שיר אלטמן 325168870

פרויקט גמר – למידת מכונה

קישור לפרויקט בגיט
קישור למאגר הנתונים

במהלך הפרויקט נבחן את מאגר הנתונים "Speed Dating Experiment", המאפשר להתבונן על דפוסי קבלת החלטות רומנטיות דרך עדשה של למידת מכונה. ננתח את מבנה הנתונים, נבין מי המשתתפים, ונבחן אילו מגבלות יכולות להשפיע על איכות המודלים שנבנה בהמשך.

מבנה המאגר וסוגי הנתונים

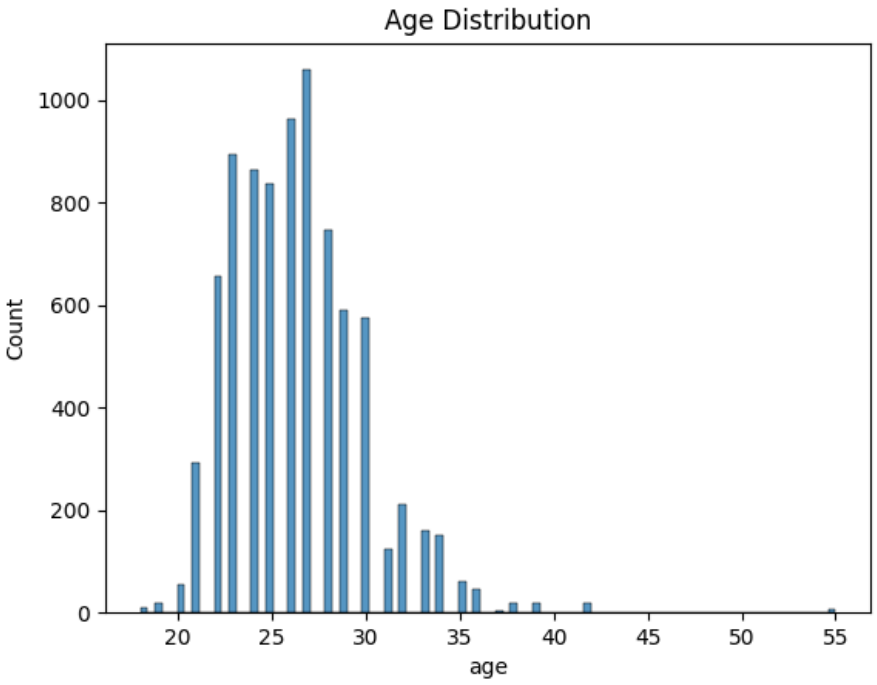
מאגר הנתונים כולל 8,378 רשומות, כאשר כל אחת מהן מתארת מפגש בין שני משתתפים באירוע הספיד-דייטינג. הנתונים נאספו לאורך 21 תאריכים שונים של ניסויים, ומכילים בסך הכול 195 תכונות, שאותן ניתן לחלק לארבע קטגוריות עיקריות:

- **פרופיל דמוגרפי**: מידע בסיסי על המשתתפים כמו גיל, מגדר, רקע אתני ותחום לימוד.
- **העדפות ותפיסה עצמית**: תשובות לשאלונים שמולאו לפני המפגש, שבהם המשתתפים ציינו מה הם מחפשים בפרטנר פוטנציאלי, וכן כיצד הם רואים את עצמם.
- **תחומי עניין ופעילויות**: דירוגי עניין של המשתתפים במגוון פעילויות פנאי כמו ספורט, סרטים, קריאה ועוד.
- **נתונים לאחר האינטראקציה** : הליבה של המאגר. אחרי כל דייט שהתמשך 4 דקות, כל משתתף התבקש לדרג את הפרטנר לפי שש תכונות שונות, ולציין אם היה מעוניין להיפגש איתו שוב.

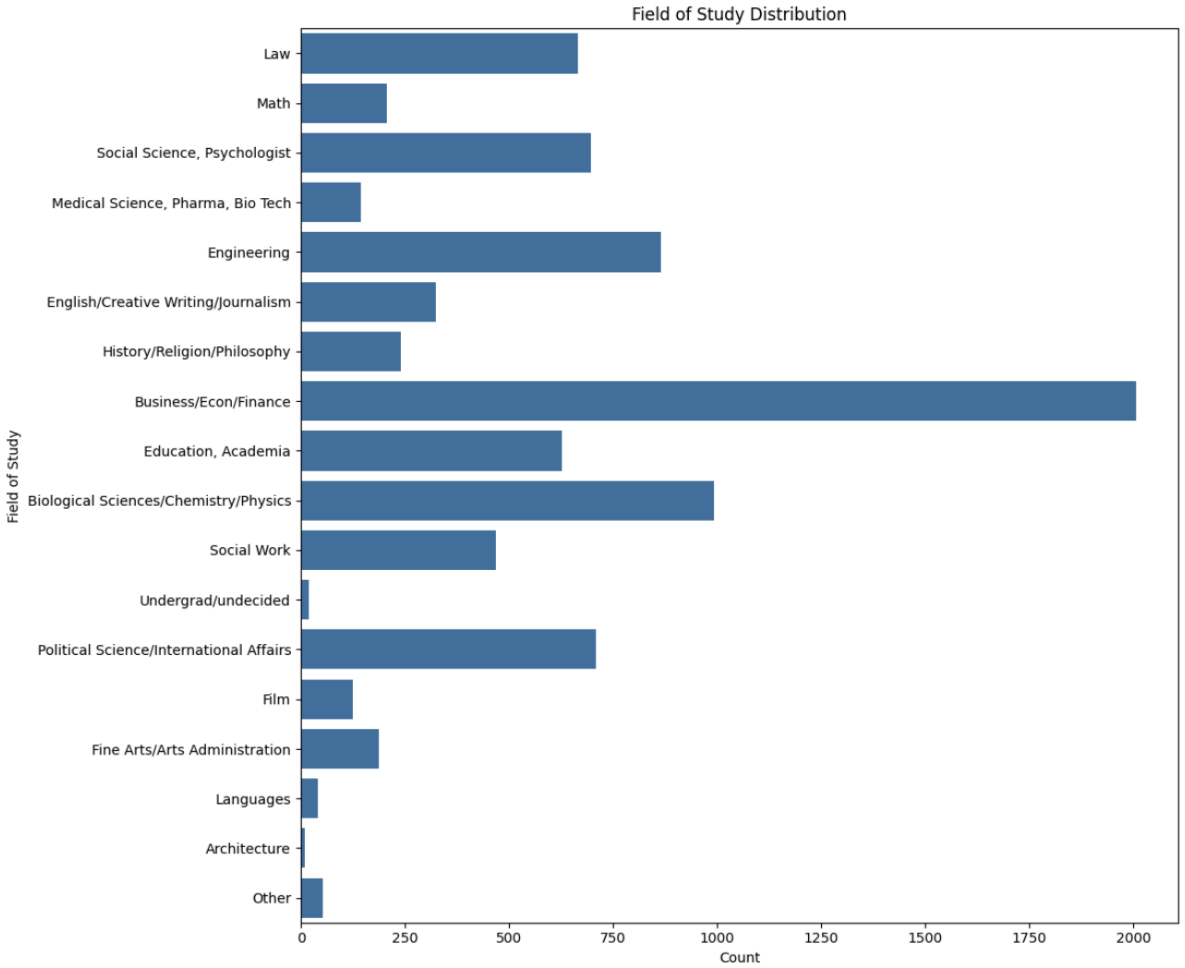
ניתוח פרופיל המשתתפים

מגדר- המאגר מאוזן כמעט לחלוטין מבחינת חלוקה בין גברים לנשים, מה שמפחית הטיות פוטנציאליות הקשורות לייצוג לא שוויוני.

דמוגרפיה- רוב המשתתפים הם סטודנטים לתארים מתקדמים באוניברסיטת קולומביה, בגיל ממוצע של כ-26, עם טווח גילאים יחסית צר. המשמעות ברורה: המודלים שנבנה יהיו מכוילים לאוכלוסייה צעירה, משכילה ועירונית, והיכולת להכליל את הממצאים לקהלים אחרים מוגבלת.



תחומי לימוד- יש הטיה לתחומים מסוימים: מקצועות כמו כלכלה, עסקים ומדעים מדויקים מיוצגים בתדירות גבוה, בעוד שמקצועות כמו אמנויות, שפות או חינוך מופיעים בתדירות נמוכה יותר.



העדפות מוצהרות מול התנהגות בפועל

פערים מגדריים: ניתוח משתני ההעדפות (כגון attr1_1) מצביע על הבדלים צפויים: גברים ייחסו חשיבות גבוהה יותר למראה חיצוני, נשים נטו להעדיף אינטליגנציה ושאפתנות.

פער בין הצהרות להתנהגות: אחד הממצאים הבולטים במאגר הוא הסתירה בין מה שאנשים מצהירים עליו לבין מה שמנבא בפועל את ההחלטות שלהם. לדוגמה, דירוג האטרקטיביות של הפרטנר (attr_o) נמצא כגורם מנבא חזק מאוד אצל שני המינים, גם בקרב נשים שטענו כי אינטליגנציה חשובה להן יותר.

