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. 1

Domanda 1:

Quale è il centro di una distribuzione secondo la norma L_1?

a) La moda
b) La mediana
c) La media aritmetica
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 68,3% della distribuzione. Quale è il valore di k?

a) 1/2
b) 1
c) 2
d) 3

Domanda 3:

La distribuzione di una variabile casuale (v.c.) ha media positiva e deviazione *standard* > 0. Quali delle seguenti conseguenze possono essere tratte?

a) La v.c. ha sempre valore positivo
b) La v.c. ha sempre valore non-negativo
c) La v.c. non segue la distribuzione uniforme
d) Nessuna delle 3

Domanda 4:

Usando un generatore di numeri pseudocasuali non è possibile generare la stessa sequenza di numeri più volte.

a) Vero
b) Falso
c) Dipende dalla lunghezza della sequenza
d) Dipende dal seme del generatore

Domanda 5:

Un'urna contiene 4 palline rosse e 2 palline bianche. Le palline sono estratte una ad una e poi reintrodotte nell'urna. Quale è la probabilità di estrarre una pallina rossa, poi una bianca, poi una rossa, in tre estrazioni successive?

a) 7/36
b) 4/27
c) 2/27
d) 6/27

Domanda 6:

Sia x una variabile casuale, e $y = a*x$ con $a < 0$. La correlazione tra x e y è pari a:

a) 1
b) 0
c) -1
d) Nessuna delle 3

Domanda 7:

Nei test statistici si indica tipicamente con il termine *p-value*:

a) La probabilità di estrarre a caso un campione molto distante dalla media della distribuzione
b) Il rapporto tra il numero di casi positivi e il numero totale dei casi
c) La probabilità del valore osservato sotto l'ipotesi nulla
d) Nessuna delle 3

Domanda 8:

Si lancia per 3 volte un dado a sei facce. Indicare la probabilità che la somma dei 3 valori sia pari a 6.

a) 1/24
b) 10/216
c) 2/36

d) 1/36

Domanda 9:

Il metodo della *silhouette* si applica unicamente per la valutazione di metodi di *clustering* basati sul *k-means*.

a) Vero
b) Falso
c) Dipende dai casi applicativi
d) Dipende dalla dimensione dei dati

Domanda 10:

La propensione all'*overfitting* di un albero di decisione può essere controllata tramite *pruning*.

a) Vero
b) Falso
c) Solo per problemi di grandi dimensioni
d) Dipende dal metodo di *pruning* impiegato

Domanda 11:

Il tasso di falsi positivi di un algoritmo di classificazione misurato tramite *cross-validation* aumenta sempre all'aumentare dell'*overfitting* dell'algoritmo.

a) Vero
b) Falso
c) Dipende dall'algoritmo di classificazione impiegato
d) Dipende dal numero di *fold* della *cross-validation*

Domanda 12:

La capacità di generalizzazione dell'algoritmo *k-nearest neighbor* aumenta all'aumentare del valore di k.

a) Vero
b) Falso
c) Dipende dalla dimensione dello spazio
d) Solo per $k > 3$

Domanda 13:

Il metodo *bootstrap* si basa su campionamento con ripetizione.

a) Vero
b) Falso

c) Dipende dalla sua implementazione
d) Dipende dalla dimensione dei dati

Domanda 14:

Quali delle seguenti affermazioni è sempre vera:

a) Una rete neurale ottimizza una funzione obiettivo convessa
b) Una rete neurale viene addestrata con il metodo del gradiente stocastico
c) Una rete neurale può usare differenti tipi di funzioni di attivazione
d) Una rete neurale ha almeno 2 strati nascosti

Domanda 15:

L'errore di classificazione di una rete neurale convoluzionale descresce ad ogni epoca di addestramento.

a) Vero
b) Falso
c) Dipende dalla dimensione della rete neurale
d) Dipende dalla dimensione dei dati

Domanda 16:

Un algoritmo di classificazione viene addestrato su una percentuale p dei dati disponibili e testato sulla resto dei dati (percentuale 1-p). Quale è il comportamento atteso dell'errore sull'insieme di test all'aumentare di p?

a) Aumenta sempre
b) Diminuisce sempre
c) Diminuisce o non aumenta
d) Aumenta o non diminuisce

Domanda 17:

Si consideri il seguente algoritmo "inizializza" che costruisce una stringa di lunghezza n i cui caratteri sono tutti uguali ad '*A*':

```
algoritmo inizializza (n) {
        inizializza s a una stringa vuota
        for i = 1 to  n
                s = concatena(s,"A")
        return s;
 }
```

Si dica quale delle seguenti relazioni descrive il tempo di esecuzione dell'algoritmo "inizializza", sotto l'ipotesi che la funzione "concatena" (x,y), che appende la stringa y al termine di x, abbia una complessità $\Theta(|x|+|y|)$ pari alla somma delle lunghezze delle stringhe x e y.

a) $T(n) = \Theta(n)$
b) $T(n) = \Theta(n^2)$
c) $T(n) = \Theta(n^3)$
d) Nessuna delle precedenti

Domanda 18:

Si consideri il problema della ricerca di un elemento in un insieme. A seconda della struttura dati utilizzata, si specifichi la funzione f(n) che descrive il tempo di esecuzione $\Theta(f(n))$ dell'operazione di ricerca nel caso peggiore, utilizzando il miglior algoritmo noto.

