**ONLINE\_TEST\_3\_2025**

**-------------------------------------------------------------------------------------------**

**Question 1:** Which of the following is (are) correct?

   ⬜ Median is used to represent the spread of the data from the centre

   ⬜ The standard deviation is used to represent the relatie position of specific measurement in the data

   ✔ Mean and median are used to measure the central  tendency of the data

   ✔ Descriptive statistics quatiatively describe the main features of data

**Question 2:** Which of the following descriptions is generally used in a statistics test?

   ⬜ Assume the body weight of the participant is lighter after the diet control.

   ⬜ Assume the body weight of the participant is heavier after the diet control.

   ✔ Assume the body weight of the participant is the same before and after the diet control

   ✔ Assume the body weight of the participant is different before and after the diet control

**Question 3:**

John flies frequently and likes to upgrade his seat to first class. He has determined that if he checks in for his flight at least two hours early., the probability that he will get an upgrade is 0.65, otherwise, the probability that he will get an upgrade is 0.3. While his busy schedule, he checks in at least two hours before his flight only 45% of the time.

**Suppose John did not receive an upgrade on his most recent attempt, what is the probability that he did not arrives two hours only?**

**Answer:**

Step: Let’s define some events

A: means that john checks in at least two hours early

A’: means that John **does not** check at least two hours early

U: means that Johmn gets an upgrade

U’: means that John **does not** get an upgrade.

Step: Calculate probabilities from problem statement:

+ P(A) = 0.45 => Checks in at least two hours early.

+ P(A’) = 1 – 0.45 = 0.55 => Does not check in at least two hours early.

+ P(U) = 0.3 => Gets an upgrade

+ P(U’) = 1 – 0.3 = 0.7 => Does not get un upgrade

Step: Calculate important stuff things

**Question 4:** Tutors (Teachers) may ask about the formula of Confidence, Lift, and Leverage

**Question 5:** Given the example of tossing two coins (Bài toán về xác suất tung đồng xu)

The possible sample space is {HH, HT, TH, TT} where H and T represent the head and tail respectively. Select the descriptions(s) that are(is) TRUE.

✔ P(At least one H) = 0.75

Explaination: This sentence is true because the probability of H (Head) display is calculated by formula P(H) =

    ⬜ P(The second coin being T given the first coin is H) = 0.25

✔ P(HH) = 0.5

Explaination: This statement is true because it is calculauted by the following formula: P(HH) =

⬜ P(Getting two heads given the first coin is H) = 0.5

**Question 6:** Select the description(s) that is (are) TRUE

🗹 The success or failure criteria are determined by whether the data and the chosen analytics models are able to accept or reject the initial hypotheses formulated in data analytics lifecycle in phase 1.

⬜ Rejecting a hypothesis means the project is failed. The possible option is to restart the project from the beginning again with other assumptions.

⬜ The success or failure criteria are determined by whether the data and the chosen analytics models are able to accept or reject the initial hypotheses formulated in data abalytics lifecycle in phase 2.

🗹 Rejecting a hypothesis does not always equate a failure. Instead, a failure usually refers to the inability to strike the balance between two possible analytics ectremes.

**Question 7:** Given two datasets as in the figures below:

A diagram of numbers and red dots

AI-generated content may be incorrect.

Which of the following is(are) correct?

⬜ Both data 1 and data 2 have the same mean value if the outlier is excluded.

⬜ The median value of data 2 is better representative than the median value.

⬜ Both data 1 and data 2 have the same mean value

⬜ Both data 1 and data 2 have the same median value

**Question 8:** Select the description(s) that is(are) TRUE

🗹 The neuron with the minimum distance is the self-organising map training is the winning neuron.

🗹 The Kohoren layer is the self-organising map neural network is a 2-D layer.

⬜ The neurons in the Kohoren layer of a self-organizing map should have full connections to other neiron in the samle layer

⬜ The neuron with the maximum distance is the self-organising map training is the winning neuron.

**Question 9:** Assume that the Gini Indies 0.73 and 0.57 are corresponding to split criteria X and Y respectively. Select the description(s) that is(are) TRUE.

⬜ if the datase after the split is pure, the value of Gini index would be 1

🗹 if the datase after the split is pure, the value of Gini index would be 0.

⬜ Criterion X is a better split option for purify the data.

🗹 Criterion Y is a better split option for purify the data.

**Question 10:** Select the description(s) that is (are) TRUE

⬜ Normalisation is used to change the observations so that the data can be described as a normal distribution.

🗹 Standardisation is also called z-score normalisation, which transforms data so that the resulting distribution has a mean of 0 and a standard deviation of 1

⬜ Min-Max normalisation is the only normalisation method

⬜ The normalisation cannot be applied if the data contains negative values.

**Question 11:** Align the correct sequence for a general ethics approval process:

(1): Familirise yourself with….

(2): Develop a research proposal

(3): Obtain informed consent.

(4): Submit your application.

(5): Wait for approval.

A couple of circles with different colors

AI-generated content may be incorrect.**Question 12:** Select the description(s) that is(are) TRUE

🗹 The entorpy of the yellow circle (on the right) group is 0.

⬜ If the entropy is large indicating that the data is pure.

🗹 The entropy of the blue circle (on the left) group is approximately 0.503

⬜ The entropy of the blue cicrle (on the left) group is 0.

**Question 13:** Given the table below, calculate the entropy for the interested goal from the specified observations.

🗹 The entropy of go jogging on Raining day is about 0.918 when using log2 in the calculation.