אתגרים ומגבלות מהותיות של המאגר

ערכים חסרים- זו כנראה המגבלה המשמעותית ביותר של המאגר. עמודות רבות, ובעיקר דירוגים שניתנו על ידי הפרטנר (כמו attr_o, sinc_o), כוללות שיעור גבוה של ערכים חסרים, ההנחה שלי שזה נבע מסיבות תפעוליות לדוגמא שלא כל המשתתפים מילאו את כל השדות בכל סבב.יש חשיבות לנתון הזה משום שאם נשתמש רק בשורות מלאות, נאבד כ-95% ממקרי ה"מאץ'". לכן נדרשים פתרונות כמו **Imputation**, אך עלול להוביל להטיה של המודל.בכך כל ניתוח שנבצע נצטרך לקחת בחשבון את המגבלה הזו .

חוסר איזון בין המחלקות - רק כ-16% מהמפגשים הסתיימו ב"מאץ'", מה שיוצר הטיה ברורה כלפי רוב הדגימות.ההשפעה שאני מצפה לראות היא שמודלים מתאמנים על הנתונים ייטו לנבא "אין מאץ" ובכך יתקשו לזהות מקרים נדירים וזה בעיקר מה שחשוב לי :) .ולכן כדי להתמודד עם המכשול הזה השתמשתי בSMOTE.

הטיית מדגם - המשתתפים לא מייצגים את כל האוכלסייה כפי שניתן לראות לפי הגרפים למעלה מדובר בעיקר בסטודנטים רווקים,צעירים ומשכילים מניו יורק. ולכן התובנות שנפיק בעיקר יתאימו לקבל דומכה ולכן לא בהכרח ניתן לבצע הכללה לשאר האוכלסייה

ניסויים ראשוניים על סט נתונים מצומצם והצורך בשינוי אסטרטגי

בשלב הראשון של הפרויקט בחרתי לאמץ גישת עיבוד מקדים נוקשה, וזאת כדי לבנות סט נתונים "מושלם" מבחינת שלמות המידע. לשם כך ביצעתי שתי פעולות מרכזיות: הסרתי את כל העמודות שבהן למעלה מ־60% מהערכים היו חסרים. לאחר מכן הסרתי כל שורה שבה הופיע אפילו ערך חסר אחד (dropna()).
כתוצאה מהפעולות הללו הייתה צמצום דרמתי בהיקף הנתונים ,מתוך 8378 שורות במאגר המקורי נשארתי עם 329 שורות בלבד.
הסט שנשאר לאחר שלב הניקוי האגרסיבי כלל פחות מ־4% מהמידע המקורי, ואיבד כ־95% מהדוגמאות החיוביות (מאצ'ים). למרות זאת, בחרתי להפעיל את המודלים כדי לבחון את יכולת החיזוי הבסיסית.

תוצאות המודלים על הסט המצומצם

למרות המגבלות הברורות, הורצו מספר מודלים על סט הנתונים המצומצם הזה כדי לבחון את יכולת החיזוי הראשונית.

- רגרסיה לוגיסטית: המודל הציג ביצועים נמוכים מאוד, עם F1-Score של 0.27 עבור ניבוי מאצ'ים. ה-Recall היה 0.23 בלבד, מה שמצביע על כך שהמודל מפספס את רוב המאצ'ים האמיתיים.

	precision	recall	f1-score	support
0	0.82	0.89	0.85	53
1	0.33	0.23	0.27	13
accuracy			0.76	66

- Random Forest: מודל זה הציג שיפור קל, עם F1-Score של 0.33. הוא היה מדויק יותר (Precision של 0.60), אך עדיין סבל מ-Recall נמוך מאוד של 0.23.

	precision	recall	f1-score	support
0	0.84	0.96	0.89	53
1	0.60	0.23	0.33	13
accuracy			0.82	66

התוצאות הנמוכות והעקביות בכל המודלים, בשילוב ההבנה שאיבדתי את רוב הדוגמאות החיוניות, הובילו למסקנה שהגישה האגרסיבית לעיבוד הנתונים הייתה שגויה. המודלים לא כשלו בגלל מגבלות אלגוריתמיות, אלא בגלל מחסור במידע. סט הנתונים שנותר היה קטן ולא מייצג, ולכן לא אפשר למידה אפקטיבית של הדפוסים.
כתוצאה מכך, הבנתי שנדרש שינוי מהותי בגישת ה preprocessing את אותו השינוי נציג בפרק הבא.

המפנה האסטרטגי

המעבר לשיטת ההשלמה נובע מההכרה שמחיקת ערכים חסרים פוגעת בצורה קריטית ביכולת המידול. במקום להמשיך למחוק שורות ועמודות, בחרתי לעבור לאסטרטגיית **Imputation** , שמטרתה לסגור את הפערים בנתונים מבלי לוותר עליהם.

האסטרטגיה התבססה על שני עקרונות מרכזיים:

- **שמירה על מידע**: רציתי לשמר כמה שיותר מהנתונים המקוריים, ובמיוחד את הדוגמאות החיוביות הנדירות (מאצ'ים), שבלעדיהן לא ניתן ללמוד דפוסים משמעותיים.
- **יציבות המודל**: ידעתי שמודל שנאמן על אלפי דוגמאות יהיה יציב, כולל ומדויק הרבה יותר ממודל שמבוסס על מאות שורות בלבד.

כדי להבטיח שההשלמה תהיה מדויקת ככל האפשר, יישמתי אסטרטגיה שמבחינה בין סוגי עמודות:

- **עמודות מספריות** (כמו גיל, דירוגים): נעשה שימוש ב Median. מדד זה נבחר על פני הממוצע בשל עמידותו ל Outliers, ובכך הוא מספק אומדן מרכזי יציב יותר.
- **עמודות קטגוריאליות** (כמו תחום לימוד): הושלמו לפי הערך השכיח ביותר.

הקוד שבניתי טען את סט הנתונים לאחר שלב הניקוי הראשוני (שבו הוסרו עמודות ריקות לחלוטין), ועבר בלולאה על כל שדה. עבור כל עמודה זיהיתי אם היא קטגוריאלית או מספרית, והחלתי את שיטת ה Imputation .

לאחר אימות סופי שלא נותרו ערכים חסרים, יצרתי קובץ אקסל חדש שכלל את כל 8,378 השורות המקוריות, כשהן יוצרות שורה מלאה. סט נתונים זה שימש אותי כבסיס לכל שלבי המידול, הניתוחים וההשוואות שנראה בהמשך.

	precision	recall	f1-score	support
0	0.84	0.87	0.85	53
1	0.36	0.31	0.33	13
accuracy			0.76	66

שיר אלטמן 325168870

רגרסיה לוגיסטית

הקוד:

המטרה המרכזית של הקוד היא לנבא את התוצאה של מפגש ספיד דייטינג – האם יתרחש "מאץ'" (התאמה) או לא – בהתבסס על מאפיינים ודירוגים שנאספו מהמשתתפים.

תהליך המימוש:

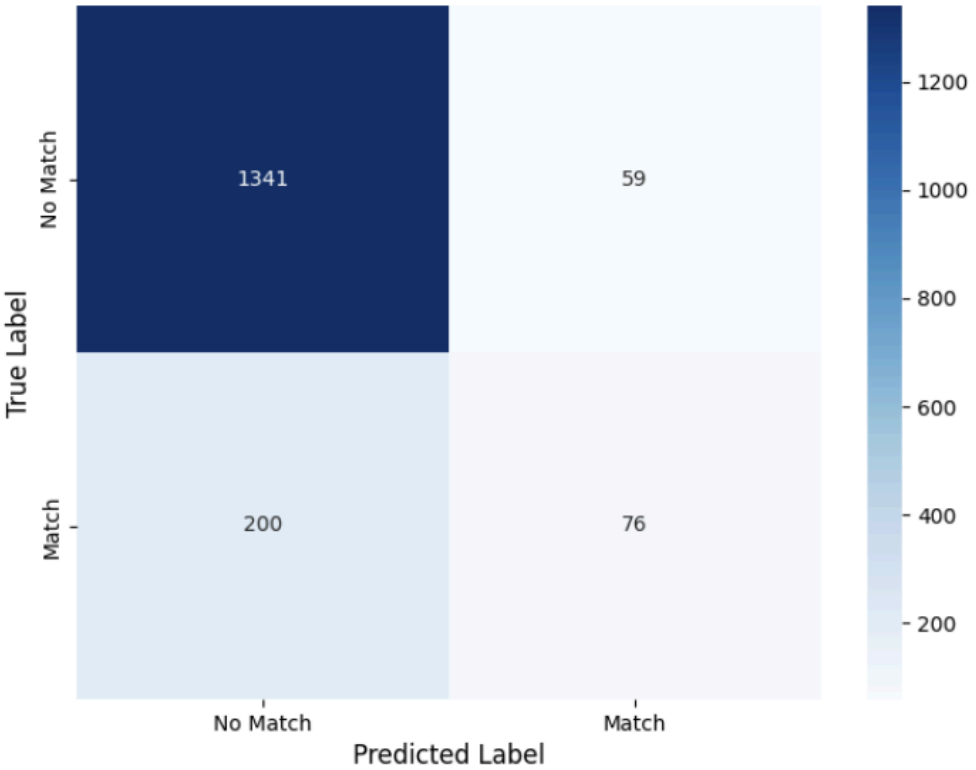
1. **תכונות כמאפיינים (Features):** המודל משתמש במגוון רחב של תכונות כדי לבצע חיזוי. בין התכונות ניתן למצוא:
   - **נתונים דמוגרפיים:** גיל, מגדר, תחום לימוד.
   - **העדפות אישיות:** חשיבות של גזע ודת בבחירת בן/בת זוג.
   - **מטרות והרגלים:** מה המטרה בדייט, תדירות יציאה לדייטים.
   - **דירוגים עצמיים:** כיצד המשתתף מדרג את עצמו בתכונות כמו אטרקטיביות, אינטליגנציה, כיף וכו'.
   - **דירוג הפרטנר:** כיצד המשתתף דירג את הפרטנר במפגש על אותן תכונות, ובמיוחד, מידת החיבה הכללית (like).
2. **המודל הלוגיסטי:** בחרתי במודל הלוגיסטי מכיוון שהוא מתאים במיוחד לבעיות סיווג בינארי, כלומר כאשר יש רק שתי תוצאות אפשריות (מאץ' או לא). המודל מנסה למצוא קשר בין התכונות לבין ההסתברות שהמשתתף יסמן כן – כלומר, שירצה מאץ'.
   - **1**: כן, היה מאץ'.
   - **0**: לא, לא היה מאץ'.
3. **הדרישה – חיזוי התאמות:** המטרה היא למזער את כמות התחזיות השגויות מסוג False Negatives – כלומר, מקרים שבהם המודל לא צפה מאץ' למרות שהוא כן התרחש בפועל.ולכן נרצה לזהות כמה שיותר התאמות אמיתיות ,לזהות כמה שיותר מאץ' כשאכן באמת היה מאץ'
4. **הבעיה והפתרון: חוסר איזון ו-SMOTE:** בסט הנתונים המקורי, כמות המפגשים שהסתיימו ב"אין מאץ'" גדולה משמעותית מכמות המפגשים שהסתיימו ב"מאץ'". ולכן אנו צריכים לטפל בחוסר איזון , הדבר גורם למודל להיות מוטה לנבא בעיקר את התוצאה השכיחה. כדי לפתור בעיה זו, השתמשתי בטכניקת SMOTE על נתוני האימון. SMOTE מייצרת דגימות סינתטיות חדשות עבור קטגוריית המיעוט , ובכך מאזנת את סט האימון ומאלצת את המודל ללמוד את הדפוסים של שתי הקטגוריות באופן שווה.

