דוח מידול נתונים

ניהול צריכה באמצעות סריקת חשבוניות קבלה

נופר גרשוני-322773680
עלית בן חמו-206851081
למנחה-גב' קרן סגל

תוכן עניינים:

1.בחירת טכניקות המודל

1.1 בחירת טכניקות נכונות:

בשלב זה של הפרויקט, לאחר הבנה מעמיקה של סוגי המודלים המתאימים לצרכי המשתמשים, התקבלה החלטה ברורה לגבי המודלים שישמשו בניתוח הנתונים. במהלך בחירת המודלים, נבחנו מספר שיקולים מרכזיים:

חלוקת הנתונים: הנתונים מתחלקים לסטים של אימון ובדיקה (Train/Test) כדי להבטיח הערכה מהימנה של ביצועי המודלים.

כמות הנתונים: נעשה מאמץ לאסוף ולהרחיב את מאגר הנתונים ככל האפשר, כולל איסוף של מספר רב של חשבוניות קבלה מגוונות. מאגר הנתונים שנוצר, איפשר אימון המודלים בצורה שהביאה לתוצאות מהימנות.

איכות הנתונים: ביצענו עיבוד נתונים לפני תהליך המידול, כדי להבטיח שהנתונים יהיו אמינים יותר. בנוסף, אספנו מספר רב של חשבוניות קבלה על מנת להגדיל את הנתונים ולהגביר את הדיוק והמהימנות של המודלים.

התאמת סוג הנתונים: הנתונים עברו עיבוד קפדני על מנת להתאים לדרישות המודלים שנבחרו. כחלק מעיבוד, יצרנו מאפיינים חדשים ברמת הלקוח וברמת המוצר, כגון סטטיסטיקות על כמויות ומחירים (ממוצע, סטיית תקן, טווחים ומספר רכישות), במטרה לשקף טוב יותר את דפוסי הצריכה וההתנהגות של הלקוחות. בנוסף, השתמשנו בטכניקה לאיזון הנתונים (SMOTE) כדי להתמודד עם חוסר איזון בין התוויות (כגון חריגות לעומת נתונים רגילים), דבר ששיפר את היכולת של המודלים לזהות אי-התאמות. נתונים קטגוריאליים הומרו לייצוג מספרי באמצעות קידוד One-Hot, ונתונים מספריים עברו נרמול (StandardScaler) כדי לשפר את ביצועי המודלים. תהליך זה יצר מאגר נתונים מגוון, מוכן ללמידה, ואפשר למודלים לבצע חיזוי מדויק ומבוסס יותר של חריגות.

1.2 הנחות המודל:

בסעיף זה נתעד את ההנחות שנעשו במהלך בחירת טכניקות מידול, יחד עם מניפולציות הנתונים שבוצעו, על מנת לעמוד בדרישות המודל.

תהליך קבלת החלטות:

במהלך תהליך קבלת ההחלטות בבחירת טכניקת המידול, שקלנו מספר גורמים מרכזיים, ובהם אופי הנתונים, מטרות הפרויקט והתאמתן של גישות מידול שונות למשימת הסיווג. מאחר שהנתונים המקוריים כללו גם שדות קטגוריים כמו מספר לקוח, מוצר וחנות – ביצענו המרה לעמודות מספריות (באמצעות קידוד), כך שהנתונים יתאימו לשיטות מידול שונות.

מכיוון שהמטרה בפרויקט הייתה לחזות חריגות בכמויות רכישה – כלומר לזהות האם כמות מסוימת חורגת מהטווח התקין שהוגדר לקוח – התמקדנו במודלים שמסוגלים ללמוד דפוסים מורכבים ולבצע סיווג בינארי מדויק.

