**פרויקט מסכם נושאים מתקדמים בלמידת מכונה:  
  
שם המרצה:** ד"ר חן חג'ג

**מגישות:**

ירדן אוחנה-315321851

עמית ששון - 322759648

[**קישור לגיטהאב**](https://github.com/AmitSasson123/Project_machine_learning)

[**קישור לדאטה**](https://data.gov.il/dataset/demand_survey_2019-br7/resource/d245414c-5231-4e86-a9d9-d2cf724b4be0)

**מבוא-**

**שאלת מחקר:**

הפרויקט שלנו מתמקד בניתוח סקר ביקוש לעסקים שונים בבאר שבע באמצעות נתונים דמוגרפיים והעדפות צרכנים.   
המטרה העיקרית היא לזהות מאפיינים אישיים המשפיעים על הבחירה בקטגוריית מוצרים מסוימת ולבנות מודל חיזוי מדויק לעמודת היעד category\_name.

**תיאור כללי של הדאטה**

הדאטה מכיל נתונים על 5,012 משתמשים וכולל 34 עמודות. הוא משלב מידע דמוגרפי (כגון גיל, מגדר, סטטוס משפחתי, מספר ילדים, ושכונה) עם העדפות צרכנים (קטגוריות מוצרים כמו מסעדות, תחביבים ושירותים).

**עיבוד מקדים (Preprocessing):**

1. **הסרת עמודות:**

* **City:** העמודה הוסרה כי כל הערכים בה זהים (באר שבע) ולכן אינה תורמת למודל.
* **:Unnamed, open notes** אחוז הערכים החסרים בעמודות אלו היה גבוה מאוד, ולכן הוסרו.
* **Choices** : למרות שהייתה לה השפעה מובהקת על עמודת היעד, העמודה הוסרה כדי למנוע מצב של**overfitting** ולאפשר למודל ללמוד קשרים כלליים יותר.

1. **תהליך מילוי ערכים חסרים:**

תמונה שמכילה צילום מסך, ריבוע, טקסט, מלבן

התיאור נוצר באופן אוטומטי

ה-Heatmap מראה קשרים משמעותיים בין עמודות דמוגרפיות מסוימות,

למשל: age ו-children (0.58): השתמשנו בקשרים האלו לקביעת גילאים בקבוצות דומות.

family\_status ו-children (0.67): הקשר החזק ביניהם מסביר למה מילוי ערכים חסרים התבסס על סטטוס משפחתי ומספר ילדים.

age ו-family\_status (0.47): שילוב גיל וסטטוס משפחתי יצר חלוקה הגיונית לקבוצות שמהן ניתן להסיק ערכים חסרים.

מסקנה: מילוי הערכים בעמודות כמו age, family\_status, children, ו-vehicle התבסס על קשרים חזקים ומשמעותיים בין המשתנים, כפי שמוצג במטריצת ה-PPS.

תמונה שמכילה קו, עלילה, תרשים, צילום מסך

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, קו, גופן

התיאור נוצר באופן אוטומטי

מסקנה נוספת מה- Heatmap העמודה category\_name (עמודת היעד) אינה מושפעת משמעותית מעמודות אחרות:  
ה-Heatmap מראה שערך ה-PPS בין category\_name לכל שאר העמודות הוא נמוך מאוד (0 או קרוב לכך).  
המשמעות הוא אין משתנה אחד שמכתיב באופן מובהק את ערכי היעד.   
אין משתנה "דומיננטי" שעלול ליצור בעיות כמו overfitting. המודל יוכל לנצל שילוב רחב של משתנים כדי להסביר את המידע.   
המסקנה היא ש- category\_name אינה מושפעת באופן מובהק ממשתנה יחיד, וזה מחזק את הגישה שלך להשתמש בכל המשתנים בתור פיצ'רים ליצירת מודל כללי וחזק.

