

Select one:

- a. TP/(TP+FP)
- b. TP+TN/(TP+FP+FN+TN)
- c. (TP+TN)/(TP*TN)
- d. 2*(Recall * Precision) / (Recall + Precision)

The correct answer is: 2*(Recall * Precision) / (Recall + Precision)