תוצאות:

	precision	recall	f1-score	support
0	0.87	0.96	0.91	1400
1	0.56	0.28	0.37	276
accuracy			0.85	1676

- **Accuracy :** נשמע שעל פניו 0.85 , נראה כמו ציון מצוין.
- **הבעיה:** הציון הגבוה מטעה. כאשר מסתכלים על Recall עבור 'מאץ' (1), רואים ציון של 0.28 בלבד. משמעות הדבר היא שהמודל מפספס 72% מההתאמות האמיתיות והוא למעשה לא שימושי למטרה שלנו.

Confusion Matrix:



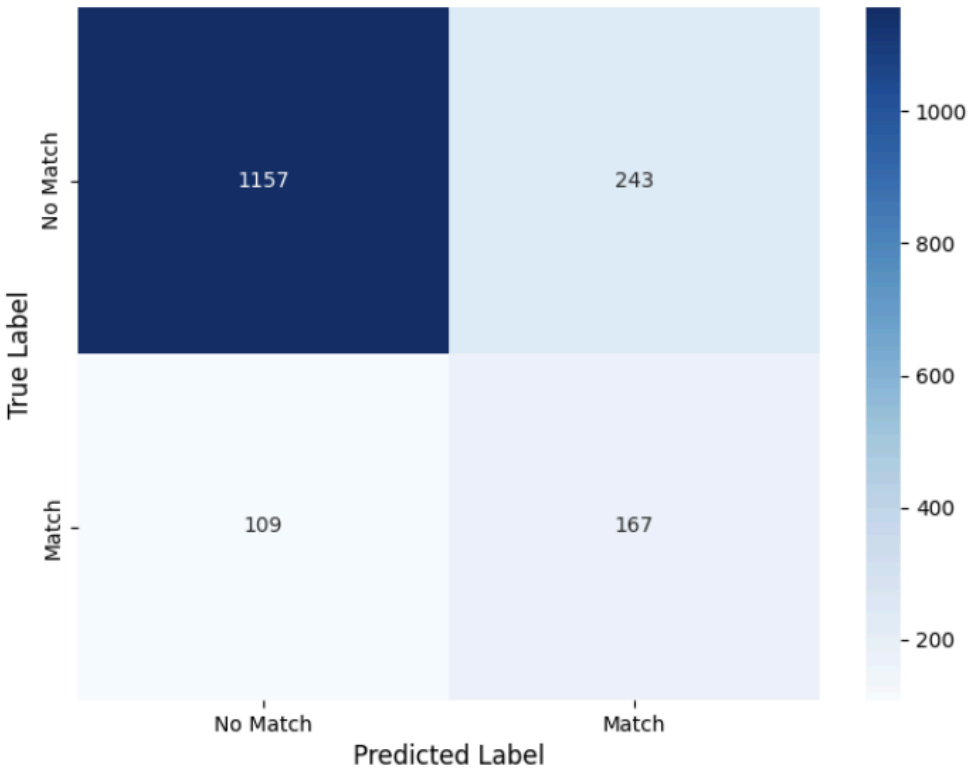
הגרף מאשר את הבעיה: מתוך 276 התאמות אמיתיות, המודל זיהה נכון רק 76 ופספס 200 (False Negatives).
המודל פשוט "למד" שהכי בטוח לנבא "אין מאץ'" וזו כמובן לא המטרה שלנו :) .

ניתוחי המודל המאוזן (לאחר SMOTE):

	precision	recall	f1-score	support
0	0.91	0.83	0.87	1400
1	0.41	0.61	0.49	276
accuracy			0.79	1676

- **Accuracy :**. הדיוק הכללי ירד מעט והוא על 0.79, אך המודל כעת הרבה יותר שימושי.
- **Recall עבור מאץ' (1):** 0.61 זהו שיפור משמעותי. המודל מצליח כעת לזהות 61% מההתאמות האמיתיות.
- **Precision עבור מאץ' (1):** 0.41 זה ה trade-off עבור היכולת לזהות יותר התאמות.

Confusion Matrix:

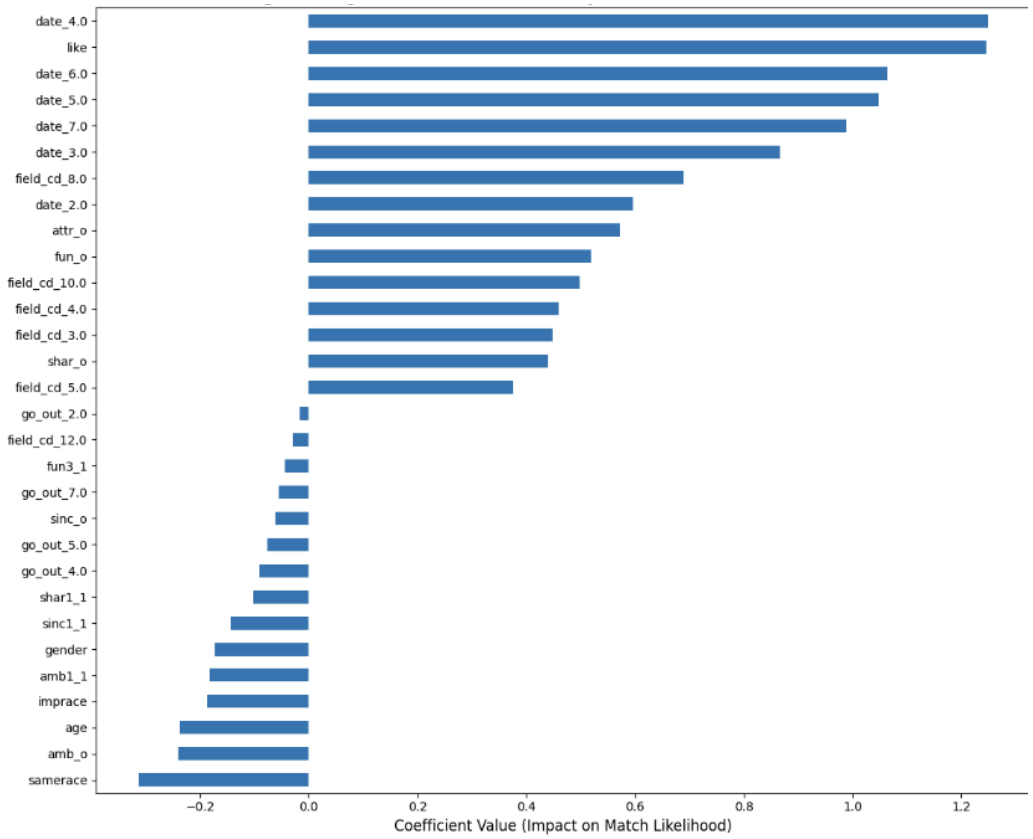


הגרף מראה בבירור את השיפור: מספר המאצים שהמודל זיהה נכון (True Positives) עלה מ-76 ל-167, ומספר הפספוסים (False Negatives) ירד מ-200 ל-109.

שיר אלטמן 325168870

מענה על שאלות המחקר באמצעות מודל הרגרסיה הלוגיסטית

ניתוח המקדמים (Feature Importance) מאפשר לזהות אילו משתנים תורמים באופן המשמעותי ביותר לחיזוי התאמה בין משתתפים. באמצעות מיפוי הגורמים הבולטים, ניתן להתחיל לגבש תמונה ברורה יותר של מה באמת משפיע על יצירת מאץ' – האם אלו העדפות מוצהרות, תכונות דמוגרפיות, או דווקא התרשמות מהפרטנר.