1. **זיהוי וטיפול באנומליות:**

* **זיהוי אנומליות:**
  + זיהינו 75 ערכים חריגים בעזרת Isolation Forest ו־87 ערכים חריגים לפי שיטת IQR.
  + ערכים נדירים בעמודות כמו neighborhood ו־ family\_status זוהו כקטגוריות   
    בתדירות נמוכה מ־0.01.
* **טיפול באנומליות:**

טיפלנו בקטגוריות נדירות בתדירות נמוכה מ־0.01

* + בעמודת **neighborhood** ערכים נדירים שולבו עם שכונות סמוכות או אוחדו עם אזורי תעשייה קרובים
  + בעמודת **family\_status** קטגוריות נדירות אוחדו עם קטגוריות דומות מבחינת משמעות.
  + תמונה שמכילה טקסט, צילום מסך, גופן, מספר

    התיאור נוצר באופן אוטומטיבעמודת **age**  הערכים נורמלו כדי לשמר מידע מדויק.

\*הערה- אחרי תיקון אנומליות מבחינה הגיונית החלטנו להשאיר רק קטגוריה אחת-   
א. תעשייה כי הקטגוריה לא קרובה לשכונות אחרות שנמצאות בדאטה והיא לא מייצגת שכונות מגורים כמו שאר הקטגוריות

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, תרשים, צילום מסך, עלילה

התיאור נוצר באופן אוטומטי

אחרי טיפול בקטגוריות נדירות ואחרי מילוי ערכים חסרים

אחרי טיפול בקטגוריות נדירות ואחרי מילוי ערכים חסרים

אחרי הנרמול:

לפני הנרמול:

תמונה שמכילה טקסט, קו, עלילה, תרשים

התיאור נוצר באופן אוטומטיתמונה שמכילה תרשים, קו, עלילה, טקסט

התיאור נוצר באופן אוטומטיתמונה שמכילה תרשים, עלילה, קו

התיאור נוצר באופן אוטומטי

Classification

הגרפים הראו שעמודת age לא מתפלגת בצורה מושלמת, אבל יש לה צורה שקרובה לפעמון. בגלל זה, מילאנו את הערכים החסרים לפי ממוצע או חציון בתוך קבוצות כמו מגדר ומצב משפחתי.   
כך שמרנו על קשר הגיוני בין הנתונים בלי לתת משקל גדול מדי לערכים חריגים.

**מהלך העבודה**

* עיבוד נתונים: הנתונים נורמלו באמצעות MinMaxScaler, למעט עמודת age והפיצול לנתוני אימון (80%) ובדיקה (20%) בוצע כדי להבטיח הערכה אמינה.
* התמודדות עם :Multi-Label נעשה שימוש ב-MultiLabelBinarizer להמרת התוויות לפורמט בינארי, וב MultiOutputClassifier להתאמת המודלים להתמודד עם משימות רב-תווית. מחלקה זו מאפשרת לכל מודל לטפל בתוויות מרובות בצורה יעילה, בכך שהיא מתאימה לכל תווית תת-מודל נפרד ובכך ממקסמת את ביצועי הסיווג.
* שימוש ב -Param Grid בוצע Grid Search על מנת למצוא את הפרמטרים המיטביים לכל מודל. תהליך זה כלל בדיקת ערכים שונים של פרמטרים מרכזיים, כמו מספר האיטרציות, עומק העצים וקצב הלמידה, כדי לשפר את ביצועי המודלים.
* מדדים שנבדקו:
  + **CV F1 (Micro)** מדד המשקף את האיזון בין Precision ל-Recall במבחן Cross Validation. הוא מאפשר להעריך את ביצועי המודל בצורה עקבית על פני מספר קיפולים, מה שמפחית השפעה של חלוקה מקרית של הנתונים.
  + **Accuracy** מודד האם המודל ניבא בצורה נכונה את *כל* התוויות עבור כל דגימה. זהו מדד מחמיר שמציג אחוז הצלחה נמוך יותר, אך נותן תמונה שלמה על הצלחת המודל בדגימה מלאה.
  + **Precision (Micro)** חשוב לצמצום False Positives
  + **Recall (Micro)** משמעותי לאיתור מלא של דגימות חיוביות.
  + **F1 Score (Micro)**איזון בין Precision וRecall
  + **F1 Score (Samples)**מדד המתייחס לרמת התחזיות המדויקות לכל דגימה בנפרד.
  + **Hamming Loss** מדד ייחודי למשימות Multi-Label המשקף שיעור טעויות פר תווית.

