**מבוא למדעי הנתונים שיעור ראשון**

חצי ראשון של השיעור- אשלים(כנס אימפקט)

**מה זה דאטה?**

דאטה (Data) הוא אוסף של עובדות, מספרים, טקסטים או מדידות.  
לדוגמה:

* רשימה של ציונים של תלמידים
* כמויות מוצרים שנמכרו בחנות
* טמפרטורות שנמדדו בעיר במהלך השנה

**📌 דוגמאות לסוגי דאטה**

1. **דאטה מובנה (Structured Data)** – טבלאות כמו Excel או מסדי נתונים (SQL).
2. **דאטה לא מובנה (Unstructured Data)** – טקסטים, תמונות, סרטונים, אודיו.
3. **דאטה חצי-מובנה (Semi-Structured Data)** – JSON, XML (משמש לאחסון מידע).

**🔹 מה זה מדעי הנתונים (Data Science)?**

מדעי הנתונים זה תחום שמשלב **סטטיסטיקה, תכנות וניתוח נתונים** כדי להסיק מסקנות.  
מטרתו להבין מגמות מתוך הנתונים ולקבל החלטות חכמות.

**📌 דוגמאות ליישומים במדעי הנתונים**

✔️ המלצות סרטים (Netflix)  
✔️ זיהוי פנים בטלפון  
✔️ ניתוח מגמות כלכליות

**🔹 ההבדל בין דאטה למידע**

דאטה הם רק עובדות גולמיות, אך **מידע** הוא פרשנות לדאטה.  
לדוגמה:

* **דאטה:** מספר הצפיות בסרטון מסוים הוא 10,000.
* **מידע:** הסרטון הזה פופולרי מאוד בקרב צעירים בגילאי 18-25.

**🔹 הקשר לסטטיסטיקה**

סטטיסטיקה עוזרת לנו להבין דאטה על ידי ניתוח מגמות, חישוב ממוצעים, סטיות תקן ועוד.  
למשל: אם אנחנו רוצים לדעת מהי **המשכורת הממוצעת בחברה**, נשתמש בנוסחאות סטטיסטיות.

**🔸 תרגול מעשי (Python)**

💻 **נכתוב קוד שמחשב ממוצע ציונים של תלמידים:**

python

CopyEdit

# רשימה של ציונים

grades = [85, 90, 78, 92, 88]

# חישוב ממוצע

average = sum(grades) / len(grades)

# הצגת התוצאה

print("ממוצע הציונים הוא:", average)

**👉 תוצאה צפויה:**

CopyEdit

ממוצע הציונים הוא: 86.6

תריץ את הקוד ב-Google Colab או ב-Jupyter Notebook ותראה איך זה עובד! 😊

**🔹 לסיכום השיעור:**

✅ דאטה הוא אוסף של עובדות ומדידות.  
✅ מדעי הנתונים הוא תחום שמשתמש בסטטיסטיקה וניתוח נתונים להסיק תובנות.  
✅ ההבדל בין דאטה למידע הוא שהמידע נותן משמעות לנתונים.  
✅ סטטיסטיקה עוזרת לנו להבין נתונים ולנתח אותם.

**מה זה אשכולות (Clustering)?**

אשכולות הוא תהליך בו אנחנו מנסים לחלק נתונים לקבוצות (אשכולות) על פי דמיון ביניהם. כל נתון בקבוצה נחשב לדומה יותר לשאר הנתונים באותה קבוצה מאשר לנתונים מקבוצות אחרות.  
כדי לעשות זאת, אנחנו משתמשים באלגוריתמים כמו **K-Means** שיבצעו את החלוקה הזו באופן אוטומטי.