לצורך כך שקלנו את יכולת ההסבר של המודל, את ביצועי המודל לפי מדדים כמו Precision, Recall, F1 Score ו-Accuracy, את יכולת ההתמודדות עם מבנה הנתונים לאחר ההמרה המספרית, ואת מידת המורכבות של המודל מבחינת יישום והבנה. על בסיס שיקולים אלה, בחרנו לבחון, מכונת וקטורים תומכים (SVM), עצי החלטה ויער רנדומלי – כל אחד מהם תורם היבט שונה בהבנת וחיזוי אי־התאמות בכמויות.

תאימות המודל: היה חשוב לוודא טכניקות המידול שנבחרו תואמות למבנה הנתונים ולסוג המשימה – חיזוי אי־התאמות בכמויות רכישה על בסיס נתונים היסטוריים הכוללים מאפיינים כמו מספר לקוח, חנות, מוצר, כמות ומחיר. מאחר ו-הנתונים כוללים ערכים מספריים, נבחרו מודלים מונחים המתאימים במיוחד למשימות סיווג, תוך התאמה לאופי ומורכבות של הנתונים, כדי לאפשר זיהוי מדויק של דפוסים המצביעים על מקרים חריגים.

יעדים עסקיים: הבחירה שלנו בטכניקות המידול תואמת את היעדים הספציפיים של הפרויקט, כגון זיהוי אי-התאמות בכמויות עבור הלקוחות, חיזוי חריגות בהתנהגות רכישה, כל זאת מתוך מטרה לספק כלי יעיל לבחינת נתונים באופן מהיר ולתמוך בקבלת החלטות מושכלת.

ההנחות שהזכרנו בחלק זה מבטיחות שקיפות בתהליך המידול שלנו, ומביאות לתהליך יעיל ואיכותי יותר, שמקל על פרשנות תוצאות המודל ויישום המסקנות בצורה פרקטית ויעילה.

2. תכנון בדיקות והערכת ביצועים:

במהלך בניית המודלים, גובש תהליך בדיקה שיטתי להערכת הביצועים, אשר כלל הגדרת קריטריונים ברורים למדידת טיב המודל (Precision, Recall, F1 Score ו־Accuracy), ובחינה של שיטות חלוקה בין מערכי אימון והבדיקה (למשל, 70%-30% 60%-40% ו־50%-50%). במסגרת התהליך נבנו מספר מודלים שונים, וכל אחד מהם הוערך לפי אותם מדדים ובאותה חלוקת נתונים, במטרה לזהות את המודל שמניב את התוצאות המדויקות המאוזנות ביותר. מכיוון שמדובר בתהליך איטרטיבי, בכל שלב בוצעה השוואה בין המודלים והתוצאות שופרו בהתאם עד להשגת מודל סופי יעיל ומדויק.

חלוקה לנתונים-

לצורך אימון ובחינה של המודלים, בחנו מספר אפשרויות לחלוקת הנתונים, ביניהן חלוקות ביחסים של 60% לאימון ו־40% לבדיקה (0.6/0.4), וכן חלוקה שווה של 50% לאימון ו־50% לבדיקה (0.5/0.5). לאחר השוואת ביצועי המודלים תחת כל אחת מהחלוקות, מצאנו כי החלוקה של 70% לאימון ו־30% לבדיקה סיפקה את התוצאות האופטימליות. חלוקה זו אפשרה למודלים ללמוד מדגימה רחבה של הנתונים, תוך שמירה על סט בדיקה מייצג בחינה אובייקטיבית של יכולת הכללה על נתונים שלא נחשפו להם בתהליך האימון.

הערכת ביצועים לפי סוג המודל-

מודלים מונחים:

בפרויקט נעשה שימוש במודלים מונחים כדי לחזות חריגות בכמויות רכישה, בהתבסס על היסטוריית הרכישות של כל לקוח הכולל שדות כמו מספר לקוח, חנות, מוצר ומחיר. מודלים מונחים מתאימים במיוחד למשימה זו, שכן כל שורת נתון במערך האימון מסומנת מראש בתווית – כלומר, האם הכמות שנרכשה נמצאת מחוץ לטווח התקין של הלקוח, כלומר מחוץ לתחום שנקבע לפי ממוצע הכמויות של הלקוח בתוספת או הפחתה של שתי סטיות תקן מהממוצע. המודל לומד קשרים בין משתני הקלט לתוצאה, וכך מסוגל לזהות דפוסים המאפיינים מקרים חריגים. המודלים שנבחרו הם:

<u>מכונת וקטורים תומכים (SVM):</u> השתמשנו במודל זה כדי לחזות האם קיימת אי־התאמה בכמות הרכישה. מודל זה מתאים במיוחד למשימות סיווג בינאריות, ומצטיין בהפרדה בין קבוצות נתונים גם כאשר הגבול ביניהן אינו ליניארי, מה שמאפשר לזהות תבניות חריגות על סמך מאפיינים רבים.

<u>עצי החלטה:</u> מודל שמסייע לחזות אי התאמות, באמצעות תהליך קבלת החלטות ברור- כל סיווג מבוסס על סדרת תנאים פשוטים הקשורים למאפיינים, כך שניתן להבין בקלות את הקריטריונים שהובילו לסיווג כל דוגמה.

<u>יער רנדומלי:</u> שילוב של מספר עצי החלטה שנועד לשפר את הדיוק והיציבות בחיזוי אי התאמות, על ידי הפחתת השפעת רעשים וטעויות בנתונים.

בהערכת ביצועי המודלים, שמנו דגש על מדדים סטנדרטיים בתחום הסיווג, המאפשרים הבנה מדויקת של יכולת המודל להתמודד עם מקרים של חריגה. לכן, מדדים כמו Precision, Recall, Accuracy ו-F1-Score חשובים במיוחד בהקשר זה, משום שהם לא רק בוחנים את אחוז התחזיות הנכונות, אלא גם את רמת האיזון בין זיהוי נכון של חריגות לבין הימנעות התרעות שווא.

בהמשך לכך, תהליך הבנייה כלל אימון מודלים שונים והשוואתם זה לזה לפי המדדים שהוזכרו, תוך הרצה במספר איטרציות כדי לזהות את הדגם שנותן את התוצאה המדויקת והמאוזנת ביותר.

מודלים לא מונחים:

בפרויקט נעשה שימוש גם בגישת למידה לא מונחית, במטרה לזהות קבוצות שונות בנתונים – כאשר שאיפתנו המרכזית הייתה להבחין בין קבוצת לקוחות עם חריגות בכמות הרכישה לבין קבוצה של לקוחות עם רכישות תקינות. הדגש היה על זיהוי חריגות בכמות הרכישה של כל לקוח, מבלי להסתמך על תוויות מוגדרות מראש.

באמצעות שימוש במודל K-Means, ביקשנו לזהות חלוקה טבעית של הנתונים שתשקף את ההבדל בין הרכישות החריגות לרכישות התקינות. לצורך הערכת איכות החלוקה וההפרדה בין הקבוצות, התמקדנו במדד Silhouette Score, אשר בוחן עד כמה כל נקודת נתון דומה יותר לאשכול שלה מאשר לאשכולות אחרים. שימוש במדד זה אפשר לנו לקבל אינדיקציה כמותית לרמת האיכות של מבנה האשכולות שנוצרו, ולבחון האם קיימת הבחנה ברורה בין קבוצת החריגות לקבוצה התקינה.

3. תיאור המודל:

