Clustering – K-means

Clustering – K-means

אתם עובדים כחברי צוות אנליסטים בחברת – "CardWise" חברת כרטיסי אשראי מובילה בישראל.
בשנים האחרונות חל גידול משמעותי בבסיס הלקוחות של החברה, אך יחד עם ההתרחבות ניצבה החברה בפני אתגר: להבין טוב יותר את התנהגות הלקוחות השונים כדי לייעל שיווק, למזער סיכונים ולאפשר הצעות ממוקדות.

בהתבסס על מערך הנתונים הפנימי של החברה (קובץ Customer Data.csv), תנסו לזהות קבוצות לקוחות דומות זו לזו (segments) באמצעות שיטות clustering, בעיקר KMeans. כך תוכלו בתום התרגיל להציע לקהלי יעד מותאמים: למשל, מבצעים ללקוחות שמשתמשים בתשלומים חד-פעמיים, או הצעות אשראי ללקוחות בעלי מסגרות גבוהות אך משתמשים בתדירות נמוכה.

לאחר שתבצעו ניתוח חקר נתונים, קדם־עיבוד ובחירת מספר הקלאסטרים, המטרה היא:

1. **לספק פרופיל תמציתי לכל קבוצה**
2. **להמליץ לצוות השיווק על מסרים ומבצעים**

להלן פירוט ה-features:

תיאור	משתנה
מזהה ייחודי של הלקוח.	**CUST_ID**
סך היתרה (במטבע) בחשבון הכרטיס, ממוצע חודשי לאורך תקופת ה־TENURE.	**BALANCE**
תדירות העדכון של היתרה (0–1): ערך קרוב ל-1 אומר שהיתרה מתעדכנת כמעט בכל חודש; ערך קרוב ל-0 – לעתים רחוקות.	**BALANCE_FREQUENCY**
סכום כל הרכישות שבוצעו בכרטיס (מבטא את היקף ההוצאות החודשיות הכולל).	**PURCHASES**
סכום הרכישות שבוצעו כעסקות חד-פעמיות (ללא תשלום בתשלומים).	**ONEOFF_PURCHASES**
סכום הרכישות שבוצעו בתשלומי פיצול (installments).	**INSTALLMENTS_PURCHASES**
סכום המשיכות במזומן (Cash Advance) שבוצעו בכרטיס.	**CASH_ADVANCE**
מספר המשיכות במזומן שבוצעו.	**CASH_ADVANCE_TRX**

תיאור	משתנה
תדירות המשיכות במזומן (0–1): ערך גבוה מצביע על קיום משיכות קרובות זו לזו.	**CASH_ADVANCE_FREQUENCY**
תדירות כללית של רכישות (0–1): כמה קרובות זו לזו העסקאות ביחס לאורך ה־TENURE.	**PURCHASES_FREQUENCY**
תדירות עסקאות חד-פעמיות.(0–1)	**ONEOFF_PURCHASES_FREQUENCY**
תדירות עסקאות בתשלומים.(0–1)	**PURCHASES_INSTALLMENTS_FREQUENCY**
מספר העסקאות הכולל שבוצעו ברכישות (כולל חד-פעמיות ובתשלומים).	**PURCHASES_TRX**
מסגרת האשראי המוקצית ללקוח (Credit Card Limit).	**CREDIT_LIMIT**
סך התשלומים שבוצעו על ידי הלקוח (Pay-off Amount), לאורך תקופת ה־TENURE.	**PAYMENTS**
סכום התשלום המינימלי שדרש הבנק בכל חודש, ממוצע לאורך כל ה־TENURE.	**MINIMUM_PAYMENTS**
אחוז החודשים שבהם הלקוח שילם את היתרה במלואה.(0–1)	**PRC_FULL_PAYMENT**
משך ההחזקה של הכרטיס (במספר חודשים), לאורכם נמדדו שאר המשתנים.	**TENURE**

חלק א׳ – טעינת הנתונים ובדיקה ראשונית

1. טען את הקובץ ל־DataFrame
2. הצג info(), head() ו־describe() לקבלת תמונה על סוגי העמודות, ערכים חסרים וסטטיסטיקות כלליות.
3. בדוק האם קיימים ערכי חסר בעמודות, וכיצד תטפל בהם (הסבר שיטת טיפול – מחיקה או השמה בערך ממוצע/מדיאני).

חלק ב׳ – ניתוח חקר נתונים (EDA)

1. חשב ופרש את מטריצת הקורלציה (correlation) בין התכונות העיקריות) למשל : BALANCE, PURCHASES, CASH_ADVANCE, CREDIT_LIMIT, PAYMENTS).
2. ויזואליזציה:

- היסטוגרמות והתפלגות של BALANCE ו־PURCHASES כדי לזהות א-סימטריות או אאוטליירים.
- Scatter plot בין PURCHASES לבין CASH_ADVANCE

3. זהה נקודות קיצון (Outliers) בעזרת תיבות (boxplot) whisker עבור עמודת CREDIT_LIMIT ועמודת.CASH_ADVANCE

חלק ג' – קדם־עיבוד והתאמה למודל

1. בחר את קבוצת העמודות הרלבנטיות להצבה במודל (לדוגמה, תכונות כספיות ותדירות רכישות). בצעו מחקר קטן באמצעות deep research לראות מה יכול להיות רלוונטי ומה פחות.
2. בצע Scale/Standardization (StandardScaler) כדי לאזן את סקאלת התכונות (במידת הצורך)

חלק ד' – מציאת מספר הקלאסטרים Elbow

1. בחר את הערך האופטימלי של k והסבר את הבחירה.

חלק ה' – אימון מודל KMeans ופרופילינג של קלאסטרים

1. אמן את KMeans עם k הנבחר ו־random_state=42
2. הוסף ל-dataframe עמודת Cluster עם התוויות שהוקצו לכל לקוח.
3. חשב את המאפיינים הממוצעים (centroids) של כל קלאסטר והצג בטבלה (אפשר להשתמש ב-groupby):
   - גודל קלאסטר (מספר לקוחות)
   - ממוצע BALANCE, PURCHASES, CASH_ADVANCE, CREDIT_LIMIT, PAYMENTS
4. פרש כל קלאסטר:
   - למשל: איזה סגנון הוצאה ותדירות רכישות מאפיין אותו? האם זה "High-Value Customers", "Frequent Spenders", "Cash-Advance Seekers" וכד'?

חלק ו' – סיכום ומסקנות עסקיות

1. אילו תובנות עסקיות ניתן להסיק מהתפלגות הקלאסטרים?
2. כיצד ניתן ליישם את המודל לשיפור שיווק ממוקד, ניהול סיכוני אשראי או מבצעים מותאמים?

בהצלחה!