**מה ציפינו ומה קרה בפועל**

הציפייה הייתה שמודלים מורכבים, כמו **CatBoost** ו-**LightGBM**, יתבלטו בביצועיהם בשל התאמתם לנתונים קטגוריאליים וחסרים, בעוד שמודלים פשוטים יותר, כמו **Decision Tree**, יציגו ביצועים נמוכים יותר עקב מגבלותיהם בטיפול במבנים מורכבים. בפועל, **CatBoost** הצטיין בכל המדדים המרכזיים. גם  **LightGBM** הציג ביצועים גבוהים במדדים רבים, כולל דיוק מה שמצביע על התאמתו למשימה. מנגד, מודלים פשוטים כמו  **Decision Tree** הראו ביצועים מספקים אך פחות מדויקים, כפי שצפינו. פיצ'רים מרכזיים כמו neighborhood, selection, ו-children נמצאו חשובים בעקביות עבור כל המודלים, דבר המצביע על תרומתם הגדולה למשימה.

השוואת ביצועים

| **model** | **CV F1 (Micro)** | **Test Accuracy** | **Precision (Micro)** | **Recall (Micro)** | **F1 Score (Micro)** | **F1 Score (Samples)** | **Hamming Loss** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **XGBoost** | 0.5296 | 0.7411 | 0.6300 | 0.4701 | 0.5384 | 0.5241 | 0.2589 |
| **Decision Tree** | 0.5197 | 0.7379 | 0.6150 | 0.4922 | 0.5468 | 0.5312 | 0.2621 |
| **Random Forest** | 0.5250 | 0.7336 | 0.6156 | 0.4545 | 0.5230 | 0.5102 | 0.2664 |
| **CatBoost** | 0.5356 | 0.7496 | 0.6505 | 0.4767 | 0.5502 | 0.5346 | 0.2504 |
| **LightGBM** | 0.5311 | 0.7422 | 0.6313 | 0.4745 | 0.5418 | 0.5269 | 0.2578 |

ניתוח תוצאות ומסקנות

כל המודלים הציגו ביצועים קרובים, אך CatBoost בלט בזכות איזון מיטבי בין המדדים. קרבה זו מעידה על יציבות הנתונים, בעוד היתרון של CatBoost ניכר בעיקר ביכולת להתמודד עם משימות מורכבות ורב-תווית.

תמונה שמכילה צילום מסך, מלבן, ריבוע, פתקייה

התיאור נוצר באופן אוטומטיתמונה שמכילה קו, תרשים, צילום מסך, עלילה

התיאור נוצר באופן אוטומטי

Clustering

**Clustering**

**הקדמה**  
במסגרת עבודה זו בוצע ניתוח חלוקה לאשכולות (Clustering) על נתונים מעורבים (נומריים וקטגוריאליים) במטרה לזהות מבנים נסתרים ולבחון חלוקה משמעותית לפי קטגוריות מרכזיות כמו *category\_name*. נבחנו ארבעה מודלים, תוך התאמתם למאפייני הנתונים:

1. **K-Prototypes** משלב נתונים נומריים וקטגוריאליים, מתאים לנתונים הטרוגניים.

מתאים מאוד לדאטה שלנו.

