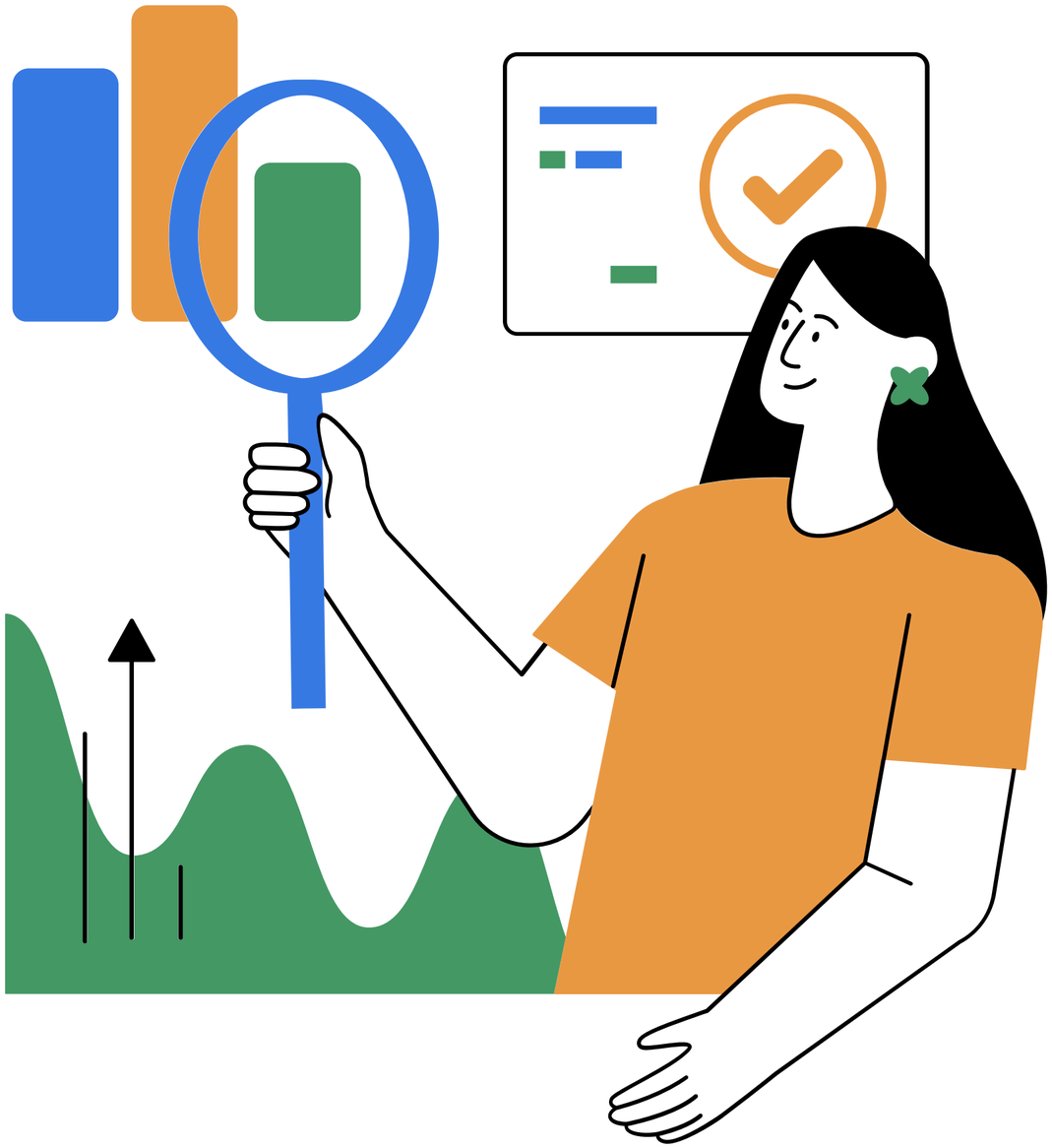
פרויקט - למידת מכונה

שם המרצה: ד"ר חן חג'ג'

שמות חברי הצוות: יהודה זגורי, אדר סבן.