שאלת מחקר 1: מהם הגורמים המרכזיים המנבאים התאמה זוגית ראשונית?

ה Feature Importance מצביע על מספר גורמים מרכזיים. המנבא החזק ביותר הוא like שזה מידת החיבה הכללית שהביע המשתתף כלפי הפרטנר. מיד אחריו, דירוגים גבוהים שניתנו לפרטנר על fun_o כיפי ו-attr_o אטרקטיביות התבררו כמגדילים משמעותית את הסיכוי למאץ'. אלו הגורמים הדומיננטיים ביותר שהמודל זיהה.

שאלת מחקר 2: כיצד משפיעים מאפיינים דמוגרפיים (גיל, רקע אתני) על הסיכוי להתאמה?

המודל מראה כי למאפיינים דמוגרפיים ישנה השפעה, אם כי מתונה יותר מהגורמים שהוזכרו לעיל. age (גיל) ו-samerace (רקע אתני זהה) נמצאו כבעלי מקדמים שליליים. משמעות הדבר היא שככל שהגיל עולה, או כאשר שני המשתתפים מגיעים מאותו רקע אתני, הסיכוי הסטטיסטי למאץ' במערך הנתונים יורד מעט.

שאלת מחקר 3: אילו תכונות אישיות ותחומי עניין נמצאו כבעלי השפעה משמעותית?

התפיסה של תכונות אישיות על ידי הפרטנר, הן קריטיות. כפי שצוין, fun_o ו-attr_o (כיפיות ואטרקטיביות) הן תכונות אישיות שנמצאו כבעלות השפעה חיובית וברורה על הרצון להיפגש שוב .

שיר אלטמן 325168870

Random Forest

הקוד:

בחרתי במודל זה משום שהוא ידוע בהתמודדות טובה עם מגוון רחב של מאפיינים ובשל עמידותו לרעש ו overfitting. בניגוד לרגרסיה הלוגיסטית שמתבססת על קשרים ליניאריים, Random Forest מאפשר לזהות יחסים מורכבים ולא ליניאריים בין המאפיינים לבין הסיכוי להתאמה בין משתתפים. המודל הותאם למשימת סיווג בינארית: חיזוי האם כל מפגש ספיד דייטינג יסתיים בהתאמה או לא – בהתבסס על העדפות, דירוגים ותכונות דמוגרפיות של המשתתפים.

תהליך המימוש:

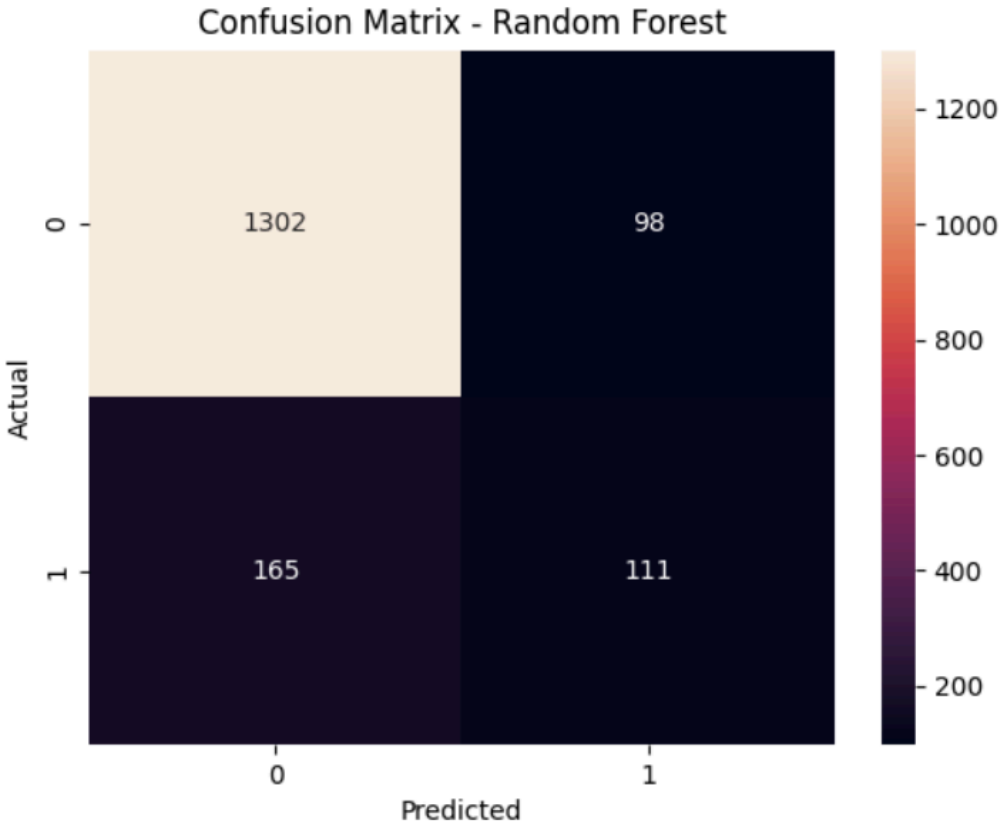
1. **תכונות כמאפיינים (Features):** השתמשנו באותו סט של מאפיינים כמו במודל הרגרסיה הלוגיסטית, הכולל נתונים דמוגרפיים, העדפות אישיות, ודירוגים הדדיים בין המשתתפים.
2. **המודל:** Random Forest הוא מודל סיווג המבוסס על שיטת Ensemble שזו גישה המשלבת מספר מודלים פשוטים ליצירת חיזוי מדויק יותר. המודל מורכב ממספר רב של עצי החלטה, כאשר כל עץ מאומן על תת-מדגם שונה של הנתונים ושל המאפיינים. בעת חיזוי, כל עץ נותן תוצאה, והיער כולו מחזיר את ההחלטה שנבחרה על ידי רוב העצים. שילוב זה הופך את המודל לעמיד יותר לרעש, יציב יותר, ופחות רגיש ל Overfitting. בנוסף, אין צורך בנרמול של המשתנים – יתרון משמעותי בעבודה עם מאפיינים מגוונים.
3. **הדרישה:** המטרה נותרה זהה- לזהות כמה שיותר התאמות אמיתיות תוך שמירה על רמת דיוק סבירה.
4. **שימוש ב-SMOTE:** כדי להבטיח השוואה הוגנת למודל הרגרסיה הלוגיסטית וכדי למנוע מהמודל להיות מוטה כלפי קטגוריית הרוב, הפעלנו גם כאן את טכניקת **SMOTE** על נתוני האימון.

תוצאות:

	precision	recall	f1-score	support
0	0.89	0.93	0.91	1400
1	0.53	0.40	0.46	276
accuracy			0.84	1676

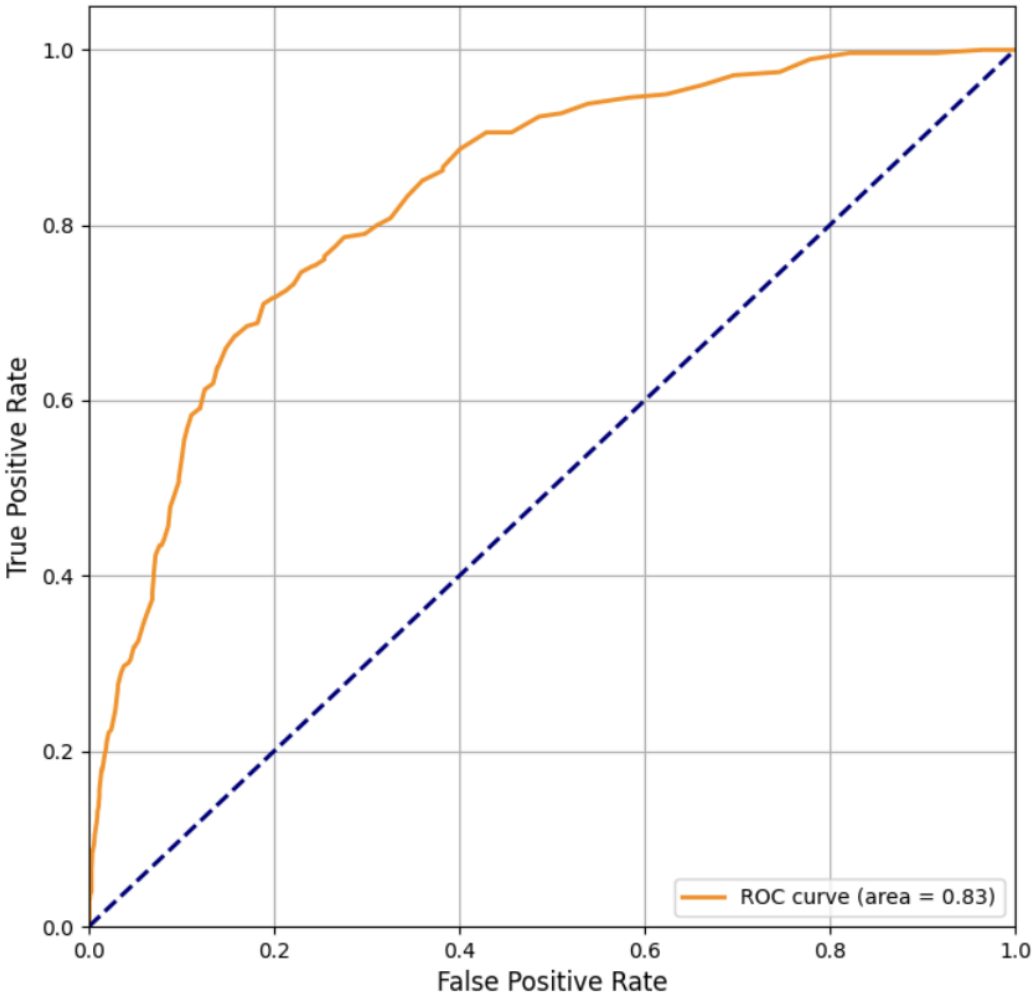
- **Accuracy :** 0.84. דיוק כללי גבוה, בדומה למודל הלוגיסטי.
- **Recall עבור מאץ' (1):** 0.40. המודל זיהה 40% מההתאמות האמיתיות. זהו שיפור לעומת מודל ללא איזון, אך נמוך יותר מה-Recall שהושג במודל הרגרסיה הלוגיסטית (0.61).
- **Precision עבור מאץ' (1):** 0.53 כאשר המודל חוזה "מאץ'", הוא צודק ב-53% מהמקרים. זהו ציון גבוה יותר מה-Precision של המודל הלוגיסטי (0.41).
- **F1-Score עבור מאץ' (1):** 0.46. ציון זה, המהווה ממוצע הרמוני של Precision ו-Recall, משקף את הפשרה שהמודל מבצע. הוא נמוך יותר מה-F1-score של המודל הלוגיסטי (0.49), מה שמצביע על כך שהמודל הלוגיסטי מצא איזון מעט טוב יותר בין שני סוגי הטעויות.