1. **K-Means** פופולרי ומתאים לנתונים נומריים, שימש להשוואה לאחר עיבוד קטגוריאלי.
2. **DBSCAN** מזהה מבנים מורכבים ורעשים, ללא צורך בהגדרת מספר אשכולות מראש.
3. **Hierarchical Clustering** מספק מבנה עץ המאפשר ניתוח היררכי.

כל מודל נבחן לפי Silhouette Score ביצועים ויכולת לזהות הבדלים משמעותיים בין אשכולות.

**הכנת הנתונים**

1. **הגדרת סוגי משתנים**: חלוקה למשתנים נומריים וקטגוריאליים.
2. **עיבוד משתנים קטגוריאליים**: המרה למחרוזות לעיבוד אחיד.
3. **שילוב נתונים**: שילוב למבנה אחיד לצורך ניתוח, כולל הגדרת אינדקסים לקטגוריות עבור K-Prototypes.

**השוואת מודלים ומאפייני אשכולות**

| **יתרונות** | **חסרונות** | **מספר אשכולות** | **Silhouette** | **מודל** |
| --- | --- | --- | --- | --- |
| משלב נתונים נומריים וקטגוריאליים | ציון נמוך, חפיפה בין קבוצות | 8 | 0.19 | **K-Prototypes** |
| פשוט להבנה ויישום | מתאים לנתונים נומריים | 8 | 0.3559 | **K-Means** |
| מזהה מבנים מורכבים ורעשים | רגיש לפרמטרים | 4 | 0.5267 | **DBSCAN** |
| ניתוח היררכי | דורש משאבים חישוביים | 8 | 0.3660 | **Hierarchical** |

לאחר ההשוואה, נבחרו **K-Prototypes** ו-**K-Means** בשל תרומתם המעשית לניתוח:

* **K-Prototypes:** מתאים לנתונים הטרוגניים, מאפשר הבנה מעמיקה של קשרים בין קטגוריות ודפוסי צריכה מגוונים.
* **K-Means:** מספק חלוקה ברורה של אשכולות, במיוחד עבור נתונים נומריים, ומדגיש מגמות רחבות באוכלוסייה.

**לא נבחר DBSCAN** למרות הציון הגבוה (0.5267), מאחר שהמודל רגיש להגדרות פרמטרים, מה שעלול להקשות על הפקת תובנות יציבות.

**ניתוח אשכולות עם K-Prototypes**

**מאפיינים דמוגרפיים והתנהגותיים:**  
אשכול 3, הכולל צעירים בגיל ממוצע 29, מתאפיין ברובו בסטודנטים וגברים המתמקדים בצריכה במסעדות ובתי קפה (68%). אוכלוסייה זו מתאימה למוצרים ושירותים בתחום הבילוי והפנאי.  
אשכול 6 מאופיין בצעירים ללא ילדים (גיל ממוצע 24), עם דגש על אופנה ופנאי, ומייצג קהל יעד מתאים לקמפיינים שיווקיים לצעירים.

**קבוצת קצה:**  
אשכול 7, הכולל קשישים בגיל ממוצע 91, מתמקד בצרכים בסיסיים, עם דגש על "מצרכים ומוצרים" (80%).

**ניתוח אשכולות עם K-Means**

**חלוקה ברורה לפי גיל ודפוסי צריכה:**  
אשכול 0 (גיל ממוצע 26) מתמקד בצריכה בפעילויות פנאי ובילוי במסעדות (68%) ומהווה קהל יעד מרכזי לקמפיינים לצעירים.  
אשכול 2 (גיל ממוצע 49) מציג צרכים מגוונים, כולל דגש על "מצרכים ומוצרים" (32%), ודורש גישה שיווקית רחבה יותר.

**קבוצות קצה ויוצאות דופן:**  
אשכול 7 (גיל ממוצע 94) מתמקד בצרכים בסיסיים, עם צריכה גבוהה של "מצרכים ומוצרים" (75%). קבוצה זו מתאימה לשירותים מותאמים כמו משלוחים.

