ΕΡΩΤΗΣΕΙΣ ΣΥΝΤΟΜΗΣ ΑΝΑΠΤΥΞΗΣ

1. Αναφέρετε τους τύπους Αναλυτικής και περιγράψτε συνοπτικά την κάθε μια, μαζί με ένα παράδειγμα.

2. Αναφέρετε τους τύπους των Μεγάλων Δεδομένων, περιγράψτε συνοπτικά τον κάθε τύπο μαζί με ένα παράδειγμα.

3. Ποια είναι τα χαρακτηριστικά των Μεγάλων Δεδομένων που γνωρίζετε; Γράψτε μια σύντομη ανάπτυξη για το καθένα.

4. Σε ποιους τύπους δεδομένων πρέπει να δίνουμε προτεραιότητα στο Ψηφιακό Μάρκετινγκ; Γράψτε μια σύντομη ανάπτυξη για τον καθένα.

5. Ποια εργαλεία αξιοποίησης Μεγάλων Δεδομένων γνωρίζετε; Αναπτύξτε σύντομα το καθένα.

6. Ποια είναι τα πλεονεκτήματα των Παράλληλων Βάσεων Δεδομένων;

7. Πως λειτουργεί ο αλγόριθμος MapReduce; Αναφέρετε ένα κλασικό παράδειγμα.

8. Ποια είναι τα δομικά στοιχεία των Apache Hadoop; Γράψτε μια σύντομη ανάπτυξη για το καθένα.

9. Πως λειτουργεί το Apache Hadoop;

10. Ποια είναι η δομή του Apache Spark; Γράψτε μια σύντομη περιγραφή για κάθε δομικό στοιχείο του.

11. Ποιες μεθόδους clustering γνωρίζετε; Περιγράψτε την κάθε μια.

12. Από ποια βήματα αποτελείται ο αλγόριθμος k-means; Δώστε ένα παράδειγμα.

13. Δίνεται ο ακόλουθος πίνακας γειτνίασης:

Στοιχείο	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	3	4	3	0	3
E	3	3	5	3	0

Εφαρμόστε **NN Clustering** με κατώφλι **t=2**. Αποτυπώστε όλα τα ενδιάμεσα βήματα.

14. Δίνεται ο παρακάτω πίνακας:

NAME	INCOME (EUROS)	OUTPUT 1
A	1600	**LOW**
B	2000	HIGH
C	1900	MEDIUM
D	1880	MEDIUM
E	1700	**LOW**
F	1850	MEDIUM
G	1600	**LOW**
H	1700	**LOW**
I	2200	HIGH
J	2100	HIGH
K	1800	MEDIUM
L	1950	MEDIUM
M	1900	MEDIUM
N	1800	MEDIUM
O	1750	**MEDIUM**

Θεωρώντας τα δεδομένα του output1 ως δεδομένα εκπαίδευσης, θέλουμε να κατηγοριοποιήσουμε την πλειάδα <P, 1600>. Εδώ η απόσταση είναι η απλά η απόλυτη τιμή της διαφοράς των τιμών του ύψους (ευκλείδειο και Manhattan μέτρο απόστασης έχουν ακριβώς τα ίδια αποτελέσματα). Εφαρμόστε τον **kNN Classifier** για K=5. Αποτυπώστε όλα τα ενδιάμεσα βήματα.

15. Σας δίνεται ο παρακάτω κώδικας σε PySpark:

```
from pyspark.sql import SparkSession
# Δημιουργία SparkSession
spark = SparkSession.builder.appName("SparkSQLExample").getOrCreate()

from pyspark.sql import Row

data = [

Row(id=1, name="Alice", age=25, city="Athens"),

Row(id=2, name="Bob", age=30, city="Thessaloniki"),

Row(id=3, name="Charlie", age=35, city="Patras")

]

df = spark.createDataFrame(data)

df.createOrReplaceTempView("people")
```

Απαντήστε, γράφοντας κώδικα σε Spark SQL, στα παρακάτω ερωτήματα:

Q1:Επιλογή όλων των δεδομένων του πίνακα people

```
result1 = spark.sql("SELECT * FROM people")

result1.show()
```

Q2: Φιλτράρισμα δεδομένων (άτομα με ηλικία > 28)

```
result2 = spark.sql("SELECT name, age FROM people WHERE age > 28")

result2.show()
```

