

תשובות לתרגיל 22 – DATA SCIENCE

מגיש דוד פתאל

02-11-2024

1. DATA SCIENCE הוא תחום בינתחומי המשתמש בשיטות, תהליכים, אלגוריתמים ומערכות מדעיות כדי לחלץ ידע ותובנות מנתונים מובנים ובלתי מובנים. הוא משלב אלמנטים מסטטיסטיקה, מדעי המחשב ומומחיות בתחום כדי לנתח ולפרש מערכי נתונים מורכבים, ובסופו של דבר עוזר לארגונים לקבל החלטות ותחזיות מושכלות.

2.

ההבדלים העיקריים בין נתונים לא מובנים לנתונים מובנים הם:

1. פורמט:

- נתונים מובנים: מאורגנים בפורמט מוגדר מראש, בדרך כלל בשורות ובעמודות (כמו מסדי נתונים או גיליונות אלקטרוניים). הוא כולל סוגי נתונים כמו מספרים, תאריכים ומחרוזות.

- נתונים לא מובנים: חסר פורמט או מבנה מוגדרים מראש, מה שהופך אותם למורכבים יותר לניתוח. זה כולל טקסט, תמונות, סרטונים, פוסטים במדיה חברתית ומיילים.

2. אחסון:

- נתונים מובנים: מאוחסנים במסדי נתונים יחסיים וניתן לגשת אליהם ולנהל אותם בקלות באמצעות SQL.
- נתונים לא מובנים: מאוחסנים בדרך כלל במסדי נתונים לא יחסיים או באגמי נתונים ודורשים פתרונות אחסון מורכבים יותר.

3. ניתוח:

- נתונים מובנים: קל יותר לנתח באמצעות כלים וטכניקות ניתוח נתונים מסורתיים בשל אופיו המאורגן.
- נתונים לא מובנים: דורש טכניקות ניתוח מתקדמות, כגון עיבוד שפה טבעית (NLP) או למידת מכונה, כדי להפיק תובנות.

4. דוגמאות:

- נתונים מובנים: שמות לקוחות, רשומות עסקאות ונתוני מלאי.
- נתונים לא מובנים: מסמכי טקסט, תמונות, קבצי שמע ותוכן מדיה חברתית.

5. מקרי שימוש:

- נתונים מובנים: מתאים ליישומים הדורשים שאלות ודיווח נתונים קפדניים, כמו מערכות פיננסיות.
- נתונים לא מובנים: ערך עבור תובנות לגבי סנטימנט לקוחות, מגמות והתנהגויות, המשמשים לעתים קרובות בשיווק ובשירות לקוחות.

הבדלים אלה משפיעים על האופן שבו ארגונים מנהלים, מנתחים ומנצלים נתונים לצורך קבלת החלטות.

3.

EDA הוא שלב קריטי בתהליך ניתוח הנתונים הכולל בדיקה וסיכום של מערכי נתונים כדי לגלות דפוסים, לזהות חריגות, לבחון השערות ולהשיג תובנות. המטרות העיקריות של EDA הן:

1. **הצגת נתונים חזותיים:** שימוש בגרפים ותרשימים כדי להמחיש התפלגות נתונים וקשרים בין משתנים.

2. **סטטיסטיקות בסיסיות:** חשב נתונים סטטיסטיים בסיסיים, כגון ממוצע, חציון, שונות וכמותיות, כדי להבין את הנתונים והשונות המרכזיות של הנתונים.

3. **זיהוי קשרים:** חיפוש מתאמים ואסוציאציות בין משתנים כדי להבין כיצד הם מקיימים אינטראקציה זה עם זה.

4. **זיהוי חריגים:** זהה נקודות נתונים חריגות שעשויות להצביע על שגיאות או תופעות מעניינות שכדאי לחקור עוד יותר.

5. **הערכת איכות הנתונים:** בדוק אם חסרים ערכים, כפילויות וחוסר עקביות שיש לטפל בהם לצורך ניתוח מדויק.

6. **יצירת השערות:** ניסוח שאלות או השערות פוטנציאליות על סמך תצפיות כדי להנחות ניתוח נוסף.

בסך הכל, EDA מסייעת למדעני נתונים ואנליסטים לבנות הבנה מוצקה של הנתונים, ומאפשרת מודלים מושכלים יותר וקבלת החלטות בשלבים הבאים של הניתוח.

4.

בחברות עסקיות, שתי המטרות של מדעי הנתונים הן:

1. **שיפור קבלת ההחלטות:** מדעי הנתונים מסייעים לארגונים לנתח נתונים כדי להפיק תובנות ניתנות לפעולה, המאפשרות החלטות מושכלות יותר. על ידי שימוש בניתוח, חיזוי והדמיית נתונים, חברות יכולות לזהות מגמות, לחזות תוצאות ולהעריך סיכונים, מה שמוביל לתכנון אסטרטגי ויעילות תפעולית טובה יותר.

2. **שיפור חווית הלקוח:** באמצעות ניתוח נתונים, חברות יכולות להבין את התנהגות והעדפות הלקוחות, מה שמאפשר שיווק מותאם אישית, המלצות למוצרים ושיפור שירות הלקוחות. על ידי מינוף תובנות מנתוני לקוחות, עסקים יכולים להתאים את ההצעות שלהם כדי לענות על צרכי הלקוחות בצורה יעילה יותר, ובסופו של דבר להגביר את שביעות הרצון והנאמנות של הלקוחות.

יעדים אלו מאפשרים לחברות למנף את הנתונים שלהן כנכס אסטרטגי, מטפח צמיחה ותחרותיות בשוק.

5.