**מגמות כלליות:**  
אשכולות גדולים, כמו אשכול 0 (415 תצפיות), מדגישים את דפוסי הצריכה המרכזיים באוכלוסייה, עם דגש על פנאי ובילוי בקרב צעירים.

**סיכום התובנות**

* **K-Prototypes:** מספק תובנות מעמיקות על קשרים בין מאפיינים דמוגרפיים לדפוסי צריכה. מתאים במיוחד לניתוח אוכלוסיות מגוונות ולזיהוי קבוצות ייחודיות.
* תמונה שמכילה צילום מסך, טקסט, צבעוני

  התיאור נוצר באופן אוטומטיתמונה שמכילה צילום מסך, קו, תרשים, טקסט

  התיאור נוצר באופן אוטומטי**K-Means:** מצטיין בזיהוי מגמות רחבות עם חלוקה ברורה.

תמונה שמכילה צילום מסך, טקסט, ריבוע, מלבן

התיאור נוצר באופן אוטומטי

k-means

K-prototype

גרף K-prototype

**מבוא למערכות ההמלצה:**

מערכות המלצה משמשות להציע למשתמשים פריטים (כגון מוצרים, קטגוריות או שירותים) בהתאם להעדפותיהם. בפרויקט פותחו ונבחנו שלוש מערכות המלצה שונות:

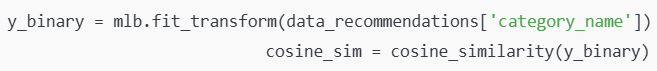
1. מבוססת תוכן (Content-Based)
2. מבוססת שיתוף פעולה (Collaborative Filtering)
3. מבוססת פופולריות (Popularity-Based)

**מערכת 1: מבוססת תוכן (Content-Based Recommendation)**

**על המערכת:**

הדגש הוא על הקשר בין התוכן (הקטגוריות), ללא קשר ישיר למשתמשים אחרים.

**תיאור המערכת:**

* המערכת ממליצה למשתמש על קטגוריות שדומות לקטגוריות שכבר בחר בעבר.
* חישוב הדמיון נעשה בין קטגוריות (תוכן) על בסיס המאפיינים שלהן.
* הדמיון מחושב באמצעות **Cosine Similarity** על הקטגוריות שנבחרו על ידי המשתמש.

תמונה שמכילה טקסט, גופן, צילום מסך

התיאור נוצר באופן אוטומטי**פלט המערכת:**

* הפלט מציג למשתמש קטגוריות חדשות הדומות לקטגוריות שכבר בחר.  
  **לדוגמה-** אם משתמש 1 בחר: [אופנה וטיפוח, תחביבים, ספורט ופנאי]   
  ומשתמש 2 בחר: [תחביבים, ספורט ופנאי, מסעדות ובתי קפה] אז מערכת   
  ההמלצה מזהה דמיון בקטגוריות שנבחרו ולכן ממליצה למשתמש 1 על  **מסעדות ובתי קפה**, אבל מזכירה שגם משתמש 2 בחר בקטגוריות דומות.

**מערכת 2: מבוססת שיתוף פעולה (Collaborative Filtering)**

**על המערכת:**

הדגש הוא על הדמיון בין המשתמשים, ולא על התוכן עצמו.

**תיאור המערכת:**

* המערכת ממליצה על קטגוריות שמשתמשים דומים בחרו, אך המשתמש עצמו עדיין לא בחר בהן.
* תמונה שמכילה טקסט, צילום מסך, גופן, מספר

  התיאור נוצר באופן אוטומטיהדמיון בין משתמשים מחושב על בסיס מאפייני הפרופיל שלהם (למשל, מגדר, שכונה, מצב משפחתי).