Q3: Ταξινόμηση των δεδομένων κατά ηλικία

```
result3 = spark.sql("SELECT * FROM people ORDER BY age DESC")

result3.show()
```

Q4: Ομαδοποίηση των δεδομένων κατά πόλη και υπολογισμός πλήθους

```
result4 = spark.sql("SELECT city, COUNT(*) as count FROM people GROUP BY city")
```

result4.show()

16.Δίνονται τα παρακάτω δυαδικά διανύσματα

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

Βρείτε την ομοιότητα των 2 διανυσμάτων με τις μεθόδους: (α) Simple Matching (β) Συντελεστές Jaccard

ΛΥΣΗ

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1
$M_{01} = 2$
$M_{10} = 1$
$M_{00} = 7$
$M_{11} = 0$

$SM = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$
$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

17.Στα παρακάτω κείμενα D1, D2 και D3, βρείτε με τη μέθοδο COSINE SIMILARITY το ζευγαράκι κειμένων με τη μεγαλύτερη ομοιότητα.

ΠΑΡΑΔΕΙΓΜΑ

D1: "Shipment of gold damaged in a fire"

D2: "Delivery of silver arrived in a silver truck"

D3: "Shipment of gold arrived in a truck

Doc ID	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
D1	1	0	1	0	1	1	1	1	1	0	0
D2	1	0	0	1	0	0	1	1	0	2	1
D3	1	1	0	0	0	1	1	1	1	0	1

REMINDER:

Αν d_1 και d_2 είναι δυο διανύσματα εγγράφων τότε:

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

όπου • σημαίνει εσωτερικό γινόμενο και $||d||$ το μήκος του διανύσματος d.

Παράδειγμα:

$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

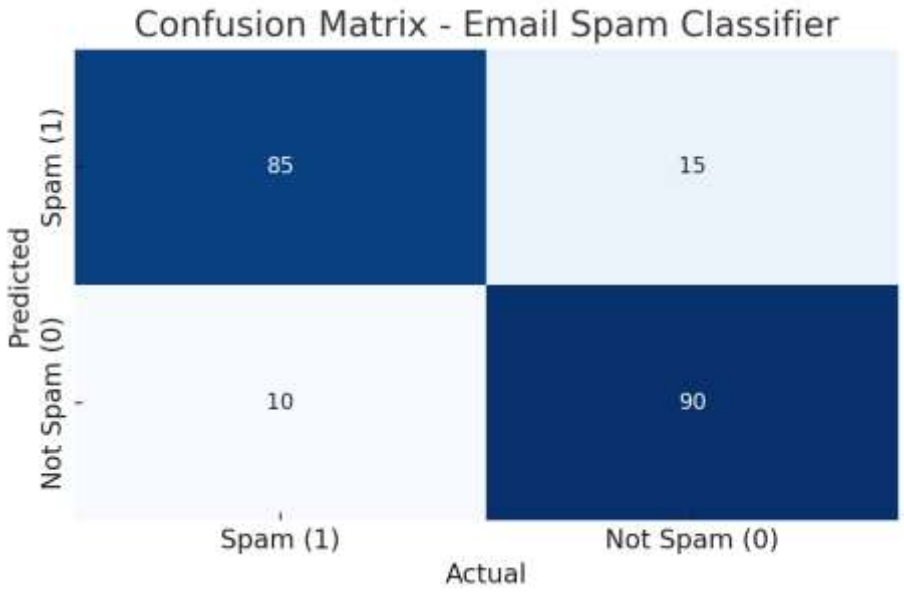
$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

18. Ένας ταξινομητής χρησιμοποιείται για την αναγνώριση ανεπιθύμητων μηνυμάτων (spam). Ο παρακάτω πίνακας σύγχυσης παρουσιάζει τα αποτελέσματα της ταξινόμησης:



Ζητούμενο:

Αξιοποιώντας τον πίνακα, υπολογίστε τις παρακάτω μετρικές αξιολόγησης:

1. **Accuracy (Ακρίβεια συνολική)**
2. **Precision (Ακρίβεια θετικών προβλέψεων)**
3. **Recall (Ανακλητικότητα)**
4. **F1 Score**
5. **False Positive Rate (FPR)**
6. **False Negative Rate (FNR)**