בשלב זה בנינו מספר מודלים שונים, במטרה לזהות את המודל המתאים ביותר למשימת הסיווג של חריגות כמות הרכישה. תחילה השתמשנו במודל K-Means לצורך אשכול הנתונים, מתוך רצון לזהות חלוקה טבעית בין רכישות תקינות לרכישות חריגות. תהליך זה כלל בחינה של טווחים שונים של מספר אשכולות (2 עד 10), תוך שימוש במדד Silhouette Score להערכת איכות ההפרדה בין הקבוצות. בהמשך, עברנו למודלים מונחים, בהם מודל מכונת וקטורים תומכת (SVM) לצורך סיווג בינארי – האם קיימת חריגה בכמות או לא. בנוסף, בחנו מודלים נוספים כגון עצי החלטה ויער רנדומלי, כדי להשוות בין ביצועי המודלים השונים. עצי ההחלטה בנויים כרשת של שאלות פשוטות על תכונות הנתונים, המובילות בסופו של דבר להחלטות לגבי התאמה או אי התאמה של הדוגמה לסיווג מסוים, מה שמאפשר הבנה פשוטה של הקריטריונים שהמודל משתמש בהם.
יער רנדומלי מורכב מאוסף של עצי החלטה קטנים, שפועלים יחד כדי לשפר את הדיוק ולהפחית הטיות אפשריות של עץ בודד. כל מודל נבדק לפי מדדי דיוק שונים כגון Precision, Recall, Accuracy ו־F1 Score, ותועדו תוצאות הניסויים.

3.1 הגדרות הפרמטרים-

3.1 הגדרות הפרמטרים

במהלך תהליך בניית מודלים בפרויקט, הגדרנו פרמטרים שונים המרכיבים כל מודל, על מנת לשפר את הביצועים, להפחית טעויות, ולהתמודד עם בעיית חוסר האיזון בנתונים. בתחילה הפעלנו את המודלים עם ערכי ברירת המחדל, אך שלב זה שימש כבסיס בלבד לבחינה ראשונית. לאחר ניתוח התוצאות, ביצענו ניסויים חוזרים בהם שינינו בכל פעם פרמטר אחד או יותר, כדי לבדוק כיצד כל שינוי משפיע על תוצאות המודל. תהליך זה, נעשה בצורה שיטתית ואיטרטיבית, איפשר לנו לדייק בהגדרות ולהתאים את המודלים טוב יותר לנתונים.

במודל מכונת וקטורים תומכת (SVM), השתמשנו בפרמטר C שערכו הוגדר כ־1.0, כדי לאזן בין דיוק הסיווג לבין שמירה על מרווח הפרדה יציב בין הקבוצות. בחרנו בגרעין מסוג 'rbf' (פונקציה רדיאלית), שמתאים למצבים בהם ההפרדה בין הקבוצות אינה לינארית. כמו כן, השתמשנו באופציה 'class_weight='balanced כדי לאפשר למודל להתמודד טוב יותר עם הבדל במספר הדוגמאות בין הקבוצות – כך שכל קבוצה תשפיע בצורה שווה על תהליך הלמידה.

מודל עץ החלטה, המבוסס על קבלת החלטות לפי תנאים, הגדרנו את העומק המרבי של העץ (max_depth) ל-5, כדי למנוע מצב של למידה מדויקת מדי, תפגע ביכולת הכללה.

במודל יער רנדומלי, המורכב ממספר רב של עצים, בחרנו להשתמש ב-100 עצים (n_estimators=100) כדי לשפר את היציבות והדיוק של התחזיות. בדומה לעץ הבודד, גם כאן הגבלנו את עומק כל עץ ל-5, כדי למנוע מורכבות מיותרת.

במודל ה־K-Means, יישמנו תהליך איטרטיבי לבחירת מספר האשכולות האופטימלי באמצעות לולאה שבודקת ערכים שונים של n_clusters בטווח שבין 2 ל־10. בכל איטרציה נבנה מחדש המודל עם מספר אשכולות שונה. מלבד פרמטר זה, שאר ההגדרות במודל נשארו כברירת מחדל. משמעות הדבר היא שהמודל השתמש באתחול חכם של מרכזי האשכולות באמצעות k means++.

ביצענו 10 הרצות שונות של האשכולות עם נקודות התחלה שונות כדי לבחור את הפתרון הטוב ביותר, והשתמשנו ב־500 איטרציות מקסימליות בכל הרצה.

3.2 תיאור המודלים-

במהלך פיתוח מודלים, ביצענו הערכה של הביצועים, זיהינו תובנות חשובות והתמודדות עם אתגרים שונים.