⬜ The entropy of go jogging from all observations is about 0.647 when using log2 in the calculation

⬜ The entropy of go fishing on raining day is 1.0 when using log2 in the calculation.

⬜ The entropy of go fishing on Sunny day is 1.0 when using log2 in the calculation.

|  |  |  |
| --- | --- | --- |
| **Outlook** | **Jogging** | **Fishing** |
| Cloudy | Yes | No |
| Windy | No | No |
| Sunny | No | Yes |
| Rain | Yes | No |
| Storm | No | No |
| Rain | No | No |
| Windy | No | No |
| Sunny | No | No |
| Rain | Yes | No |
| Cloudy | No | Yes |

Answer:

- Total transactions (Total Rows): 10

- Now consider each statement:

The entropy of go jogging on Raining day is about 0.918 when using log2 in the calculation.

+ Jogging on Raining Day: Yes(2) and No(1) => Total: 3

* Propabilities of Yes: => Propabilities of No:
* Entropy: H = - ≈0.918 => True

The entropy of go jogging from all observations is about 0.647 when using log2 in the calculation

+ Total observations of Jogging: Yes(3) and No(7) => 10 in total observations

+ P(Yes) = 3/0 and P(No) = 7/10

* H ≈ 0.881

**Question 14:** Accurancy formula

**Question 15:** Match the contents to their correct groups

- True positive rate (TPR): Sentivity

- True negative rate (TNR): Specificity

- Precision: Calculated by the true…..

- Area under the curve (AUC): The plasma concentration….

**Question 16:** Select the description(s) that is (are) TRUE

🗹 Stacking often considers heterogeneous weak learners, learns them in parallel and combines them by training a meta-model to output a prediction based on the different weak models predictions.

⬜ Stacking often considers homogeneous weak learners, learns them in parallel and combines them by training a meta-model to output a prediction based on the different weak models predictions.

🗹 Bagging is also called Bootstrapping.

⬜ The ensemble learning result must be obtained by taking the average of multiple outputs.

**Question 17:** A forensic expert is owrking on samples collected from the crime scene and the suspect. The samples are provided in the table below

|  |  |
| --- | --- |
| **From a suspect** | **From a crime scene** |
| 37 | 36 |
| 39 | 38 |
| 37 | 36 |
| 38 | 38 |

Assume the pooled variance is 1.345, the critical value of the Student’s t-test is

If the foresic expert examines the significance level of α = 0.01two-tailed t-test, the null hypothesis should be\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(accepted/rejected)

Mean of “**From a suspect”: 37.75**

Mean of “**From a crime scene”: 37**

**Question 18:** Remember the positive and negative skew diagram.

**Question 19:**

1. Type one error: Rejection of the null hypothesis….
2. Type two error: Acceptance of the null hypothesis…
3. β: The probability of a type…
4. β-1: The power of a test

**Question 20:** Standard Deviation (SD) formula

**Question 21:** If you find that the Student’s t-test is used in the data analysis, the test datasets should have **equal** variance. On the other hand, the Welch’s t-test will be used if the test datasets have **unequal** variance.

**Question 22: CNN is insensitive to which of the following characteristic(s)? =>** Shifting, rotation, mirrowing

**Question 23:** Given a feed-forward neural network structure:

A diagram of a network

AI-generated content may be incorrect.

This kind of feed-forward neural network is called a **multiplier perceptron.**

**Question 24:** In the Radial Basis Function (RBF) neural network, the RBF can be defined by φ(*D*) = , where *D* is the distance between the data point and the centre of a circle and .

Given the two RBF plots below, the

A graph with a blue curve

AI-generated content may be incorrect.

**Question 25:** Given a feature map below

|  |  |  |  |
| --- | --- | --- | --- |
| -3 | 1 | 3 | 2 |
| 0 | -2 | -4 | 2 |
| -2 | 0 | 1 | -4 |
| -1 | 0 | -3 | -3 |

If the step size is 2, the output after max pooling would be:

|  |  |
| --- | --- |
| (a) | (b) |
| (c) | (d) |

**Question 26:**

Given one-dimensional sequence A:[2,3,4] and a mask sequence B:[-1,0,1]

The output of the 1-D convolution result A\*B is [\_\_\_\_\_\_\_\_\_\_\_\_]

**Question 27:** Which of the following content(s) should be included in the analyst presentation?  
 🗹 Describe expected model performance and any vaceal.

⬜ Support key points with simple charts and graphics.

⬜ List the relevant and actionable points to champion the proposed analytic solution within the organiation.

⬜ Explain the business impact, including risks and ROI

🗹 Discuss variables, scope, and predictive power.

🗹 Supplement recommendations with implications for the modelling.

**Question 28:**

[\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_] is usually to judge whether a chart is overfitted with information. It also rquals [\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_] minus the proportion of a graphic that can be erased without being data information.

**Question 29:**

In the model deployment, it is recommended that back-testing and now-testing infrastructure be developed.

A [\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_] infrastructure tests any new update to a model on historical data to determine whether the update has truly made improvement to the older model.

The [\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_] is a test of the new model on the latest data to check whether the new model is ready for deployment.

**Question 30:** Match the best fit of data to be visualized to the type of charts from left to right.

+ Components (part of whole): Donus chart

+ Item: Bar chart

+ Time-series data: Line chart

+ Frequency: Histogram

+ Correlation: side-by-side bar chart

**Question 31**: When assessing the significance of results, we still need to consider the\_\_\_\_significance, as it depends on the subject matter and can be different in other applications

**Question 32 -> 35: Using KNIME to create MP model, Naïve Bayer,…**