Operazione di ricerca	f(n)
In un array qualunque	
In un array ordinato	
In un albero binario di ricerca bilanciato (ad esempio, AVL o red-black)	

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 peggiore, è comunque possibile che su qualche istanza termini in O(n) passi
b) Se si dimostra che un algoritmo ha tempo di esecuzione $\Omega(n^2)$ nel caso migliore, è comunque possibile che su qualche istanza termini in O(n) passi
c) Esistono algoritmi di ordinamento basati su confronti che impiegano tempo O(n) nel caso peggiore
d) Ogni algoritmo per il calcolo del cammino minimo tra due nodi in un grafo con pesi positivi richiede tempo almeno esponenziale 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 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 $n=2^{30}$ nodi e $m=2^{40}$ archi. Si dica quale delle seguenti affermazioni è vera:

a) Al più 2^{21} nodi di G possono avere grado $\geq 2^{20}$
b) Il grafo deve necessariamente essere un albero
c) Il grado medio è 2^{20}
d) In un tale grafo, tutti i nodi devono avere lo stesso grado

Domanda 22:

Dire quale delle seguenti affermazioni è vera. In una rete che esibisce il fenomeno di *Small World*:

a) Il diametro è piccolo rispetto al numero di nodi
b) Il grado medio dei nodi è elevato
c) La densità è elevata
d) La distribuzione dei gradi segue una legge binomiale

Domanda 23:

Usando il sistema decimale (potenze in base 10), 4500 *Megabyte* sono equivalenti a:

a) 4500000 Byte
b) 4,5 GB
c) 45000 KB
d) 0,45 TB

Domanda 24:

Nella modellazione di un *database* di studenti, quale potrebbe essere una chiave primaria?

a) Numero di matricola
b) Media dei voti conseguiti agli esami
c) Città di nascita
d) Data di nascita

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 i nomi e gli indirizzi dei dipendenti che guadagnano più di 42,000€?

(a)
SELECT e.employee_name, e.address
FROM employee e, works w
WHERE w.salary > 42,000 AND w.employee_name = e.employee_name

(b)
SELECT e.employee_name, e.address
FROM employee e, works w
WHERE w.salary > 42,000 AND w.employee_name = w.employee_name

(c)
SELECT e.employee_name, e.address
FROM employee e, works w
WHERE e.salary > 42,000 AND w.employee_name = e.employee_name

(d)
SELECT e.employee_name
FROM employee e, works w
WHERE w.salary > 42,000 AND w.employee_name = e.employee_name

Domanda 26:

Dovendo generare la seguente sequenza di numeri, quale espressione è corretta in Python?
Sequenza: -10 -6 -2 2 6 10 14 18

a) range (-10, 19, 3)
b) range (-11, 20, 4)
c) range (-10, 20, 4)
d) range (-10, 19)

Domanda 27:

Quale delle seguenti affermazioni è vera?

a) La conversione esplicita di tipo viene eseguita automaticamente dall'interprete Python
b) Le funzioni di conversione di tipo in Python permettono di convertire direttamente un tipo di dati in un altro
c) Python non sempre può evitare la perdita di dati nelle conversioni automatiche di tipo
d) Quando un programmatore esegue il casting di tipo, non è possibile avere perdita di dati

Domanda 28:

In Python, un *data frame* è:

a) Una struttura dati n-dimensionale
b) Una lista di liste di valori scalari

c) Una struttura gerarchica
d) Nessuna delle precedenti

Domanda 29:

In Python, se A è una lista di quattro valori tutti uguali a 1, e B è una lista di quattro valori tutti uguali a 0, quale sarà il contenuto di A e B dopo l'esecuzione del seguente codice?

```
for i in range(len(A)):
    for j in range(i+1):
        B[i] = B[i] + A[j]
        A[i] = A[i] * 2
```

a) Non saranno modificati
b) A sarà [2,4,8,16] e B sarà [1,4,10,22]
c) A sarà [2,2,2,2] e B sarà [1,4,6,8]
d) Non si può eseguire il codice perché c'è un errore sintattico

Domanda 30:

Quale dei seguenti tipi non è un tipo di dato fondamentale (*core*) in Python?

a) Lista
b) Tupla
c) Classe
d) Dizionario

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	148	26
POS	29	69

Tabella 2: *Algoritmo B, Risultati Cross Validation*

	NEG	POS
NEG	156	18
POS	19	79

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	75	11
POS	30	180

Tabella 4

Algoritmo B, Training

	NEG	POS
NEG	82	4
POS	12	198

Algoritmo A, Testing

	NEG	POS
NEG	128	19
POS	51	306

Algoritmo B, Testing

	NEG	POS
NEG	125	22
POS	74	283

Domanda 33:

Avete svolto l'esame scritto del corso di Algoritmi, cui hanno partecipato 100 studenti. Dopo aver raccolto gli elaborati stampati su fogli A4, dovete ora ordinarli *in ordine alfabetico*. Quale algoritmo utilizzereste: Bubblesort, Mergesort 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 *min-heap* (*heap* binario in cui il minimo è memorizzato nella radice) contenente n valori interi distinti. Discutere dove può trovarsi il *terzo intero più piccolo*, usando al massimo 200 parole.