- **Confusion Matrix**



ניתן לראות כי המודל זיהה נכון 111 התאמות (True Positives) ו-1,302 מקרים של חוסר התאמה (True Negatives). לצד ההצלחות, המודל טעה ב-98 מקרים שבהם חזה מאץ' בטעות (טעות מסוג FP, או אזעקת שווא), ופספס 165 התאמות אמיתיות (טעות מסוג FN, או פספוס). הנתונים הללו מראים שהמודל מציג פשרה בין Recall ל-Precision בהשוואה למודל הלוגיסטי. הוא נוקט בגישה זהירה יותר בחיזוי מאצ'ים, ולכן מייצר פחות אזעקות שווא אך במקביל גם יותר פספוסים.

- ניתוח עקומת ROC וציון AUC:

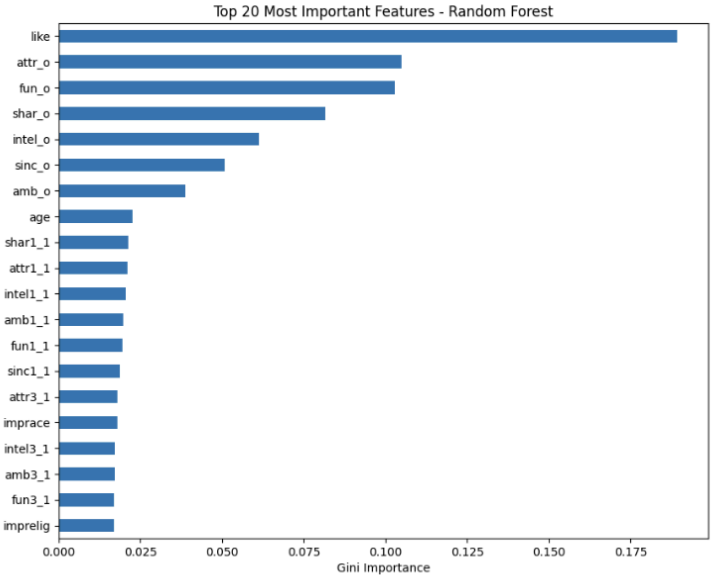


עקומת ה-ROC מספקת מבט ויזואלי על יכולת ההבחנה של המודל. העקומה מציגה את הקשר בין מספר הזיהויים ה-TP בין מספר הזיהויים ה-FP. ככל שהעקומה קרובה יותר לפינה השמאלית העליונה, כך ביצועי המודל טובים יותר. ציון ה-AUC, המייצג את השטח שמתחת לעקומה, עומד על 0.83. ציון זה נחשב טוב מאוד משום שהוא קרוב ל-1 המייצג מודל מושלם ומצביע על כך שלמודל Random Forest יש יכולת הבחנה גבוהה בין מקרים של מאץ' ואין מאץ'. בכך הוא מחזק את המסקנה כי המודל למד דפוסים משמעותיים מהנתונים.

שיר אלטמן 325168870

מענה על שאלות המחקר באמצעות Random Forest

בRandom Forest, חשיבות המאפיינים נמדדת בדרך כלל באמצעות ממד Gini – כמה כל מאפיין תורם להחלטות בעצי היער.



שאלת מחקר 1: מהם הגורמים המרכזיים המנבאים התאמה זוגית ראשונית?

בדומה למודל הרגרסיה הלוגיסטית, גם כאן התוצאות חד-משמעיות. הגורם המרכזי והחשוב ביותר בפער ניכר הוא מידת החיבה. אחריו, דירוגים גבוהים של אטרקטיביות, כיף, ותחומי עניין משותפים הם המנבאים החזקים ביותר. העקביות בין שני המודלים מחזקת מאוד את מהימנות הממצא.

שאלת מחקר 2: כיצד משפיעים מאפיינים דמוגרפיים (גיל, רקע אתני) על הסיכוי להתאמה?

המודל מראה גם הוא שלמאפיינים דמוגרפיים יש השפעה. גיל מופיע כמאפיין חשוב יחסית, אם כי משמעותית פחות מהדירוגים ההדדיים. גם חשיבות הגזע מופיע ברשימה, מה שמצביע על כך שלרקע אתני יש תפקיד בתהליך קבלת ההחלטות.

שאלת מחקר 3: אילו תכונות אישיות ותחומי עניין נמצאו כבעלי השפעה משמעותית?

המודל מדגיש את החשיבות העצומה של תכונות אישיות שניתן היה לחוש בעת הפגישה . כל ששת הדירוגים שהמשתתפים נתנו לבני זוגם, כגון אטרקטיביות, כנות, אינטליגנציה, כיף, שאפתנות ותחומי עניין משותפים, מופיעים ברשימת 20 המאפיינים החשובים ביותר. הדבר מצביע על כך שהרושם הראשוני המבוסס על תכונות אלו הוא קריטי לחיזוי התאמה.

שיר אלטמן 325168870

Clustering

הקוד:

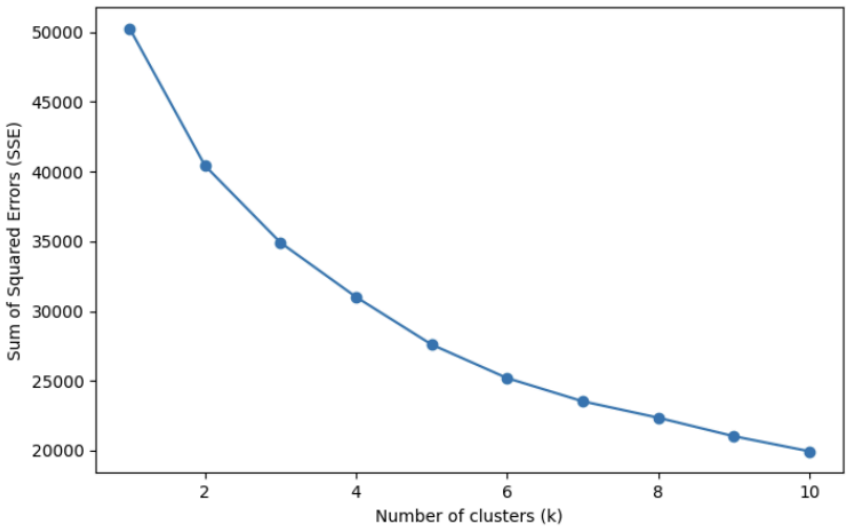
בשלב זה, השתמשנו במודל K-Means, שהוא אלגוריתם למידה Unsupervised Learning. מטרתו היא לא לנבא תוצאה ספציפית כמו האם קיים מאץ' או שאין מאץ', אני רוצה לזהות ולקבץ משתתפים לקבוצות (Clusters) בעלות מאפיינים דומים. בניתוח התמקדתי בהעדפות של המשתתפים שהם הצהירו עליהם כדי להבין אם קיימים טיפוסים שונים של משתתפים בספיד דייטינג.

תהליך המימוש:

1. **בחירת מאפיינים לאשכול:** התמקדתי במאפיינים המייצגים את מה שמשתתפים מצהירים שהם מחפשים בבן/בת זוג בתחילת האירוע. מאפיינים אלו כוללים את החשיבות שהם מייחסים לאטרקטיביות, כנות, אינטליגנציה, כייפיות, שאפתנות ותחומי עניין משותפים.
2. **נרמול הנתונים:** מכיוון ש-K-Means הוא אלגוריתם מבוסס-מרחק, הייתי חייבת לנרמל את המאפיינים. בלי נרמול, מאפיין עם טווח ערכים גדול (למשל, 1-100) ישפיע על המודל הרבה יותר ממאפיין עם טווח קטן (למשל, 1-10), גם אם הוא לא יותר חשוב.
3. **מציאת מספר האשכולות האופטימלי :** כדי להחליט לכמה קבוצות לחלק את המשתתפים, השתמשנו ה-Elbow Method. בשיטה זו נבחנת סכום ריבועי השגיאות (SSE) כפונקציה של מספר האשכולות. אנו מחפשים את הנקודה על הגרף שבה הוספת אשכול נוסף כבר לא מביאה לירידה משמעותית בשגיאה וזה יגדיר את המספר האופטימלי של אשכולות.