להלן פירוט המודלים שבנינו:

מכונת וקטורים תומכת(SVM)

מטרה: סיווג תוצאות בינאריות, לדוגמה, חיזוי האם קיימת אי־התאמה בכמות הרכישה.

חוזקות: מתמודד היטב עם מרחבים בעלי מספר משתנים רב, מתאים גם כאשר ההפרדה בין הקבוצות אינה לינארית, ויעיל במצבים בהם הגבולות בין הקבוצות אינם ברורים באופן מובהק.

שימוש: במסגרת בניית מודל מכונת וקטורים תומכת, סיווגנו את הנתונים המקוריים כך שכל שורת נתונים ייצגה מקרה של "יש אי התאמה" או "אין אי התאמה" בכמות הרכישה. הסיווג בוצע לפי חישוב סטטיסטי מוקדם – לדוגמה, אם הכמות חריגה מהטווח התקין שנקבע לפי ממוצע וסטיית תקן, סומנה כשורה חריגה.

תהליך זה הפך את הנתונים הגולמיים לנתוני סיווג בינאריים, המתאימים למודל SVM, שנועד להפריד בין שתי קבוצות באמצעות גבול חכם שממקסם את המרחק בין הקבוצות. לצורך כך, נעשה שימוש בגרעין מסוג(RBF) המתאים למצבים בהם ההפרדה בין הקבוצות אינה קווית. באופן זה, ניתן היה לאמן את המודל על בסיס מאפיינים כמו כמות, מחיר, ממוצעים, סטיות תקן והפרשים, ולקבל תחזית אם רשומה מסוימת מצביעה על אי התאמה.

עצי החלטה (Decision Trees)

מטרה: סיווג או חיזוי על בסיס סדרת תנאים פשוטים וברורים.
חוזקות: פרשנות נוחה, זיהוי בולט של גורמים קריטיים המשפיעים על התוצאה.
שימוש: עצי ההחלטה סייעו לנו לחזות האם תתרחש אי התאמה בכמות הרכישה על בסיס מאפייני הקלט שנבחרו, כמו כמות ומחיר. המודל למד לזהות דפוסים בקשר בין המאפיינים לבין משתנה המטרה (חריגה).

Random Forest

מטרה: חיזוי מדויק יותר על ידי שילוב תוצאות של מספר עצי החלטה.
חוזקות: עמיד בפני התאמת יתר, מטפל היטב בנתונים גדולים עם מאפיינים מרובים.
שימוש: השתמשנו ב-Random Forest כדי לשפר את הדיוק של חיזוי על ידי שילוב של מספר עצי החלטה. במקום להסתמך על עץ בודד שעלול להיות מושפע מרעש בנתונים, Random Forest מאפשר לבצע חיזוי יציב יותר, מהימן יותר, ומקטין את הסיכון להתאמת יתר.

אשכולות (Clustering -K-Means)

מטרה: קיבוץ תצפיות דומות יחדיו, לצורך זיהוי קבוצות טבעיות בנתונים.
חוזקות: אין צורך בהגדרה מוקדמת של תגיות (ללא פיקוח), קל להבנה ויישום.
שימוש: ביצענו אשכולות כדי לבדוק אם יש קבוצות מובחנות של רכישות שהתנהגו בצורה דומה. באמצעות זה זיהינו קבוצות של רכישות "חריגות" מול "רגילות" עוד לפני החיזוי, מה שנתן לנו תובנה נוספת על דפוסי רכישה לא שגרתיים, שאותם יכולנו לחבר אחר כך גם לחיזוי האי התאמות.

אתגרים והתמודדות-

במהלך העבודה נתקלנו במספר אתגרים משמעותיים:

מחסור במאפיינים:

בתחילת הדרך, מערך הנתונים הכיל מספר מוגבל של מאפיינים, מה שהקשה על בניית מודלים חזקים. זיהינו את הצורך להעשיר את הנתונים והוספנו מאפיינים חדשים שנגזרו מן הנתונים הקיימים, מה שישפר את איכות התחזיות.