תאריך: 21.01.2025



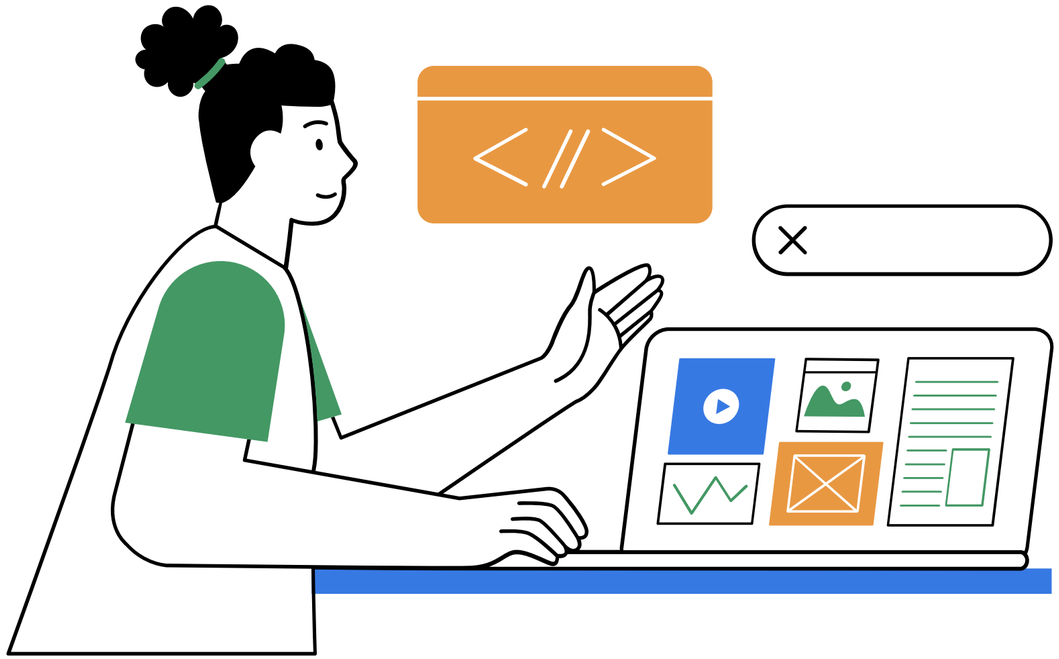
הבעיה

הפרויקט מתמקד בחיזוי סטטוס ההסמכה האנרגטי של מבנים בישראל. מדובר בפרמטר מרכזי בתכנון עירוני ובבנייה בת-קיימא, המספק תובנות חשובות לגבי עמידה בתקנים סביבתיים. הבנה טובה יותר של הדפוסים המשפיעים על עמידה בתקנים יכולה לתרום לשיפור תהליכי תכנון, להפחתת עלויות תפעול ועוד.

מטרות הפרויקט

מטרת הפרויקט היא לבצע תהליך חיזוי מלא, כולל בחירת תכונות, עיבוד מוקדם של הנתונים, בניית מודלים מפוקחים ולא-מפוקחים, והשוואת הביצועים שלהם. בנוסף, ניתוח הגורמים המשפיעים על סטטוס ההסמכה האנרגטי נועד לספק תובנות מעמיקות לצורך שיפור תהליכי בנייה ותכנון עירוני.

שיטות & טכניקות בפרויקט



הפרויקט עושה שימוש בטכניקות מתקדמות לניתוח נתונים, כגון הנדסת תכונות ליצירת משתנים חדשים ושימוש בטכניקות איזון כמו SMOTE לטיפול בחוסר איזון בקטגוריות היעד.
בבחירת המודלים נבדקו מספר אלגוריתמים:

מודלים מפוקחים: Decision Tree, Random Forest, Gradient Boosting, XGBoost, SVM, KNN.

מודלים לא-מפוקחים: K-Means, DBSCAN

כל מודל נבחר על בסיס התאמתו לנתונים ולאתגרי הפרויקט.

