Autorità Garante
della Concorrenza e del Mercato

Commissione esaminatrice del concorso pubblico, per titoli ed esami, per l'assunzione straordinaria a tempo indeterminato di 2 funzionari in prova, al livello 6 della tabella stipendiale dei funzionari dell'Autorità, per lo svolgimento di attività di *data engineering* e *data science* (F6DS).

Prova scritta del 7 maggio 2024

TRACCIA N. 3

Domanda 1:

In una distribuzione bimodale la media è sempre uguale a:

a) La moda
b) La mediana
c) Il minimo della distribuzione
d) Nessuno dei casi precedenti

Domanda 2:

Sia x una variabile casuale (v.c.) continua che segue la distribuzione normale standardizzata. L'intervallo (-k,+k) contiene circa il 99,7% della distribuzione. Quale è il valore di k?

a) $0 < k < 1$
b) 1
c) 2
d) 3

Domanda 3:

La differenza tra il terzo e il primo quartile di una distribuzione è una misura della:

a) Dispersione
b) Dimensione
c) Simmetria
d) Normalità

Domanda 4:

Due dadi vengono lanciati indipendentemente per 10 volte, generando le seguenti due sequenze di valori: {2,2,2,6,6,4,4,1,5,3}, {1,3,2,1,6,6,3,4,5,4}. Possiamo concludere che:

a) Il primo dado è truccato
b) Il secondo dado è truccato
c) I due dadi non sono truccati
d) Non possiamo trarre conclusioni sui due dadi

Domanda 5:

In una scatola di cioccolatini sono contenuti 9 pezzi, di cui 3 alla menta, 3 al cocco e 3 al caffè, tutti di uguale forma e colore. Quale è la probabilità che, scegliendo e mangiando 1 cioccolatino a caso per 3 volte, se ne mangi almeno uno al caffè?

a) 16/21
b) 10/42
c) 9/28
d) 1/3

Domanda 6:

Sia x una variabile casuale, e $y = \alpha x + \varepsilon$, dove l'errore ε è distribuito secondo una distribuzione normale con media nulla e deviazione standard > 0. Si ha che:

a) La correlazione tra y e x è nulla
b) La correlazione tra y e x è pari a 1
c) Il segno della correlazione tra y e x è pari al segno di α
d) Il valore della correlazione dipende dal valore della deviazione standard di ε

Domanda 7:

Per eseguire un *test* di ipotesi sulla differenza tra media di due distribuzioni, si impiega:

a) La disuguaglianza di *Chebyshev*
b) La distribuzione esponenziale
c) La distribuzione uniforme
d) La distribuzione binomiale

Domanda 8:

Un giocatore deve scegliere tra il gioco A e il gioco B. Con il gioco A potrà vincere 100 euro con probabilità 0,05; con il gioco B potrà vincere 200 euro con probabilità pari a *p*. Per partecipare al gioco A è richiesta una posta di 5 euro, per B una posta di 20 euro. Per quale valore di *p* i due giochi possono essere considerati equivalenti (i.e. caratterizzati da eguali vincite)?

a) 1/2
b) 1/10
c) 1/20
d) Non si può stabilire con le informazioni disponibili

Domanda 9:

Il metodo della *silhouette* viene impiegato per verificare la qualità dei risultati di un algoritmo di apprendimento non supervisionato.

a) Vero
b) Falso
c) Dipende dai casi applicativi
d) Solo per grandi campioni

Domanda 10:

In quale modo è possibile ridurre *l'overfitting* di un metodo di classificazione basato su alberi di decisione?

a) Applicando il *pruning* nella fase di addestramento
b) Riducendo la dimensione dell'insieme di addestramento
c) Riducendo la dimensione massima delle foglie dell'albero
d) Imponendo all'algoritmo di sviluppare un albero bilanciato nella fase di addestramento

Domanda 11:

La capacità di generalizzazione di un algoritmo di classificazione basato su *Support Vector Machine* con Kernel polinomiale è maggiore:

a) All'aumentare della dimensione dell'insieme di *test*
b) All'aumentare della dimensione dell'insieme di validazione
c) All'aumentare del coefficiente del Kernel impiegato
d) Al diminuire del coefficiente del Kernel impiegato

Domanda 12:

L'algoritmo di classificazione *Nearest Neighbor* è caratterizzato da:

a) Apprendimento oneroso e alta capacità di generalizzazione
b) Apprendimento oneroso e bassa capacità di generalizzazione
c) Apprendimento economico e alta capacità di generalizzazione
d) Apprendimento economico e bassa capacità di generalizzazione