התאמת יתר:

עקב המאפיינים המועטים בתחילה, המודלים סבלו מהתאמת יתר - כלומר, התאמה מוגזמת לנתוני האימון תוך פגיעה ביכולת הכללה לנתוני הבדיקה. כדי להתמודד עם כך, הוספנו מאפיינים נוספים ושיפרנו את איכות הנתונים באמצעות טכניקות קידוד (כגון One-Hot Encoding למשתנים קטגוריים). לאחר העשרה והקידוד, התאמת יתר פחתה והמודלים הציגו יכולת הכללה טובה יותר.

התאמות נתונים:

ביצענו נרמול למשתנים לפי הצורך, פיצלנו את הנתונים לסט אימון וסט בדיקה בצורה נכונה, כדי להבטיח שהמודלים ילמדו על חלק מהנתונים וייבחנו על חלק אחר שלא שימש לאימון.

4. הערכת המודל:

בשלב זה של העבודה התמקדנו בהערכת המודלים שבחרנו, הן מבחינת ביצועים והן מבחינת יכולת הפרשנות וההתאמה לצרכי הפרויקט.
מטרתנו הייתה לזהות מודלים שלא רק השיגו רמות דיוק גבוהות, אלא גם סיפקו תובנות מעשיות שיכולות לתרום לשיפור תהליך חיזוי אי התאמות בכמות הרכישה עבור המשתמשים.

לצורך ההערכה השתמשנו במדדים מתאימים למשימות סיווג בינארי: דיוק (Accuracy), דיוק חיובי (Precision), רגישות (Recall), ומדד האיזון (F1 Score).
מדדים אלו אפשרו לנו להעריך בצורה כמותית את יכולת החיזוי של כל מודל, את יכולתו לזהות נכון חריגות, ואת האיזון בין זיהוי חיובי נכון לבין שגיאות חיוביות.
בנוסף לביצועים הכמותיים, שקלנו גם את נוחות השימוש במודלים והיכולת לפרש את התוצאות בצורה שתשרת את המשתמשים, תוך הבנת המאפיינים המשפיעים על הסיכוי לאי התאמה, ופישוט תהליך קבלת החלטות על סמך החיזויים.

<u>חלוקה לפי פרמטרים test=0.3 train=0.7-</u>

מודל	Precision	Recall	F1 Score	Accuracy	Silhouette
מכונת וקטורים תומכת(SVM)	0.947	0.989	0.968	0.996	-
עצי החלטה	0.889	0.978	0.932	0.991	-
Radom Forest	0.607	0.956	0.742	0.960	-
k-means	-	-	-	-	0.74

בהתבסס על טבלת הביצועים שהתקבלה, ניתן להסיק שמודל מכונת וקטורים תומכת (SVM) הפגין את הדיוק הגבוה המאוזן ביותר. עם Precision של 0.947 ו־Recall של 0.989, המודל מצליח גם לזהות כמעט את כל החריגות וגם לשמור על דיוק מרבי בזיהוי אותן חריגות. ערכי F1 Score של 0.968 ו־Accuracy של 0.996 מחזקים את הרושם של ביצועי סיווג מעולים, המאוזנים היטב בין זיהוי לבין מניעת שגיאות.

מודל עצי החלטה הפגין ביצועים טובים, עם Precision של 0.889 ו־Recall של 0.978, מה שמעיד על יכולת כמעט מלאה לזהות חריגות. ערכי F1 Score של 0.932 ו־Accuracy של 0.991 מעידים על ביצועים חזקים.

מודל Random Forest מציג Recall גבוה של 0.956, מה שמעיד על זיהוי טוב של חריגות, אך Precision של 0.607 בלבד – נתון מצביע על זיהויים שגויים כחריגים. F1 Score של 0.742 ו־דיוק כללי (Accuracy) של 0.960 מצביעים על ביצועים סבירים, אך פחות מאוזנים לעומת שאר המודלים.

