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. 2

Domanda 1:

In una distribuzione simmetrica, la media è sempre uguale a:

a) La moda
b) La mediana
c) Il minimo della distribuzione
d) Nessuna delle 3

Domanda 2:

Sia x una variabile casuale (v.c.) continua che segue la distribuzione normale standardizzata. L'intervallo (-k,+k) contiene circa il 52,3% 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 dispersione della distribuzione.

a) Vero
b) Falso
c) Solo per le distribuzioni bimodali
d) Solo per le distribuzioni simmetriche

Domanda 4:

Un generatore di numeri pseudocasuali produce le due sequenze di numeri compresi tra 1 e 10: {1,3,5,7,9,8,2,4,3,4}, {1,1,1,2,2,2,3,3,3,4}. Possiamo concludere che:

a) Il generatore funzione in modo corretto
b) La seconda sequenza non è casuale
c) Il generatore non funziona in modo corretto
d) Non possiamo trarre conclusioni sul suo funzionamento

Domanda 5:

Una scatola contiene 4 caramelle rosse e 3 caramelle bianche. Pierino senza guardare, sceglie in sequenza, 3 caramelle a caso. Quale è la probabilità che Pierino mangi prima una caramella rossa, poi una bianca, poi una rossa?

a) 4/35
b) 2/35
c) 4/27
d) 6/35

Domanda 6:

Sia x una variabile casuale, e $y = 0,5x + e$, dove l'errore e è distribuito secondo una distribuzione normale con media nulla e deviazione *standard* > 0. Si ha che:

a) La correlazione tra y e x è pari a 1/2
b) La correlazione tra y e x è > 0
c) La correlazione tra y e x è pari a 1
d) Il valore della correlazione dipende dal valore della deviazione standard di e.

Domanda 7:

In assenza di informazioni sulla distribuzione di una variabile casuale, è possibile eseguire un *test* di ipotesi:

a) In nessun caso
b) Calcolando il *p-value* empirico con il metodo di Monte Carlo
c) Adottando la distribuzione normale
d) Adottando la distribuzione uniforme

Domanda 8:

Un giocatore deve scegliere tra il gioco A e il gioco B. Con il gioco A potrà vincere 100 euro con probabilità 0,03; con il gioco B potrà vincere 200 euro con probabilità 0,02. Per partecipare al gioco A è richiesta una posta di 5 euro, per B una posta di 3 euro. Quale è il gioco più profittevole?

a) Non si può stabilire con le informazioni disponibili
b) A
c) B
d) I due giochi solo equivalenti

Domanda 9:

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

a) Vero
b) Falso
c) Dipende dai casi applicativi
d) Solo nel caso di risultati ottenuti con algoritmo *k-means*

Domanda 10:

Per quale motivo si applica il *pruning* negli alberi di decisione?

a) Per rendere l'albero più efficiente nella fase di test
b) Per ridurre la dimensione dell'insieme di addestramento
c) Per evitare l'*overfitting* dell'insieme di addestramento
d) Per bilanciare l'errore sulle diverse classi

Domanda 11:

Per misurare l'affidabilità di un algoritmo di classificazione è sufficiente valutare il numero di falsi positivi e falsi negativi ottenuti nell'insieme di addestramento.

a) Vero
b) Falso
c) Dipende dalla dimensione dell'insieme di addestramento
d) Dipende dal tipo di algoritmo

Domanda 12:

L'algoritmo di classificazione *nearest neighbor* non richiede addestramento.
a) Vero
b) Falso
c) Solo per insieme di *training* di dimensioni ridotte
d) Dipende dai dati oggetto di analisi

Domanda 13:

Quali tecniche possono essere impiegate per dati di apprendimento dove una o più classi hanno frequenza molto più bassa rispetto alle altre classi?

a) Ricampionamento
b) Uso di *deep-learning*
c) Uso di *cross-validation*
d) Nessuna delle tre

Domanda 14:

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

a) Il riconoscimento di immagini
b) La previsione di serie storiche
c) La classificazione di dati qualitativi
d) La classificazione di dati biomedici

Domanda 15:

Clustering, Regressione, mappe di *Kohonen*, sono tutti metodi di apprendimento supervisionato.

a) Vero
b) Falso
c) Dipende dai dati impiegati
d) Dipende dalle modalità di impiego

Domanda 16:

L'analisi in componenti principali (ACP) è un metodo di apprendimento supervisionato.

a) Vero
b) Falso
c) Dipende dalla dimensione dei dati
d) Nessuna delle tre

Domanda 17:

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

```
j=n
while j>1 {
        FUN(j)
        j=j/2
}
```

a) $T(n) = \Theta (\log n)$
b) $T(n) = \Theta (n)$
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 *max-heap* (*heap* binario in cui il massimo è memorizzato nella radice). Si compili la tabella sottostante, specificando il tempo di esecuzione asintotico per ciascuna operazione nel caso peggiore, utilizzando il miglior algoritmo noto.

Operazione	*Max heap*	*Array* ordinato
Trovare il massimo		
Inserire un nuovo elemento		

Domanda 19:

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

a) Ogni algoritmo per visitare in profondità un grafo non orientato e non connesso richiede tempo almeno esponenziale nel caso peggiore
b) Il problema dei cammini minimi è ben definito anche in grafi orientati pesati che contengono cicli di costo negativo
c) Se si dimostra che un algoritmo ha tempo di esecuzione $O(n^2)$ nel caso peggiore, è comunque possibile che su qualche istanza termini in Θ (n) passi
d) Se si dimostra che un algoritmo ha tempo di esecuzione $\Theta(n^3)$ nel caso migliore, è comunque possibile che esistano istanze su cui termina in $O(n^2)$ passi

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 minimo numero possibile di nodi. Quale delle seguenti relazioni soddisfa il numero di nodi n, in funzione dell'altezza h?