Domanda 13:

Si intende applicare un algoritmo di apprendimento supervisionato dove la variabile *target* è di tipo continuo. Quale delle seguenti tecniche è più indicata:

a) Regressione
b) Alberi di classificazione
c) Reti neurali con neuroni di *output* sigmoidali
d) Analisi in componenti principali

Domanda 14:

Le reti neurali di tipo *Long-Short-Term-Memory* sono particolarmente indicate per:

a) Apprendimento da dati di grandi dimensioni
b) Apprendimento da dati con connotazione spaziale
c) Apprendimento da dati dipendenti dal tempo
d) Apprendimento da dati espressi da variabili qualitative

Domanda 15:

L'uso del meccanismo della *self-attention* è caratterizzante per:

a) Large Language Models
b) Reti Neurali Convoluzionali
c) Apprendimento di Immagini
d) Modelli di apprendimento con rinforzo (*reinforcement learning*)

Domanda 16:

L'analisi in componenti principali (ACP) è un metodo di:

a) *Feature Selection*
b) *Feature Extraction*
c) Apprendimento supervisionato
d) Inferenza statistica

Domanda 17:

Se la funzione FUN(n) richiede tempo $\Theta(n)$ lineare in n, qual è il tempo di esecuzione di questo ciclo?

```
j = n
while j ≥ 1 {
        FUN(n)
        j = j-1
}
```

a) $T(n) = \Theta(n)$
b) $T(n) = \Theta(n^2)$
c) $T(n) = \Theta(n \log n)$
d) Nessuna delle precedenti

Domanda 18:

Si supponga di memorizzare n valori in un *array* ordinato oppure in un albero di ricerca bilanciato (ad esempio, AVL o *red-black*). Si compili la tabella sottostante, specificando il tempo di esecuzione asintotico per ciascuna operazione nel caso peggiore.

Operazione	*Array* ordinato	Albero binario di ricerca bilanciato
Trovare il minimo		
Inserire un nuovo elemento		

Domanda 19:

Assumendo che n rappresenti la dimensione dell'*input*, si dica quale delle seguenti affermazioni è vera:

a) Se un algoritmo ha tempo di esecuzione $\Theta(n^2)$ nel caso migliore, allora anche nel caso peggiore terminerà in $\Theta(n^2)$ passi
b) Se un algoritmo ha tempo di esecuzione $\Omega(n)$ e $O(n)$ nel caso peggiore, possiamo concludere che nel caso peggiore è $\Theta(n)$
c) Il problema dei cammini minimi è ben definito anche in grafi orientati pesati che contengono cicli di costo negativo
d) Esistono algoritmi di ordinamento basati su confronti che impiegano tempo O(n) nel caso peggiore

Domanda 20:

Si consideri un albero binario T di n nodi e altezza h tale che, tra tutti gli alberi binari di altezza h, T abbia il massimo numero possibile di nodi. Quale delle seguenti relazioni soddisfa l'altezza, in funzione del numero di nodi?

a) $h = \Theta(\log n)$
b) $h = \Theta(n)$
c) $h = \Theta(n^2)$
d) Nessuna delle precedenti

Domanda 21:

Si considerino due grafi con lo stesso numero di nodi e la stessa distribuzione dei gradi. Si dica quale delle seguenti affermazioni è vera:

a) Il diametro dei due grafi è uguale
b) Il numero di archi dei due grafi è uguale
c) In un tale grafo, tutti i nodi devono avere lo stesso grado
d) Il grado medio dei due grafi può essere diverso

Domanda 22:

Dire quale delle seguenti affermazioni è vera. Un grafo casuale ottenuto secondo il modello di Erdős–Rényi G(n,p):

a) Modella bene le caratteristiche delle reti sociali
b) Non può avere nodi isolati (ovvero di grado 0)
c) Ha un grado atteso dei nodi pari a (n-1)p
d) Ha una distribuzione dei gradi che segue una legge a potenza

Domanda 23:

Usando il sistema decimale (potenze in base 10), 3 Gigabyte sono equivalenti a:

a) 3000 Byte
b) 3000 MB
c) 3000 KB
d) 3000 TB

Domanda 24:

In un diagramma E/R, quale dei seguenti potrebbe essere un attributo?

a) Data di nascita
b) Persone
c) Automobili
d) Fatture

Domanda 25:

Si immagini di avere un dataset rappresentato nelle seguenti tabelle:

- employee (id, *employee_name*, *address*, city)
- works (*employee_id, employee_name, company_name, salary*)
- company (*company_name, city*)
- manages (*code, manager_name, employee_name, age*)

Quale codice SQL trova l'azienda con il più basso payroll?