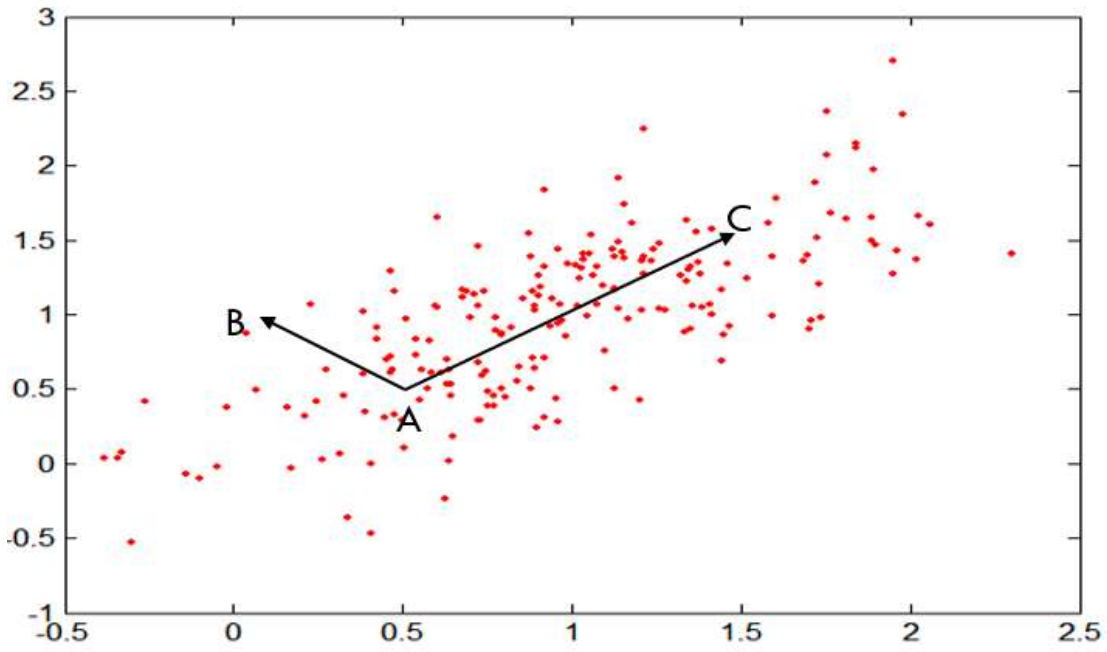
19. Μετρικές Απόστασης

(α) Δίνεται ο παρακάτω πίνακας 4 διανυσμάτων (embeddings).

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Να υπολογίσετε τις αποστάσεις σύμφωνα με τις L1, L2 και L∞ Νόρμες, μεταξύ όλων των δυνατών ζευγαριών:

(β) Τί εκφράζει η Mahalanobis Distance? Στο παρακάτω σχήμα ισχύει Mahal(A,B) > Mahal(A,C)? Ναι, όχι και γιατί.



ΛΥΣΗ

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

(β) Η Mahalanobis distance μεταξύ 2 σημείων Α και Β εκφράζει τη «δυσκολία» «μετακίνησης» από το Α στο Β και επομένως συνδέεται άμεσα με την πυκνότητα διασποράς των σημείων (διανυσμάτων) στο χώρο. Άρα, Mahal(A,C)<Mahal (A,B).

20.

Έχουμε ένα σύστημα ταξινόμησης email σε δύο κατηγορίες:

- **Spam**
- **Not Spam**

Από ένα σύνολο εκπαίδευσης γνωρίζουμε ότι:

Πιθανότητα	Τιμή
P(Spam)	0.40
P(Not Spam)	0.60
P("free" | Spam)	0.70
P("free" | Not Spam)	0.10
P("offer" | Spam)	0.60
P("offer" | Not Spam)	0.20

Έρχεται νέο email που περιέχει τις λέξεις:

"free offer"

Να ταξινομηθεί το email χρησιμοποιώντας **Naive Bayesian Classifier**.

Θεωρήστε ότι οι λέξεις είναι ανεξάρτητες μεταξύ τους.