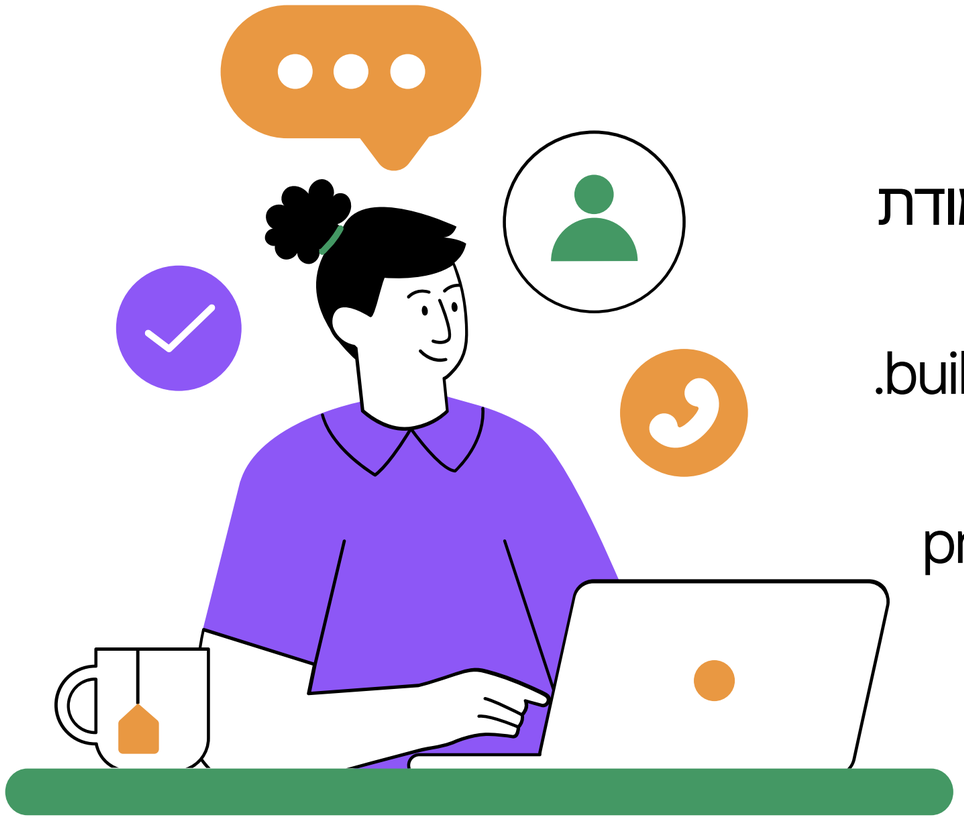
הדטאסט & תכונות עיקריות

הדאטהסט מכיל מידע על פרויקטים בתחום הבנייה בישראל, ונלקח ממאגר הנתונים הממשלתי data.gov.il. הוא כולל 4862 שורות ו-47 עמודות.

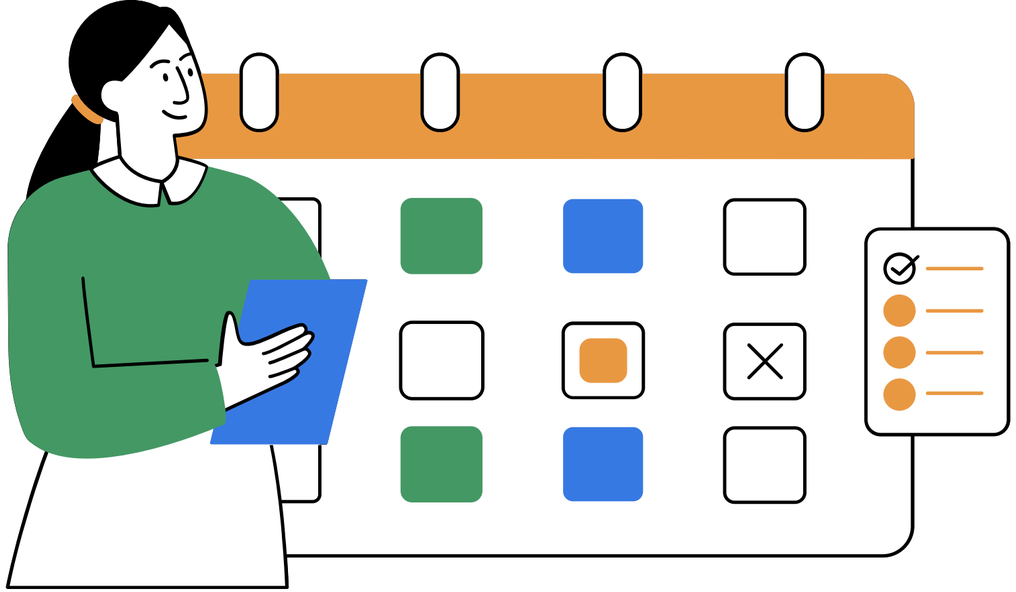
התכונות העיקריות כוללות:

1. קטגוריאליות: municipality_name, certificate_energy_pre (עמודת המטרה).
2. מספריות: building_area, residential_units, floors_above_ground.
3. גיאוגרפיות: X, Y, gush, helka.
4. תאריכים: project_general_last_update, certificate_date_pre

כמו כן, עמודת המטרה certificate_energy_pre אינה מאוזנת, דבר שדרש שימוש בטכניקות איזון נתונים.



מתודולוגיה



המתודולוגיה כוללת ארבעה שלבים עיקריים: הנדסת תכונות, איזון נתונים, בחירת מודלים מפוקחים ובחירת מודלים לא-מפוקחים.