תוצאות:

איתור מספר האשכולות (k):

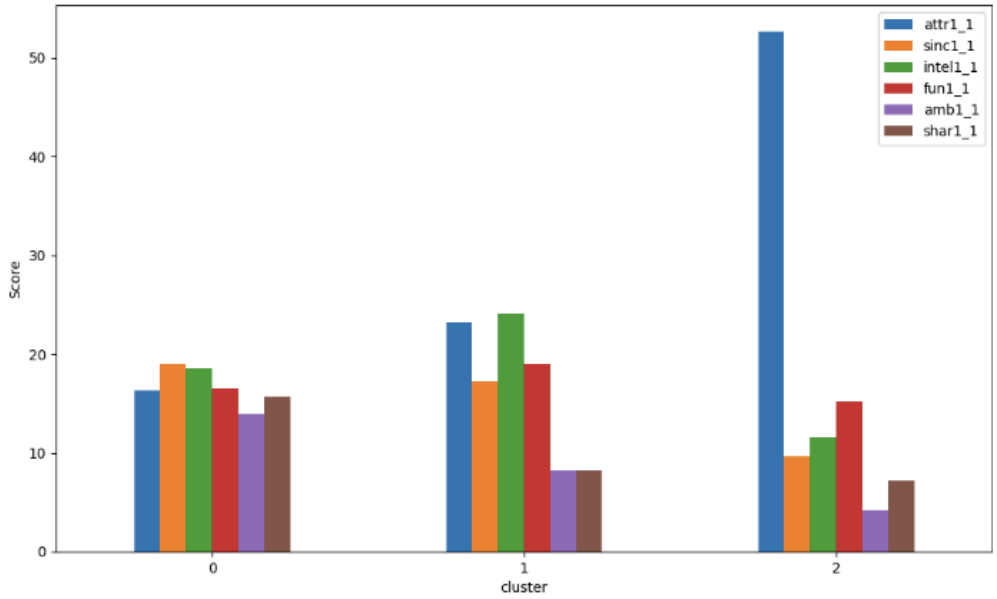


הגרף מראה ירידה חדה בשגיאה עד לנקודה של 3-4 אשכולות, ולאחר מכן הירידה מתמתנת. ולכן נבחר **k=3** כמספר האשכולות האופטימלי, מכיוון שהוא מייצג איזון טוב בין פשטות המודל לבין יכולת ההסבר שלו.

פרופיל האשכולות:

לאחר הרצת האלגוריתם עם k=3, ניתחתי את המאפיינים הממוצעים של כל אשכול כדי להבין מה מייחד כל קבוצה.

cluster	attr1_1	sinc1_1	intel1_1	fun1_1	amb1_1	shar1_1
0	16.340559	18.981237	18.618184	16.558889	13.909649	15.737446
1	23.166835	17.241131	24.101641	19.039661	8.261948	8.200135
2	52.694626	9.608126	11.634338	15.153342	4.182176	7.233814



כשצללתי לנתונים של כל קבוצה, גיליתי שיש שלושה טיפוסים יחסית ברורים של משתתפים, אז ניסיתי לתת לכל קבוצה שם שמתאר אותה. **הקבוצה הראשונה**, שקראנו לה **"הזרמנים"**, היא של אנשים שפשוט באו בשביל הכיף והחוויה, והייתה העדפה לפרטנר זורם ומהנה. **הקבוצה השנייה**, לה נקרא **" השכלתנים"**, כללה משתתפים שהדגישו שהם רוצים גם יופי וגם שכל כלומר, מישהו גם אטרקטיבי וגם אינטליגנט. **הקבוצה השלישית** הייתה הכי חד-משמעית, ולכן שמה נבחר להיות **"ממוקדי מראה"** אלו משתתפים שעבורם מה שקבע כמעט לחלוטין היה המראה החיצוני, והם ייחסו חשיבות נמוכה מאוד לשאר התכונות כמו כנות או שאפתנות.

התובנה המרכזית מניתוח האשכולות היא שהמשתתפים שונים בחשיבות שלהם לא תכונה. לכן, השלב הבא היה לבנות מודל חיזוי נפרד (מסוג יער אקראי) עבור כל אחת מהקבוצות שיצרנו, כדי לבדוק מה מנבא הצלחה עבור כל "טיפוס".

- **אשכול 0 :**

	precision	recall	f1-score	support
0	0.89	0.95	0.92	702
1	0.54	0.35	0.42	124
accuracy			0.86	826

המודל הגיע לדיוק כללי של 86%, אך מתקשה לזהות התאמות אמיתיות (0.35). כלומר, עבור קבוצה זו, קשה יותר לחזות מראש מתי קיים "קליק", אולי כי ההעדפות קשה יותר להגדיר .

- **אשכול 1 :**

	precision	recall	f1-score	support
0	0.87	0.94	0.91	574
1	0.58	0.37	0.45	123
accuracy			0.84	697

התוצאות כאן דומות, עם דיוק כללי של 84% ו-Recall נמוך של 0.37. גם כאן, נראה שהשילוב המורכב של "יופי ושכל" מקשה על המודל לנבא הצלחה באופן מדויק.

- **אשכול 2 :**

	precision	recall	f1-score	support
0	0.90	0.94	0.92	124
1	0.67	0.55	0.60	29
accuracy			0.86	153

כאן רואים את התוצאה המעניינת ביותר. הדיוק הכללי נשאר גבוה (86%), אך מדד ה-Recall קפץ ל-0.55. זוהי עלייה משמעותית, והיא מצביעה על כך שהמודל מצליח הרבה יותר לזהות התאמות אמיתיות בתוך קבוצה זו. ההסבר ההגיוני ביותר הוא שהקריטריון המרכזי שלהם הוא פשוט וקל יותר למדידה ולחיזוי.

המודלים הראשונים שהרצתי נתנו לנו תשובה כללית לשאלת המחקר: גורמים כמו חיבה, אטרקטיביות וכיף הם המנבאים המרכזיים להתאמה.

שלב הclustering חשף **למה**. הוא הראה לנו שהתשובה הכללית היא "ממוצע" שלא באמת מתאר אף אחד. בפועל, קיימות תת-קבוצות עם רפרנסים שונים. הגישה של בניית מודל נפרד לכל קבוצה הוכיחה את עצמה: גילינו שעבור קבוצת ה**ממוקדי מראה**, שמתמקדת בקריטריון פשוט, קל בהרבה לנבא הצלחה. לעומת זאת, עבור שאר הקבוצות עם העדפות מורכבות יותר, משימת החיזוי נותרה מאתגרת.

מענה על שאלות המחקר באמצעות clustering

לשאלות 1 ו 3 (מהם הגורמים המרכזיים ואילו תכונות משפיעות?): התשובה אינה אחידה, אלא תלוית בקבוצה. המודלים הראשונים הראו באופן כללי שחיבה ואטרקטיביות חשובות, אך ניתוח האשכולות מדייק את התשובה. עבור "ממוקדי מראה", הגורם המרכזי הוא מראה חיצוני. עבור "גם וגם" וה"זרמנים", התמונה מורכבת יותר, וקשה יותר לנבא הצלחה על בסיס העדפות מוצהרות בלבד. בכך, אנו עוברים מתשובה פשטנית של **מה הכי חשוב?** לתובנה מציאותית יותר של **מה הכי חשוב עבור מי?** .

לשאלה 2 (כיצד משפיעים מאפיינים דמוגרפיים?): גישה זו איפשרה לבחון את חשיבות המשתנים הדמוגרפיים באופן פרטני לכל קבוצה. ניתוח חשיבות המאפיינים לכל אשכול יכול לחשוף אם, למשל, לגיל יש השפעה שונה על הצלחה בקרב זרמנים לעומת ממוקדי מראה.

SVM

הקוד:

בשלב זה, השתמשנו במודל מסוג SVM, אלגוריתם סיווג שמטרתו למצוא את "קו הגבול" (או ליתר דיוק, המישור המפריד) הטוב ביותר שמפריד בין שתי הקבוצות – מאץ' ואין מאץ '. הרעיון הוא לא רק להפריד ביניהן, אלא למצוא את הגבול שיוצר את המרווח הרחב ביותר בין הנקודות הקרובות ביותר מכל קבוצה . השתמשנו בגרעין (Kernel) מסוג RBF, שמאפשר למודל למצוא גבולות הפרדה מורכבים ולא ישרים.

תהליך המימוש:

1. **הכנת הנתונים ושימוש ב-SMOTE:** חילקתי את הנתונים לסט אימון ומבחן, והשתמשתי בטכניקת SMOTE על סט האימון כדי לטפל בבעיית חוסר האיזון בין הקבוצות.
2. **נרמול הנתונים:** שלב זה **קריטי במיוחד** עבור מודל SVM. מכיוון שהמודל מבוסס על חישוב מרחקים בין נקודות כדי למצוא את קו הגבול האופטימלי, צריך להביא את כל המאפיינים לאותו סולם. ללא נרמול, מאפיינים עם ערכים גדולים ישפיעו על חישוב המרחק באופן לא פרופורציונלי ויטו את תוצאות המודל.
3. **אימון מודל ה-SVM:** המודל אומן על נתוני האימון המאוזנים והמנורמלים.