Ζητείται

1. Να υπολογίσετε το σκορ για την κατηγορία **Spam**.
2. Να υπολογίσετε το σκορ για την κατηγορία **Not Spam**.
3. Να αποφασίσετε σε ποια κατηγορία ανήκει το email.

Λύση

Για **Spam**:

$$P(Spam) \cdot P(free|Spam) \cdot P(offer|Spam)$$

$$0.40 \cdot 0.70 \cdot 0.60 = 0.168$$

Για **Not Spam**:

$$P(NotSpam) \cdot P(free|NotSpam) \cdot P(offer|NotSpam)$$

$$0.60 \cdot 0.10 \cdot 0.20 = 0.012$$

Επειδή:

$$0.168 > 0.012$$

το email ταξινομείται ως:

Spam.

21. Δίνεται ο παρακάτω πίνακας:

x	y
x1	y1
x2	y2
……	……
x7	y7

Όπου x=input και y=output. Να εφαρμόσετε γραμμική παλινδρόμηση (linear regression) για να υπολογίσετε την y=f(x).

Υπόδειξη: The equation can be found in any statistical analytics text book: $y=m\cdot x+b$

Where ***m*** **is called the slope** which determines the direction of the line and ***b*** **is called the intercept** which defines the distance of the line from 0.

$m=\sum(x_i - ẍ)(y_i - ÿ) / \sum(x_i - ẍ)^2 = [(x1- ẍ)(y1- ÿ)+(x2- ẍ)(y2- ÿ)+….+(x7- ẍ)(y7- ÿ)] / [(x1- ẍ)(x1- ẍ)+(x2- ẍ)(x2- ẍ)+….+(x7- ẍ)(x7- ẍ],$

$b=max\ y - max\ x \cdot m,$ where $ẍ$ and $ÿ$ are the averages of x and y.

ΕΡΩΤΗΣΕΙΣ ΠΟΛΛΑΠΛΗΣ ΕΠΙΛΟΓΗΣ

1. Ποια από τις ακόλουθες δηλώσεις περιγράφει καλύτερα τα μεγάλα δεδομένα;

α) Δεδομένα που είναι αποθηκευμένα σε μία μόνο τοποθεσία

β) Σύνολα δεδομένων μικρού μεγέθους που μπορούν να αναλυθούν εύκολα

γ) Μεγάλα και πολύπλοκα σύνολα δεδομένων που απαιτούν εξειδικευμένα εργαλεία και τεχνικές

δ) Δεδομένα που είναι εύκολα διαχειρίσιμα με τη χρήση παραδοσιακών βάσεων δεδομένων

2. Ποια τεχνολογία χρησιμοποιείται συνήθως για την επεξεργασία μεγάλων δεδομένων;

α) Εικονική πραγματικότητα

β) Blockchain

γ) Τεχνητή νοημοσύνη

δ) Hadoop

3. Ποιος είναι ο όρος που χρησιμοποιείται για να περιγράψει τη διαδικασία εξαγωγής χρήσιμων πληροφοριών από μεγάλα δεδομένα;

α) Εξόρυξη δεδομένων

β) Αποθήκευση δεδομένων

γ) Μοντελοποίηση δεδομένων

δ) Οπτικοποίηση δεδομένων

4. Ποια είναι η έννοια που αναφέρεται στην ικανότητα επεξεργασίας και ανάλυσης δεδομένων σε πραγματικό χρόνο;

α) Επεξεργασία παρτίδων

β) Επεξεργασία ροής

γ) Οπτικοποίηση δεδομένων

δ) Ολοκλήρωση δεδομένων

5. Ποιο από τα ακόλουθα αποτελεί παράδειγμα μη δομημένων δεδομένων;

α) Φύλλο εργασίας

β) Σχεσιακή βάση δεδομένων

γ) Μηνύματα ηλεκτρονικού ταχυδρομείου

δ) Δεδομένα αισθητήρων

6. Ποιο από τα ακόλουθα ΔΕΝ αποτελεί δυνητικό όφελος της ανάλυσης μεγάλων δεδομένων;

α) Βελτιωμένη λήψη αποφάσεων

β) Βελτιωμένη εμπειρία πελάτη

γ) Μειωμένο κόστος αποθήκευσης δεδομένων

δ) Αυξημένοι κίνδυνοι κυβερνοασφάλειας