לבסוף, מודל K-Means, שנבחן לצורכי אשכול (Clustering) ולא סיווג מפורש, הציג מדד Silhouette של 0.74 – נתון טוב שמצביע על חלוקה ברורה יחסית של הקבוצות. אף שאינו מיועד לחיזוי ישיר של חריגות, מודל זה תורם להבנת מבנה הנתונים ויכול לשמש ככלי עזר ראשוני לגילוי תבניות חריגות.

<u>חלוקה לפי פרמטרים test=0.4 train=0.6-</u>

מודל	Precision	Recall	F1 Score	Accuracy	Silhouette
מכונת וקטורים תומכת(SVM)	0.907	0.975	0.940	0.993	-
עצי החלטה	0.859	0.967	0.910	0.989	-
Radom Forest	0.546	0.908	0.682	0.950	-
k-means	-	-	-	-	0.75

בהתבסס על הנתונים שבטבלה, ניתן לראות כי מודל מכונת וקטורים תומכת (SVM) הציג את הביצועים הגבוהים ביותר, עם ערכים מרשימים של Precision (0.907), F1 Score (0.940), Recall (0.975) ו-Accuracy (0.993), המעידים על איזון מצוין בין זיהוי נכון של חריגות לבין שמירה על רמת דיוק גבוהה בסיווג הכללי. גם מודל עצי החלטה, הציג תוצאות טובות עם Recall גבוה (0.967) ו-F1 של 0.910, אך דיוק (Precision) מעט נמוך יותר (0.859) ביחס ל-SVM. לעומת זאת, מודל ה-Random Forest השיג Recall גבוה יחסית (0.908), אך דיוק נמוך (0.546), דבר המעיד על נטייה לסמן יותר מדי חריגות, גם כאשר הן אינן קיימות – מה שמוביל לאיזון פחות טוב בין רגישות לדיוק. מודל K-means, ששימש לניתוח אשכולות, הציג מדד Silhouette של 0.75 – ערך טוב יחסית, המעיד על מבנה אשכולות ברור

<u>חלוקה לפי פרמטרים test=0.5 train=0.5-</u>

מודל	Precision	Recall	F1 Score	Accuracy	Silhouette
מכונת וקטורים תומכת(SVM)	0.902	0.980	0.939	0.992	-
עצי החלטה	0.879	0.967	0.921	0.990	-
Radom Forest	0.579	0.920	0.711	0.955	-
k-means	-	-	-	-	0.75

בהתאם לנתונים שבטבלה, מודל מכונת וקטורים תומכת (SVM) הציג את התוצאות המרשימות ביותר עם איזון מצוין בין כל המדדים: Precision של 0.902, Recall של 0.980, F1 Score של 0.939 ו-Accuracy של 0.992. נתונים אלו מעידים על יכולת גבוהה לזהות חריגות בדיוק רב, תוך שמירה על שיעור נמוך של טעויות. מודל עצי ההחלטה הציג אף הוא ביצועים טובים, עם Recall של 0.967 ו-Accuracy של 0.990. לעומתם, Random Forest הציג Recall גבוה (0.920), אך Precision נמוך (0.579), מה שמעיד על נטייה לזהות מקרים רבים חריגים – גם כשאינם כאלה, דבר הפוגע בדיוק הכללי של המודל. בנוסף, מודל K-means שנעשה בו שימוש לצורכי אשכולות, הציג מדד Silhouette של 0.75 – ערך המעיד על חלוקה טובה וברורה של הקבוצות.