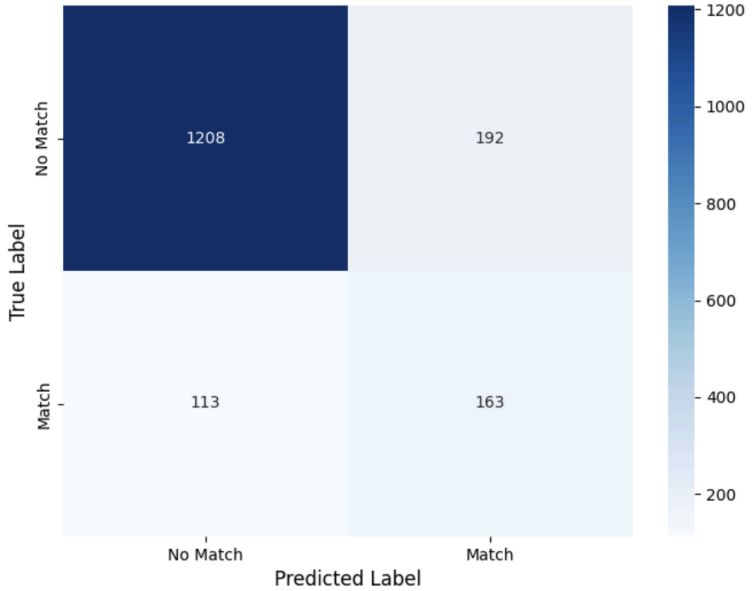
תוצאות:

	precision	recall	f1-score	support
0	0.91	0.86	0.89	1400
1	0.46	0.59	0.52	276
accuracy			0.82	1676

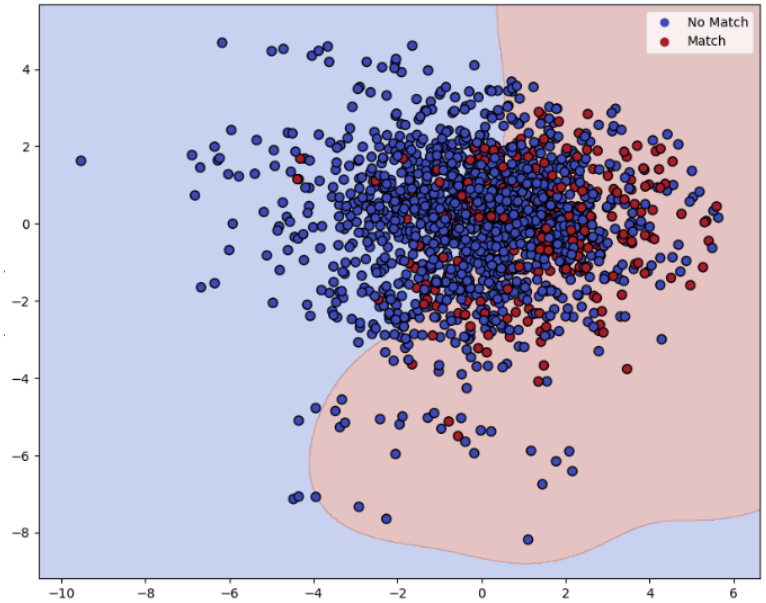
הסיווג מראה שמודל ה-SVM השיג תוצאות טובות מאוד. עם דיוק כללי של 82%, המודל מציג איזון טוב בין היכולת לזהות נכון מקרים של אין מאץ' לבין היכולת לזהות התאמות אמיתיות. מדד ה-Recall עבור קטגוריית המאץ' עומד על 0.59, מה שמצביע על כך שהמודל הצליח לזהות 59% מכלל ההתאמות האמיתיות שיפור ניכר לעומת מודל הRandom Forest. מדד ה-F1-Score, המשקלל את הדיוק וה-Recall, עומד על 0.52, ציון סביר המעיד על איזון טוב.

Confusion Matrix:

היא מראה שהמודל זיהה נכון 1157 מקרים של אין מאץ' (True Negatives) ו-163 מקרים של מאץ' (True Positives). מספר הפספוסים (False Negatives), כלומר מקרים של התאמה שהמודל לא זיהה, עומד על 113.



אחד האתגרים במודל SVM הוא שהתוצאה שלו פחות אינטואיטיבית להבנה מאשר רשימת חשיבות מאפיינים . כדי להתמודד עם זה, ניתן להמחיש באופן חזותי את גבול ההחלטה שהמודל למד. מכיוון שלא ניתן לצייר גרף עם עשרות מאפיינים, השתמשתי בטכניקת PCA כדי לצמצם את המידע הרב לשני רכיבים עיקריים (PC1 ו-PC2) שניתן להציג על גרף דו-ממדי. צמצום הממדים נועד שנוכל לראות בעין מה התרחש ולא משפיע על ביצועי המודל.



הגרף מציג את גבול ההחלטה שSVM למד. האזור הכחול מייצג את המרחב שבו המודל ינבא אין מאץ' , והאזור האדום הוא המרחב שבו ינבא מאץ' . הנקודות על הגרף הן נקודות המבחן האמיתיות. ניתן לראות שהמודל הצליח למצוא גבול מורכב שמפריד בצורה טובה בין שתי הקבוצות, אם כי קיימת חפיפה מסוימת ביניהן, מה שמסביר מדוע המודל אינו מושלם.

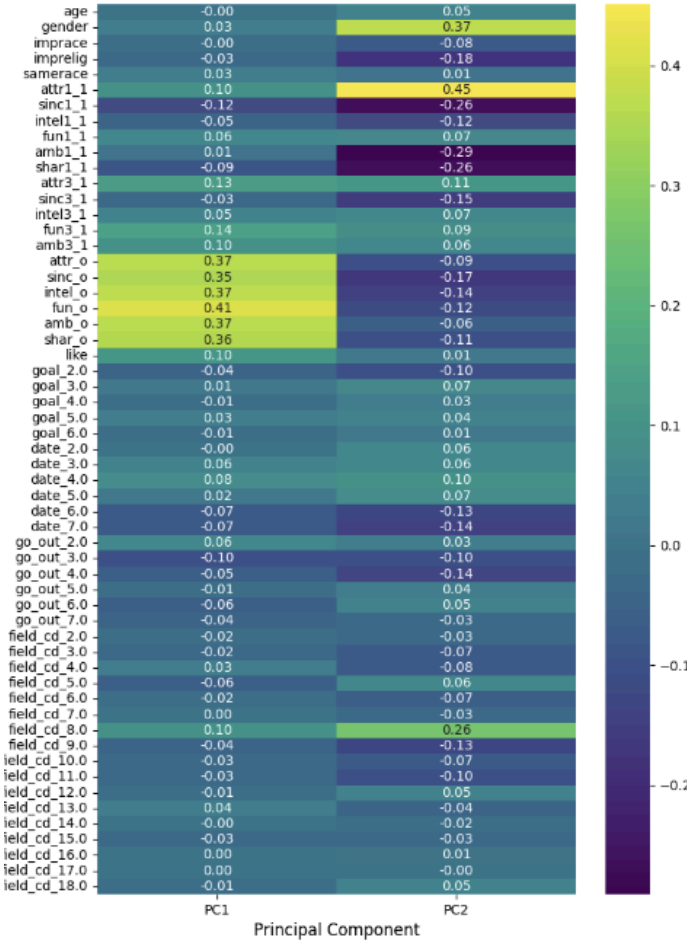
שיר אלטמן 325168870

מענה על שאלות המחקר באמצעות ניתוח PCA

מודל ה-SVM, בניגוד לרגרסיה לוגיסטית ויער אקראי, אינו מספק באופן טבעי רשימה של המאפיינים החשובים ביותר. עוצמתו היא ביכולת למצוא קשרים מורכבים בין כלל המאפיינים כדי ליצור הפרדה אופטימלית.

```
--- Top 10 Features for Principal Component 1 ---
fun_o       0.413323
intel_o     0.373703
attr_o      0.367230
amb_o       0.365912
shar_o      0.358749
sinc_o      0.351865
fun3_1      0.137961
attr3_1     0.128547
sinc1_1     0.122280
like        0.104420
Name: PC1, dtype: float64

--- Top 10 Features for Principal Component 2 ---
attr1_1          0.451317
gender           0.373886
amb1_1           0.290953
sinc1_1          0.262576
shar1_1          0.262005
field_cd_8.0     0.261029
imprelig         0.175376
sinc_o           0.168957
sinc3_1          0.154954
date_7.0         0.137375
```



- **רכיב עיקרי 1 (PC1, ציר X):** רכיב זה מורכב בעיקר מהדירוגים שהמשתתף נתן לפרטנר שלו. כלומר, ציר ה-X מייצג את "התפיסה החיובית של המשתתף כלפי הפרטנר".
- **רכיב עיקרי 2 (PC2, ציר Y):** רכיב זה מורכב בעיקר מהדירוגים העצמיים של המשתתף וממאפיינים דמוגרפיים כמו מגדר. כלומר, ציר ה-Y מייצג את "הפרופיל והתפיסה העצמית של המשתתף".