1.<u>הנדסת תכונות</u> -יצירת משתנים חדשים לשיפור איכות הנתונים וקשרים עם עמודת המטרה:

- project_days_since_update: מספר הימים מאז העדכון האחרון של הפרויקט.
- density: יחס יחידות דיור לשטח המבנה (residential_units/building_area).
- region: חלוקה גיאוגרפית לאזורים (צפון, מרכז, דרום).

2.<u>איזון נתונים</u> -הדאטהסט הציג חוסר איזון משמעותי בקטגוריות עמודת המטרה certificate_energy_pre, כאשר קטגוריות מסוימות היו דומיננטיות ואחרות בעלות ייצוג נמוך.

- שימוש בשיטת SMOTE (Synthetic Minority Oversampling Technique) לאיזון קטגוריות בעמודת המטרה.
- יצירת דוגמאות סינתטיות לקטגוריות בעלות ייצוג נמוך תוך שמירה על המבנה הכללי של הדאטהסט.

מתודולוגיה- המשך

3. בחירת מודלים מפוקחים- נבחרו אלגוריתמים מגוונים בהתאמה לנתונים:

1. Decision Tree: פשטות ופרשנות.
2. Random Forest: עמידות לחוסר איזון.
3. Gradient Boosting ו-XGBoost: התאמה לנתונים מורכבים.
4. SVM: עבור נתונים לא ליניאריים.
5. KNN: לניסויים ראשוניים (אך נמצא רגיש לרעש).

4.בחירת מודלים לא-מפוקחים

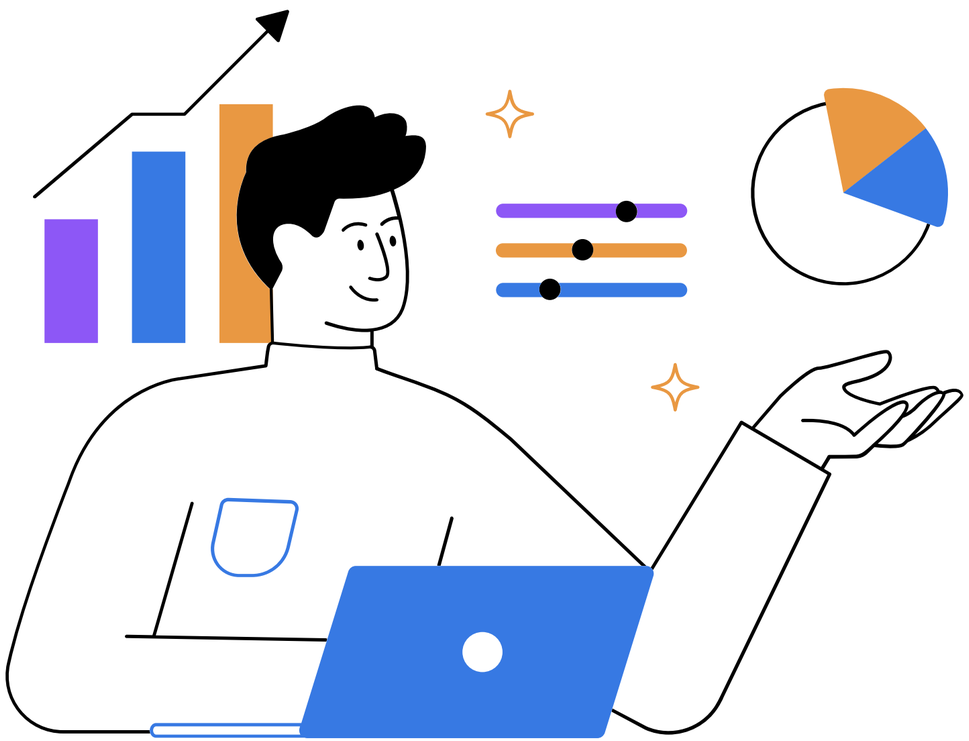
1. K-Means: חלוקה לקבוצות מבוססת מרחק אוקלידי עם בחירת K בעזרת Elbow Method.
2. DBSCAN: זיהוי אשכולות על בסיס צפיפות, כולל טיפול בנקודות חריגות (Outliers).

ניסויים ותוצאות

הדאטהסט פוצל לסט אימון (80%) וסט בדיקה (20%) באמצעות הפונקציה train_test_split, תוך שימוש ב- random_state=42 להבטחת שחזוריות הניסויים.