כלי AI גנרטיבי הם יישומים המשתמשים בטכניקות בינה מלאכותית גנרטיבית ליצירת תוכן חדש, בין אם זה טקסט, תמונות, אודיו או צורות אחרות של מדיה. כלים אלה בנויים על מודלים מתקדמים של למידת מכונה, במיוחד ארכיטקטורות למידה עמוקה כמו רשתות עצביות. להלן כמה סוגים נפוצים ודוגמאות לכלי AI גנרטיביים:

1. יצירת טקסט:

- צ'אטבוטים ועוזרים וירטואליים: כלים כמו ChatGPT יכולים ליצור תגובות כמו אנושיות בשיחות.
- כלים ליצירת תוכן: יישומים המסייעים בכתיבת מאמרים, בלוגים או עותק שיווקי, כגון Jasper או Writesonic.

2. יצירת תמונות:

- מחוללי אמנות AI: כלים כמו DALL-E ו-Midjourney יכולים ליצור תמונות מתיאורים טקסטואליים.
- כלים לשיפור תמונות: תוכנה שיכולה לשפר את איכות התמונה או לשנות תמונות בדרכים יצירתיות.

3. יצירת מוזיקה ואודיו:

- כלים שיוצרים יצירות מוזיקה או אפקטים קוליים, כמו MuseNet של OpenAI או AIVA.

4. יצירת וידאו:

- תוכנה שיכולה לייצר תוכן וידאו מסקרפטים או להפוך תהליכי עריכת וידאו לאוטומטיים.

5. יצירת קוד:

- כלים המסייעים למפתחים על ידי יצירת קטעי קוד או תוכניות שלמות על סמך מפרטים, כמו GitHub Copilot.

כלי בינה מלאכותית גנרטיבית נמצאים בשימוש יותר ויותר בתעשיות עבור יישומים יצירתיים, יצירת תוכן, ואפילו אוטומציה של משימות חוזרות ונשנות, שיפור הפרודוקטיביות והחדשנות.

6.

Pandas מאפשרת מניפולציה וניתוח נתונים ברמה גבוהה, ומציעה מבני נתונים ופעולות למניפולציה של נתוני סדרות זמן וטבלאות מספריות.

NumPy מסייעת בחישובים מתמטיים מורכבים ותומכת במערכים רב מימדיים גדולים ומטריצות.

Matplotlib היא ספרייה קריטית נוספת המאפשרת הדמיית נתונים יעילה, יצירת סטטיסטיקה והנפשות.

7.

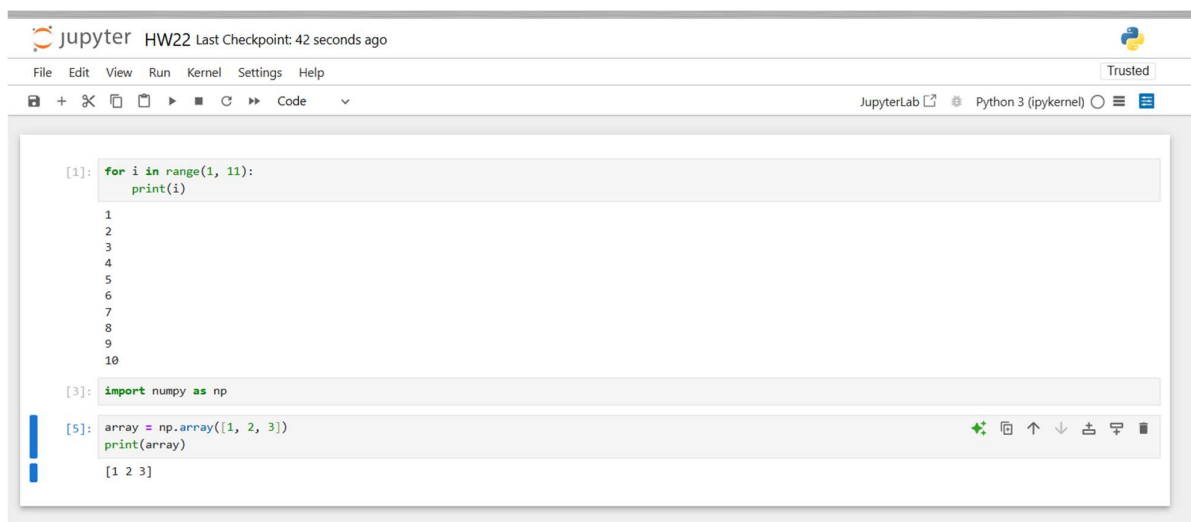
Jupyter Notebook הוא יישום אינטרנט בקוד פתוח המאפשר יצירה ושיתוף של מסמך המכיל קוד חי, משוואות, הדמיות וטקסט נרטיבי.

Jupyter Notebook מאפשר לנו ליצור מחברות אינטראקטיביות ניתנות לשיתוף עם קטעי קוד שניתן לרוץ באופן אינטראקטיבי, לצד הסברים והדמיות. בשל תכונות אלה, כלי זה הפך למאוד פופולרי ויעיל.

Jupyter Notebook Markdown היא דרך לעצב טקסט בתוך Jupyter Notebooks באמצעות Markdown, שפת סימון קלה. זה מאפשר למשתמשים ליצור תיעוד מובנה וקריא לצד הקוד שלהם, מה שהופך את המחברות ליותר אינפורמטיביות ומושכות מבחינה ויזואלית.

באמצעות ! לפני הפקודה מתאפשר לך להפעיל פקודות Terminal או Shell ישירות מהקוד. שימושי לביצוע פקודות מערכת מבלי לצאת מסביבת המחברת.

8.



```
[1]: for i in range(1, 11):
      print(i)
1
2
3
4
5
6
7
8
9
10

[3]: import numpy as np

[5]: array = np.array([1, 2, 3])
      print(array)
[1 2 3]
```