(a)
```
SELECT company_name
FROM works
GROUP BY company_name
HAVING SUM (salary) <= ALL      (SELECT SUM (salary)
                                    FROM works
                                    GROUP BY company_name)
```

(b)
```
SELECT manager_name
FROM works
GROUP BY company_name
HAVING SUM (salary) <= ALL      (SELECT SUM (salary)
                                    FROM works
                                    GROUP BY manager_name)
```

(c)
```
SELECT company_name
FROM works
GROUP BY company_name
HAVING SUM (salary) <= ALL      (SELECT AVG (salary)
                                    FROM works
                                    GROUP BY employee_name)
```

(d)
```
SELECT company_name
FROM works
GROUP BY employee_name
HAVING SUM (salary) <= ALL      (SELECT AVG (salary)
                                    FROM works
                                    GROUP BY company_name)
```

Domanda 26:

Quale espressione Python ha come valore 512?

a) 2**(3**2)
b) (2**3)**2
c) 2**2**3
d) Nessuna

Domanda 27:

In Python, quale delle seguenti affermazioni è vera?

a) Avendo due liste, è possibile moltiplicare gli elementi a coppie calcolando A*B
b) Avendo due ndarray di Numpy, è possibile moltiplicare gli elementi a coppie calcolando A*B
c) Avendo due tuple di valori, è possibile moltiplicare gli elementi a coppie direttamente come A*B
d) Avendo due dizionari, è possibile moltiplicare le chiavi a coppie direttamente come A*B

Domanda 28:

In Python, quando viene eseguita una funzione che non deve restituire alcun valore, di default viene restituito:

a) 0
b) False
c) None
d) Un valore intero arbitrario

Domanda 29:

In Python, eseguendo il codice

my_tuple.append((5, 6, 7))

su una tupla che contiene i primi quattro numeri interi positivi, quanto varrà len(my_tuple)?

a) 1
b) 2
c) None
d) Verrà lanciata un'eccezione

Domanda 30:

Si consideri il seguente codice Python.

```
def askWeather():
        print("Is it rainy?")
```

La prima linea è chiamata:

a) Intestazione della funzione
b) Corpo della funzione
c) Definizione della funzione
d) Nome della funzione

Domanda 31:

Le Tabelle 1 e 2 riportano le matrici di confusione ottenute tramite *cross-validation* dopo l'applicazione di due algoritmi di classificazione per lo stesso *dataset*. Il candidato commenti la validità relativa dei due algoritmi secondo i risultati ottenuti (usare al massimo 200 parole).

Tabella 1: *Algoritmo A, Risultati Cross Validation*

	NEG	POS
NEG	222	52
POS	18	122

Tabella 2: *Algoritmo B, Risultati Cross Validation*

	NEG	POS
NEG	195	79
POS	31	109

Domanda 32:

Le Tabelle 3 e 4 riportano le matrici di confusione dopo l'applicazione di due algoritmi di classificazione per lo stesso *dataset*, per l'insieme di addestramento (*training*) e di *test*. Il candidato commenti la validità relativa dei due algoritmi secondo i risultati ottenuti, commentando in particolare la eventuale presenza di *overfitting* (usare al massimo 200 parole).

Tabella 3

Algoritmo A, Training

	NEG	POS
NEG	120	15
POS	5	25

Tabella 4

Algoritmo B, Training

	NEG	POS
NEG	98	37
POS	1	29

Algoritmo A, Testing

	NEG	POS
NEG	450	52
POS	20	270

Algoritmo B, Testing

	NEG	POS
NEG	400	102
POS	12	278

Domanda 33:
Avete svolto l'esame scritto del corso di Big Data Analytics, cui hanno partecipato 90 studenti. Dopo aver raccolto e corretto gli elaborati stampati su fogli A4, dovete ora ordinarli per voto crescente (i voti ricadono nell'intervallo da 0 a 30 e sono annotati direttamente su ciascun elaborato). Quale algoritmo utilizzereste: Bubblesort, Mergesort o Countingsort? Scegliete la strategia che vi sembra più *pratica*, considerando che avete a disposizione una scrivania lunga 3 metri su cui disporre i compiti. Motivate poi la vostra risposta, usando al massimo 200 parole.

Domanda 34:

Sia dato un Albero Binario di Ricerca contenente n valori interi distinti. Discutere dove può trovarsi il *secondo intero più piccolo*, usando al massimo 200 parole.