**פלט המערכת:**

* הפלט מציג למשתמש קטגוריות שנבחרו על ידי משתמשים עם פרופיל דומה.  
  **לדוגמה:**  
  משתמש עם פרופיל: "גבר, נשוי, יש ילדים, שכונת נווה נוי"   
  יקבל המלצה על קטגוריות שנבחרו על ידי משתמשים עם פרופיל דומה.

**מערכת 3: מערכת המלצה משולבת (Hybrid Recommendation System)**

**על המערכת:**

משלבת את היתרונות של תוכן, שיתוף פעולה ופופולריות כדי לספק המלצות מותאמות אישית.

**תיאור המערכת:**

* מערכת זו משלבת בין שלוש שיטות: שיתוף פעולה מבוסס משתמשים, תוכן, ופופולריות.
* תמונה שמכילה טקסט, צילום מסך, גופן, מספר

  התיאור נוצר באופן אוטומטיהמטרה היא להציע למשתמש קטגוריות חדשות בצורה מותאמת אישית המבוססת על משתמשים דומים, תוך שמירה על פופולריות הקטגוריות ברקע.

**פלט המערכת:**

* הפלט מציג למשתמש קטגוריות מומלצות לפי:
  1. דמיון למשתמשים אחרים (Collaborative Filtering).
  2. פופולריות של הקטגוריות הנבחרות (Popularity-Based).

**טבלה מסכמת- מערכות המלצה:**

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי

בהצעת הפרויקט ציינו שנשתמש במדדים **Precision@K** ו-**Recall@K**, שהם מדדים מקובלים להערכת מערכות המלצה. מדדים אלו בוחנים את הדמיון בין המלצות המערכת לבין המציאות בפועל, כלומר, עד כמה המערכת מצליחה לשחזר את העדפות המשתמשים הקיימות. עם זאת, במהלך העבודה הבנו שהשאיפה שלנו ממערכות ההמלצה היא לא רק לשקף את המציאות אלא להרחיב את החשיפה של המשתמשים לקטגוריות חדשות ומגוונות.

לכן, החלטנו להתמקד במדדים **Diversification** ו-**Novelty**, שמייצגים את היכולת של המערכת להציג תכנים שונים (Diversification) וחדשים (Novelty) עבור המשתמשים. מדדים אלה מדגישים את השוני בין המערכת למציאות, ומסייעים ליצור חוויית משתמש עשירה יותר, שמתאימה למטרתנו להרחיב את האופקים של המשתמשים ולאפשר להם לגלות תוכן חדש ולא מוכר.

**מה המשמעות של התוצאה:**העובדה ששני המדדים הם **1.00** מראה שהמערכת מתפקדת בדיוק כפי שרצינו: **1.** ההמלצות מגוונות לחלוטין ואינן חוזרות על קטגוריות שכבר נבחרו (Diversification = 1.00). **2.** ההמלצות מכילות רק קטגוריות חדשות שלא נבחרו על ידי המשתמש (Novelty = 1.00).  
**מטרת המערכת:** המטרה הייתה להבטיח שהמערכת תספק למשתמש אופציות חדשות ומעניינות שלא התנסה בהן בעבר,   
כדי להעשיר את חוויית המשתמש.  
**משמעות תוצאה 1.00:** זה מדגיש שהמערכת הותאמה כראוי:  
- כל ההמלצות ייחודיות ולא חוזרות על עצמן**.**  
- המערכת לא ממליצה על קטגוריות שהמשתמש כבר ראה או בחר**.**

תמונה שמכילה טקסט, גופן, צילום מסך, קו

התיאור נוצר באופן אוטומטי

**סיכום ודיון**

במהלך הפרויקט התמקדנו בניתוח ובפיתוח מודלים בתחום מערכות ההמלצה באמצעות נתוני סקר ביקוש לעסקים בבאר שבע.   
מטרתנו הייתה לזהות דפוסי צריכה מרכזיים ולהציע מודלים שמסוגלים לספק המלצות מותאמות אישית וחדשניות.