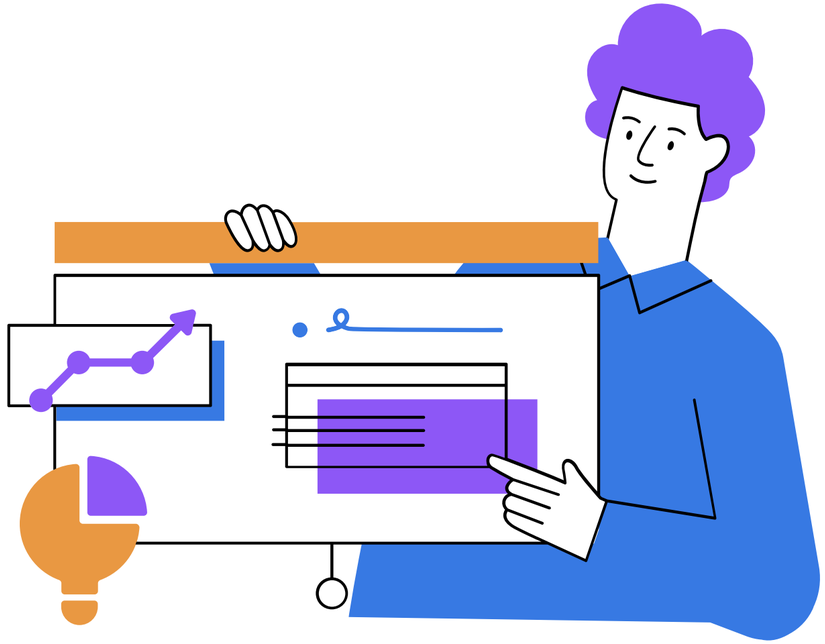
נתחיל עם הלא מפוקחים

מודל	Silhouette Score	הסבר
K-Means	0.15	התקשה ליצור אשכולות משמעותיים עקב מורכבות מבנה הנתונים.
DBSCAN	-0.01	לא זיהה אשכולות משמעותיים והיה רגיש מאוד לרעש.



Model	F1	Recall	Presicion	Accuracy	הסבר
Decision Tree	0.77	0.77	0.78	0.77	ביצועים מאוזנים, מתאים לדפוסים פשוטים.
Random Forest	0.82	0.83	0.83	0.83	ביצועים חזקים עם איזון טוב ועמידות לחוסר איזון.
Gradient Boosting	0.77	0.77	0.77	0.77	מתאים לנתונים מורכבים, אך פחות יעיל מ-Random Forest.
XGBoost	0.81	0.82	0.82	0.82	מותאם לנתונים מורכבים, ביצועים דומים ל-Random Forest.
kNN	0.65	0.63	0.69	0.63	רגיש לרעש ולחוסר איזון, ביצועים חלשים.

השוואת טכניקות ואלגוריתמים



בפרויקט נבדקו מגוון מודלים מפוקחים ולא-מפוקחים. מבין המודלים המפוקחים, Random Forest ו-XGBoost הצטיינו בהתמודדות עם הנתונים המורכבים והלא מאוזנים, תוך הצגת דיוק גבוה ואיזון מצוין בין Precision ל-Recall. לעומתם, Gradient Boosting ו-Decision Tree הראו ביצועים טובים בזיהוי דפוסים פשוטים, אך היו פחות יעילים בהתמודדות עם נתונים מורכבים. המודל kNN הציג ביצועים נמוכים יחסית, בעיקר עקב רגישותו לרעש ולחוסר איזון. במודלים הלא-מפוקחים, K-Means ו-DBSCAN התקשו לייצר אשכולות משמעותיים בשל מבנה הנתונים המורכב ורמת הרעש הגבוהה. בהתאם לכך, המודלים המבוססים על עצים הם המומלצים ביותר לפרויקט זה.

מסקנות ותובנות

בפרויקט זה הדגמנו את היכולת להשתמש בלמידת מכונה לצורך ניתוח נתונים מורכבים ובלתי מאוזנים בתחום הבנייה בישראל.
המודלים המבוססים על עצים, במיוחד Random Forest ו-XGBoost, הראו את היכולת הטובה ביותר להתמודד עם האתגרים של נתוני הפרויקט, תוך שמירה על דיוק ואיזון גבוהים בביצועים.
בנוסף, טכניקות כמו הנדסת תכונות ואיזון נתונים (SMOTE) היו קריטיות לשיפור ביצועי המודלים. יחד עם זאת, קטגוריות נדירות כמו A+ נותרו מאתגרות לזיהוי, ומדגישות את החשיבות של הרחבת הדאטהסט בעתיד.
הפרויקט מספק בסיס יציב להמשך פיתוח ושימוש בלמידת מכונה לצורך שיפור תכנון וניהול בתחום הבנייה בישראל, עם פוטנציאל לניתוח מגמות ולשיפור תהליכים קיימים.

תודה !

- https://github.com/AdarSaban/final_projecr_ml