סיכום-

מהערכת ביצועי המודלים על סמך מדדי Precision, Recall, F1 Score ו-Accuracy, אפשר להסיק מספר תובנות חשובות:
מודל מכונת וקטורים תומכת (SVM) הראה את הביצועים החזקים ביותר מבין המודלים שנבדקו. המודל השיג ערכי Precision ו-Recall גבוהים – 90.2% ו-98.0% בהתאמה. בנוסף, ניתוח ערכי ה-Precision, F1 ו-Accuracy מצביע על כך שהמודל מצליח לשמור על איזון טוב בין זיהוי מדויק של חריגים לבין הימנעות סיווג שגוי של מקרים תקינים. מדדים מעידים שהמודל מזהה בצורה מהימנה את רוב החריגים, מבלי להפעיל יותר מדי התראות שגויות, וכך שומר על ביצועים יציבים ומאוזנים.

עץ החלטה הציג ביצועים טובים, בעיקר מבחינת Recall של 96.7%, כלומר הצליח לזהות כמעט את כל המקרים שבהם הייתה אי התאמה. עם זאת, ערך ה-Precision היה נמוך יותר (87.9%), מה מרמז על כך שחלק מהזיהויים היו חיוביים שגויים - כלומר המודל נוטה במידה מסוימת להתריע על אי התאמה גם כשאין באמת בעיה. למרות זאת, האיזון הכולל בין המדדים נותר טוב יחסית והמודל הציג דיוק (Accuracy) גבוה (99.0%).

Random Forest, לעומת זאת, הפגין אומנם Recall גבוה (92.0%), מה שמעיד שגם הוא מזהה היטב את מרבית מקרי האי התאמה, אך Precision נמוך יחסית (57.9%). המשמעות היא שהרבה פעמים המודל סימן שיש בעיה גם כשבפועל לא הייתה, מה שפוגע ביכולת לסמוך עליו בקבלת החלטות. כתוצאה מכך גם ערך ה-F1

היה נמוך יותר(71.1%), והדיוק הכללי (95.5%) היה סביר בהשוואה לשאר המודלים.

סיכום, מודל מכונת הווקטורים התומכת הוכיח את עצמו כמודל החזק והמאוזן ביותר מבחינת כל מדדי הביצועים, ולכן הוא המתאים ביותר למטרת זיהוי מקרי אי התאמה בכמות הרכישה במערכת שלנו. מודל עץ החלטה מהווה אלטרנטיבה סבירה עם יתרון בפרשנות, אך ביצועיו היו מעט פחות מדויקים, ואילו Random Forest לא התאים לצורך שלנו עקב ריבוי התראות שגויות.

בנוסף, השתמשנו גם באלגוריתם K-Means כדי לנתח את הנתונים ולזהות קבוצות טבעיות (אשכולות) בקרב הרכישות.
כדי להעריך את איכות האשכולות שהתקבלו, חישבנו את Silhouette Score, שהערך שלו נע בין 1- ל-1:
ערכים קרובים ל-1 מעידים על כך שהדגימות משויכות היטב לאשכול שלהן ומובחנות היטב מאשכולות אחרים.
ערכים קרובים ל-0 מצביעים על כך שהאשכולות חופפים או לא מובחנים.
ערכים שליליים מעידים על שיוך לקוי של דגימות לאשכולות.

במקרה שלנו, התקבל ציון Silhouette של 0.75, מה שמעיד על כך שהנתונים התחלקו בצורה יחסית ברורה בין האשכולות.
המשמעות היא שכל דגימה הייתה קרובה יחסית למרכז האשכול שלה ורחוקה מהאשכולות האחרים - מה שמעיד על חלוקה טבעית ומשמעותית בנתונים.
תוצאה זו מחזקת את העובדה שהתבנית הקיימת בנתונים מאפשרת לזהות קבוצות מובהקות של רכישות, דבר שיכול לסייע בניתוח דפוסי רכישה חריגים.

לסיכום, מודל ה-K-Means הצליח להפריד את הנתונים לאשכולות איכותיים בצורה אפקטיבית, כפי שמעיד מדד ה-Silhouette, והוסיף לנו הבנה עמוקה יותר על מבנה הנתונים מעבר לחיזוי הבינארי של אי התאמות.