**אלגוריתם k-מרכזים** (k-means) הוא שיטה פופולרית עבור [ניתוח אשכולות](https://he.wikipedia.org/wiki/%D7%A0%D7%99%D7%AA%D7%95%D7%97_%D7%90%D7%A9%D7%9B%D7%95%D7%9C%D7%95%D7%AA) (Clustering) ב[כריית נתונים](https://he.wikipedia.org/wiki/%D7%9B%D7%A8%D7%99%D7%99%D7%AA_%D7%A0%D7%AA%D7%95%D7%A0%D7%99%D7%9D). מטרתו לחלק את התצפיות ל-k אשכולות לפי מרכזי כובד (k-means). כל תצפית משויכת לאחד מ"מרכזי הכובד". על ידי בחירה נכונה של מרכזי כובד ניתן לאתר את הקבוצות השונות.[[1]](https://he.wikipedia.org/wiki/%D7%90%D7%9C%D7%92%D7%95%D7%A8%D7%99%D7%AA%D7%9D_k-%D7%9E%D7%A8%D7%9B%D7%96%D7%99%D7%9D#cite_note-:0-1)

נדרשות תצפיות רבות על מנת להשתמש במודל ותוספת של תצפיות חדשות עשויה לחייב חישוב חוזר. מדובר באלגוריתם [היוריסטי](https://he.wikipedia.org/wiki/%D7%99%D7%95%D7%A8%D7%99%D7%A1%D7%98%D7%99%D7%A7%D7%94) שמשתמשים בו כדי לבצע חישובים שמובילים להתכנסות לפתרון מקומי (שאינו בהכרח הפתרון הטוב ביותר)

חצי שני:

תמונה שמכילה טקסט, צילום מסך, לוח כתיבה לבן, בתוך מבנה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

בדיקת סדר מטריצה פקודת – X.shape

וקטורים = lables

x.shape בעזרת הפקודה מקבלים את מספר השורות והעמודות שיש במטריצה

Plot.scatter(X[ : ,0],X[ : ,1]);

הסבר על הקוד:

X זה שם המטריצה

קטע קוד הסבר (X[ : ,0]:

x זה השם של המטריצה ,בתוך הסוגריים המרובעים יש מקום לשתי עמודות,העמודה הראשונה מסמלת את השורות, העמודה השנייה מסמלת את הטורים. בעמודה הראשונה הצבנו נקודותיים – פקודה שמחלצת את כל המידע ,כלומר מחלצת את כל השורות בעמודה מספר 0. בעמודה השנייה הצבנו 0 שזה בעצם מספר העמודה שמתוכה אנחנו מחלצים את המידע.

קטע קוד (X[ : ,1]:

x זה השם של המטריצה , בתוך הסוגרים המרובעים מחלצים את כל השורות של העמודה השנייה (מיקום 1= עמודה שנייה).

Plot.scatter(X[:

כאן בעצם יוצרים גרף מהמטריצה באמצעות פקודת scatter של mcplotlib.

שיעור 2

אם לא עושים פסיק רק במטריצה כלומר

X[:5]

אז יש התייחסות לכל העמודות.

**📌 שלבים ראשוניים בעבודה עם דאטה**

1. **טעינת הדאטה**
2. **הצגת נתונים כללית**
3. **חישוב מדדים סטטיסטיים**
4. **ניקוי נתונים במידת הצורך**
5. תמונה שמכילה טקסט, צילום מסך, גופן, דף אינטרנט

   תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.תמונה שמכילה טקסט, צילום מסך, גופן, דף אינטרנט

   תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**יצירת ויזואליזציות**

תמונה שמכילה טקסט, צילום מסך, גופן, דף אינטרנט

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.plt.head היא פקודה בספריית pandas שמראה את 5 השורות הראשונות של הdata.

שיעור 3

עד 11 להשלים

מ11:

Numpy היא ספרייה שמריצה קוד במהירות של c (אם כותבים בלי לולאות פור).

Axis= ציר

axis 0 מחלץ את הטורים(וקטורים)

axis 1 מחלץ את השורות (נניח יש לנו 2 וקטורים ו5 שורות אז הוא מחלץ את השורות מתוך כל וקטור)

מציאת k אופטימלי באשכול

ראינו כי שיטת המרפק איננה טובה מספיק. נרצה שיטה מדויקת / אמינה יותר.

בשיטת המרפק הסתכלנו רק על המרחקים בתוך אותו האשכול.