1. **עיבוד מקדים:** ביצענו עיבוד נתונים יסודי שכלל זיהוי אנומליות, מילוי ערכים חסרים, ונרמול. תהליך זה אפשר לנו לעבוד עם דאטה יציב ומאוזן ששימש כבסיס לכל המודלים שבנינו.
2. **קלסיפיקציה:** בדקנו מספר מודלים, וביניהם CatBoost ו-LightGBM שבלטו בביצועיהם. המודלים הצליחו לזהות פיצ'רים מרכזיים שהשפיעו על הבחירה בקטגוריות מוצרים.
3. **Clustering:** חלוקת הנתונים לאשכולות אפשרה לזהות מגמות מרכזיות בקבוצות אוכלוסייה שונות, תוך שימוש במודלים כמו K-Prototypes ו-K-Means שהבליטו את הדמיון והשוני בין קבוצות דמוגרפיות שונות.
4. **מערכות המלצה:** פיתחנו ובחנו שלוש מערכות המלצה – מבוססת תוכן, מבוססת שיתוף פעולה, ומשולבת. כל אחת מהמערכות נבדקה בעזרת מדדים ייחודיים כמו Diversification ו-Novelty, שהראו שהמערכת הצליחה להציע למשתמשים קטגוריות מגוונות וחדשות.

תהליך העבודה הראה את החשיבות של עיבוד מקדים מדויק וזיהוי משתנים משמעותיים. שילוב בין שיטות שונות (למשל, מערכות המלצה משולבות ואשכולות) מאפשר לנתח נתונים ממספר כיוונים ולספק המלצות מותאמות אישית בצורה אפקטיבית יותר.

שימוש במדדים כמו Diversification ו-Novelty תרם ליצירת מערכות המלצה מתקדמות שמעשירות את חוויית המשתמש.

שילוב טכניקות מתקדמות כמו MultiOutputClassifier ומודלים מתקדמים דוגמת CatBoost שיפרו את ביצועי המערכת.

**מגבלות:**

* חלק מהמודלים היו תלויים בהגדרת פרמטרים מדויקת למשל ( DBSCAN) מה שהקשה על ייצוב התוצאות.
* כל משתתף בסקר היה יכול לבחור מספר קטגוריות, מה שהוביל לריבוי שורות עם פרטים זהים, פרט לעמודת היעד. כאשר צמצמנו כל משתמש לשורה אחת, נאלצנו לשלב את הקטגוריות שבחר לרשימה בעמודת היעד. פעולה זו הקשתה על תהליך בניית המודלים, שכן נדרשנו להתמודד עם אתגרי Multi-Label, דבר שחייב לימוד והטמעה של שיטות עבודה ייחודיות בשלבים הבאים של הפרויקט והרחבה של מה שנלמד בכיתה.

**חשוב לנו לציין שאת כל העבודה עשינו ממש ביחד פענחנו, למדנו, התייעצנו, בכינו, צחקנו, החכמנו ולמדנו דברים חדשים לגמרי.  
  
המלצות לעתיד:**

1. הרחבת הדאטה לאזורים גיאוגרפיים נוספים.
2. בדיקה של מודלים נוספים לשיפור מערכות ההמלצה.
3. שילוב נתונים בזמן אמת (real-time) לשיפור ההתאמה האישית.

**חשוב-עבור כתיבת הקודים הטכניים נעזרנו בצ'אט GPT על מנת לייעל את העבודה. את מהלך החשיבה ביצענו בעצמנו וגם דיוקים בתוך הקוד עצמו והצ'אט עזר לנו בניסוח הקוד לפי הדרישות שלנו. כמובן שאין קוד שלא הבנו לעומק מה עומד מאחוריו ואיך הוא מתבצע ואת השגיאות תיקנו בעצמנו בצורה ידנית.**