כעת, אפשר לפרש את גבול ההחלטה בצורה חכמה יותר ולענות על שאלות המחקר:

לשאלות 1 ו-3 (מהם הגורמים המרכזיים ואילו תכונות משפיעות?): מודל ה-SVM מוכיח שההחלטה על מאץ' תלויה באינטראקציה המורכבת בין התפיסה העצמית של המשתתף לבין התפיסה שלו את הפרטנר. הגרף מראה שהגבול אינו ישר כלומר, לא מספיק רק לחבב את הפרטנר. ההצלחה תלויה בשילוב בין שני הגורמים. לדוגמה, ייתכן שאנשים עם תפיסה עצמית מסוימת (ערך גבוה או נמוך בציר Y) צריכים רמה שונה של "תפיסה חיובית כלפי הפרטנר" (ערך בציר X) כדי להחליט על מאץ'. זה מאשר שהגורמים המרכזיים הם אכן תכונות נתפסות, אך גם מדגיש שהאינטראקציה ביניהן היא המפתח.

לשאלה 2 (כיצד משפיעים מאפיינים דמוגרפיים?): ניתוח ה-PCA מראה בבירור שמאפיינים דמוגרפיים כמו מגדר הם חלק מרכזי ברכיב השני (PC2). הדבר מוכיח שלמאפיינים אלו יש תרומה משמעותית למודל, והם מהווים חלק מה"פרופיל העצמי" שהמודל לוקח בחשבון כאשר הוא קובע את סיכויי ההתאמה.

למרות חוסר היכולת לקבל רשימת "חשיבות" פשוטה, ניתוח ה-PCA מאפשר לנו להסיק שהמודל למד חוקיות מורכבת המבוססת על הקשר בין תפיסת הפרטנר לתפיסה העצמית, ומאשר את חשיבותם של מאפייני ההעדפות והדמוגרפיה בתהליך.

סיכום תוצאות ודיון

בפרויקט זה התעסקתי בשאלה מה עומד בבסיס הבחירה להיפגש שוב לאחר מפגש "ספיד דייטינג", מהם הגורמים המרכזיים המנבאים התאמה רומנטית ראשונית. כדי לענות על כך, יצאתי לתהליך אנליטי מרתק שהתחלק לשני שלבים מרכזיים: בשלב הראשון, בחנתי מודלים גלובליים של למידת מכונה שאומנו על כלל הדאטה סט. לאחר שהבנתי את מגבלותיה של גישה כללית זו, פניתי בשלב השני לגישה מתקדמת ומפולחת יותר של ניתוח אשכולות (Clustering), במטרה לבנות מודל ייעודי ומדויק יותר לכל תת-קבוצה שזיהיתי.בפרק זה אסכם את התהליך, אציג את התוצאות שהתקבלו מהמודלים השונים בכל אחד מהשלבים, ואדון במשמעותן כל זאת במטרה לתת מענה מקיף לשאלות המחקר שהגדרתי בתחילת הדרך.

תשובות לשאלות המחקר

בהתבסס על כל הניתוחים שביצעתי, החל ממודלים גלובליים ועד למודלים מותאמים אישית לכל אשכול, הגעתי לתשובות הבאות עבור שאלות המחקר :

1. **מהם הגורמים המרכזיים המנבאים התאמה זוגית ראשונית?** המסקנה החד משמעית מהמחקר היא שאין גורם מנבא יחיד, והתשובה תלויה לחלוטין ב"טיפוס" המשתתף. עם זאת, ניתן להצביע על מספר גורמים מרכזיים:
   - **חיבה כללית :** באופן לא מפתיע, המנבא החזק ביותר בכל המודלים ובכל הקבוצות היה הדירוג הכללי של המשתתף את הפרטנר.
   - **גורמים תלויי אשכול:** הגורמים המשמעותיים אחרי חיבה כללית השתנו דרמטית בין האשכולות שזיהיתי: אצל **הזרמנים** הדגש היה על כיף (fun_o), אצל **ממוקדי המראה** הדגש היה על משיכה חיצונית (attr_o), ואצל **השכלתנים** ניכר חיפוש אחר איזון בין מראה, תחומי עניין ואינטליגנציה.
   - **הפער בין הצהרות למעשים:** תובנה מרכזית נוספת היא שההעדפות המוצהרות של המשתתפים (מה שהם אמרו שחשוב להם) היו מנבאים חלשים מאוד להתנהגותם בפועל.
2. **כיצד משפיעים מאפיינים דמוגרפיים על סיכויי ההתאמה?** בניתוח חשיבות התכונות במודלים השונים, מצאתי כי למאפיינים דמוגרפיים כמו רקע אתני, השכלה ומגדר הייתה השפעה נמוכה יחסית על הסיכוי ליצירת מאץ' . גם גיל הופיע לעיתים ברשימת התכונות החשובות, אך בעדיפות נמוכה. המסקנה היא שגורמים אלו מתגמדים בחשיבותם לעומת תכונות הקשורות לאינטראקציה המיידית ולרושם הראשוני (כמו משיכה, כיף וחיבה).
3. **אילו תכונות אישיות ותחומי עניין נמצאו כבעלי השפעה משמעותית?** גם כאן, התשובה תלויה בקבוצה. ברמה הכללית, תחומי עניין משותפים היו חשובים במיוחד עבור קבוצת **השכלתנים**. תכונות כמו שאפתנות וכנות דורגו נמוך יחסית על ידי כלל הקבוצות, מה שמחזק את המסקנה שבמפגש קצר וראשוני, תכונות שטחיות יותר הקשורות לחוויה ולמשיכה הן בעלות ההשפעה הגדולה ביותר על הרצון להיפגש שוב.

תוצאות המודלים הגלובליים

לאחר טיפול בחוסר האיזון במדגם באמצעות טכניקת SMOTE, הרצתי שלושה מודלים קלאסיים. המטרה שלי הייתה לקבל תמונה כללית על יכולת החיזוי של התאמה משום שזה החלק היותר מאתגר לפיצוח. להלן סיכום ביצועי המודלים:

מדד	רגרסיה לוגיסטית	יער אקראי	SVM
Accuracy	0.79	0.84	0.82
Precision	0.41	0.53	0.52
Recall	0.61	0.4	0.59
F1-Score	0.49	0.46	0.52

מהתוצאות עלה כי למרות רמת דיוק כללית טובה, המודלים התקשו יחסית בזיהוי נכון של התאמות בערכי **Recall** ו**Precision** נמוכים יחסית עבור קלאס 1. הבנתי כי גישה אחידה לכלל המשתתפים אינה מספיק רגישה כדי ללכוד את הניואנסים בהעדפות האנושיות, מה שהוביל אותי לשלב הבא.

תוצאות המודלים מבוססי clustering

בשלב זה, השתמשתי באלגוריתם K-Means כדי לזהות "טיפוסים" שונים של משתתפים על סמך האופן שבו הם דירגו את חשיבותן של תכונות שונות בבן/בת זוג. זיהיתי שלושה אשכולות ברורים, שלכל אחד מהם נתתי כינוי המשקף את מאפייניו. לאחר מכן, בניתי מודל יער אקראי לכל אשכול.

אשכול	שם הקבוצה	Accuracy	Precision	Recall	F1-Score
אשכול 0	הזרמנים	0.86	0.54	0.35	0.42
אשכול 1	השכלתנים	0.84	0.58	0.37	0.45
אשכול 2	ממוקדי מראה	0.86	0.67	0.55	0.6

המודל שנבנה עבור ממוקדי המראה הגיע לביצועים הטובים ביותר, ככל הנראה מכיוון שהעדפותיהם הברורות והממוקדות היו קלות יותר למידול.

מסקנה סופית

הפרויקט מצליח להדגים כיצד ניתן להשתמש בכלים שלמדנו כדי לפענח דפוסים מורכבים בהתנהגות אנושית. המסקנה המרכזית היא שהדרך להבנת תופעות כמו התאמה רומנטית אינה טמונה במודל אחד שמתאים לכולם, אלא ביכולת לזהות ולפלח את האוכלוסייה לתתי קבוצות בעלות העדפות שונות. גישת האשכולות לא רק שיפרה את יכולת החיזוי, אלא חשפה את האמת המורכבת יותר: מה שגורם לקליק עבור אדם אחד, עשוי להיות חסר חשיבות לחלוטין עבור אדם אחר.

"כִּי הָאָדָם יִרְאֶה לַעֵינַיִם, וַה' יִרְאֶה לַלֵּבָב" **(שמואל א' פרק ט"ז פסוק ז').**

מגבלות המחקר וכיוונים לעתיד

במהלך המחקר נוכחתי לגלות שישנן מספר מגבלות. ראשית, הוא מבוסס על מדגם הומוגני יחסית של סטודנטים מאוניברסיטה אחת, וייתכן שהמסקנות אינן תקפות לאוכלוסיות אחרות. שנית, סביבת הספיד דייטינג היא סביבה מלאכותית המעודדת קבלת החלטות מהירות.

כיווני המחקר העתידיים שיכולים לנבוע מעבודה זו הם רבים:

1. **הרחבת המדגם:** יהיה מרתק לבחון אם אותם סוגי אנשים קיימים גם בקרב אוכלוסיות מבוגרות יותר או מרקעים תרבותיים שונים.
2. **הנדסת תכונות:** ניתן להעשיר את המודלים על ידי יצירת תכונות חדשות, כמו פער גילאים או התאמה בתחומי עניין ספציפיים.
3. **בחינת מודלים נוספים:** ערכתי ניסויים ראשוניים עם מודלים מתקדמים כמו XGBoost ורשתות נוירונים, אך הם לא הניבו שיפור בתוצאות על הדאטה-סט הנוכחי. יהיה מעניין לבחון אותם שוב בעתיד על דאטה-סט גדול יותר או בשילוב הנדסת תכונות נוספת.