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

Domanda 21:

Sia G un grafo non orientato con $n=2^{20}$ nodi e $m=2^{30}$ archi. Si dica quale delle seguenti affermazioni è vera:

a) Il grado medio è 2^{15}
b) Al più 2^{16} nodi di G possono avere grado $\geq 2^{15}$

c) Il grafo è sicuramente non connesso
d) In un tale grafo, tutti i nodi devono avere lo stesso grado

Domanda 22:

Dire quale delle seguenti affermazioni è vera. Tipicamente, una rete sociale di grandi dimensioni:

a) Ha un diametro grande rispetto al numero di nodi
b) Non contiene nodi di grado molto alto
c) È densa
d) Ha una distribuzione dei gradi che segue una legge a potenza

Domanda 23:

I dati memorizzabili in una memoria RAM hanno solitamente ordine di grandezza proporzionale a:

a) Poche decine di *Megabyte*
b) Poche decine di *Gigabyte*
c) Qualche *Terabyte*
d) Qualche *Petabyte*

Domanda 24:

Nella modellazione di un database di automobili, quale tipo di relazione può essere utilizzata per modellare la relazione tra proprietari e automobili (considerando che le automobili non possano essere cointestate)?

a) Uno a Molti
b) Uno a Uno
c) Molti a Molti
d) Nessuna

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 maggior numero di dipendenti?

(a)

```
SELECT company_name
    FROM works
    GROUP BY company_name
    HAVING COUNT (DISTINCT employee_name) >= ALL (SELECT COUNT (DISTINCT employee_name)
                                                  FROM works
                                                  GROUP BY company_name)
```

(b)

```
SELECT company_name
    FROM works
    GROUP BY company_name
    HAVING COUNT (DISTINCT employee_name) >=  (SELECT COUNT (DISTINCT employee_name)
                                                  FROM works
                                                  GROUP BY company_name)
```

(c)

```
SELECT company_name
    FROM works
    GROUP BY company_name
    HAVING COUNT (DISTINCT employee_name) >= NOT IN (SELECT COUNT (DISTINCT employee_name)
                                                  FROM works
                                                  GROUP BY company_name)
```

(d)

```
SELECT company_name
    FROM works
    GROUP BY company_name
    HAVING COUNT (DISTINCT manager_name) >= ALL (SELECT COUNT (DISTINCT manager_name)
                                                  FROM works
                                                  GROUP BY employee_name)
```

Domanda 26:

In Python, data una lista seq=[-1,2,1,4,8,5,-12], qual è il valore di -seq[1:-1][:-2][-1]?

a) -4
b) [-5,12]
c) 4
d) Viene sollevata un'eccezione durante la valutazione

Domanda 27:

Quale delle seguenti affermazioni è vera?

a) Python avrebbe la stessa espressività anche senza l'istruzione *for* (ossia, qualsiasi programma contenente un ciclo for può essere reimplementato con un *while*)
b) Python avrebbe la stessa espressività anche senza l'istruzione *while* (ossia, qualsiasi programma contenente un ciclo *while* può essere reimplementato con un *for*)

c) Python non sempre può evitare la perdita di dati nelle conversioni automatiche di tipo
d) Per iterare su una sequenza di valori di lunghezza sconosciuta, si utilizza tipicamente un'istruzione *for*

Domanda 28:

In Python, quando un insieme di istruzioni correlate progettate per eseguire un compito computazionale è raggruppato insieme, si sta implementando:

a) un modulo
b) un pacchetto
c) una funzione
d) uno *script* Python

Domanda 29:

In Python, date le liste L=[2,4,2,5,3,1,-3,7] e R=[0], qual è il contenuto di R dopo l'esecuzione del seguente codice?

for i in range (2,5): R.append(L[i])

a) [0,4,2,5]
b) [0,2,5,3,1]
c) [0,2,5,3]
d) [0]

Domanda 30:

Quale dei seguenti non sarebbe un nome di variabile corretto in Python?

a) a
b) x1
c) length
d) 123x

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*		
	NEG	POS
NEG	120	20
POS	75	120

Tabella 2: *Algoritmo B, Risultati Cross*		
	NEG	POS
NEG	135	5
POS	45	150

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	150	22
POS	60	360

Tabella 4		
Algoritmo B, Training		
	NEG	POS
NEG	164	8
POS	12	408

Algoritmo A, Testing		
	NEG	POS
NEG	256	38
POS	52	662

Algoritmo B, Testing		
	NEG	POS
NEG	250	44
POS	148	566

Domanda 33:

Avete svolto l'esame scritto del corso di Machine Learning, cui hanno partecipato 80 studenti. Dopo aver raccolto gli elaborati stampati su fogli A4, dovete ora ordinarli *in base al numero di matricola* indicato dallo studente (si tratta di un numero a otto cifre, ad esempio 20547791). Quale algoritmo utilizzereste: Selectionsort, Heapsort o Radixsort? 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 *max-heap* (*heap* binario in cui il massimo è memorizzato nella radice) contenente n valori interi distinti. Discutere dove può trovarsi il *terzo intero più grande*, usando al massimo 200 parole.