המחלקה להנדסת תעשייה וניהול
- סמסטר א' תשפ"ה -
דוח מסכם "נושאים מתקדמים בלמידת מכונה"

מגישים	אדר סבן	אליה יהודה זגורי
תעודות זהות	313174120	207313131

מרצה: ד"ר חן חג'ג'.
תאריך : 20.1.2025

תוכן עניינים

מבוא

הפרויקט עוסק בניתוח נתונים בתחום הבנייה בישראל, עם התמקדות בחיזוי סטטוס ההסמכה 'האנרגטי' של מבנים (certificate_energy_pre). סטטוס זה מציין את מידת העמידה של מבנים בתקנים אנרגטיים וידידותיים לסביבה, מהווה מדד חשוב בתכנון עירוני ובניהול פרויקטים.
שאפנו להשתמש בכלים ומודלים שנלמדו בקורס בשילוב עם ידע מקורסים קודמים.
קישור לפרויקט - https://github.com/AdarSaban/final_projecr_ml

מטרות הפרויקט

מטרת הפרויקט היא לבצע תהליך מלא של חיזוי הכולל עיבוד מוקדם של הנתונים, הנדסת תכונות, בחירה ואימון של מודלים, וכן השוואה והערכה של ביצועי המודלים. בנוסף, נבדקו ניתוחים בלתי מפוקחים נוספים, כגון זיהוי חריגות, ונעשה ניתוח מעמיק של הגורמים המשפיעים על סטטוס ההסמכה.

מוטיבציה

למען ההגינות נאמר, זה לא הנושא הראשון שאותו בחרנו, אך עקב זאת שזוג נוסף מהקורס בחר את הנושא הראשון שרצינו, החלטנו לוותר להם ולהתמקד בביצוע המשימה עם הכלים שנלמדו. שנינו אנשים העובדים בענף הבנייה אז כן יכולנו להתחבר לנושא, עם הכרות במושגים, תהליכי עבודה ועוד.
נתונים בתחום הבנייה מהווים בסיס לקבלת החלטות מושכלות בתכנון עירוני ופרויקטים סביבתיים. הבנת הדפוסים המשפיעים על סטטוס ההסמכה האנרגטי יכולה לסייע לשפר תהליכים תכנוניים בבנייה בת-קיימא, לתמוך במנהלי פרויקטים, מתכננים וגופים ממשלתיים בקבלת החלטות מבוססות נתונים, להקטין עלויות תפעול ארוכות טווח ולשפר את איכות החיים.

דאטה סט ותכונות

כאמור הדטאסט הנבחר נלקח מהאתר data.gov.il מאגר נתונים ממשלתי פתוח. הדאטהסט מורכב מנתונים טבלאיים , הכוללים מידע מובנה שמאורגן בשורות ועמודות. כל שורה מייצגת פרויקט בנייה או מבנה בודד, וכל עמודה מתארת מאפיין מסוים של המבנה, כגון פרטים גיאוגרפיים, מאפייני בנייה, ותוצאות הסמכה אנרגטית.

מפרט הנתונים

מספר שורות :4862
מספר עמודות / מאפיינים : 47.

סוגי המשתנים העיקריים

קטגוריאליים: municipality_name, certificate_energy_pre, project_halted.
מספריים: building_area, residential_units, floors_above_ground.
תאריכים: project_general_last_update, certificate_date_pre.
גיאוגרפיים: X, Y, gush, helka.

תובנות ראשוניות מהנתונים

עמודת המטרה certificate_energy_pre, המייצגת את סטטוס ההסמכה האנרגטי, היא הקריטית ביותר בפרויקט ומהווה את היעד לחיזוי. ניתוח ראשוני העלה כי העמודה אינה מאוזנת, עם דומיננטיות של קטגוריות מסוימות וייצוג נמוך של אחרות, מה שעלול להוביל להטיית המודלים.
רשומות עם ערכים חסרים בעמודת המטרה הוסרו לחלוטין, כיוון שהן אינן רלוונטיות לתהליך החיזוי. בנוסף, עמודות עם מעל 70% ערכים חסרים הוסרו מהדאטהסט, וערכים חסרים בודדים בעמודות קטגוריאליות

ומספריות מולאו בערך הנפוץ ביותר (mode) ובחציון (median), בהתאמה.
בחירה זו נבעה ממספרם המועט של ערכים חסרים בעמודות אלו ומתוך גישה מותאמת לסוג הנתונים: בעמודות קטגוריאליות, הערך הנפוץ מייצג את הדפוס השכיח בנתונים ומבטיח עקביות, בעוד שבמספריות, החציון עמיד בפני ערכים חריגים (outliers) שמסוגלים לעוות את המידע.
תהליכי ניקוי נוספים לא בוצעו, מתוך מטרה לשמור על שלמות הנתונים. עמודת המטרה נמצאה כקשורה פוטנציאלית לתכונות מרכזיות אחרות, כמו building_area, floors_above_ground, ו-region, שייבחנו לעומק בהמשך תהליך המודלים. ביצענו מספר אופציות נוספות למילוי הנתונים אך דווקא שיטות אלו התגלו בהמשך כטובות ביותר למודל.

מתודולוגיה

המתודולוגיה תחולק ל- 4 חלקים, כמובן לאחר שבוצע שלב עיבוד הנתונים.

- הנדסת תכונות -שלב זה נועד לשפר את איכות הנתונים על ידי יצירת תכונות חדשות או התאמת התכונות הקיימות, כך שידגישו קשרים רלוונטיים בין הנתונים לעמודת המטרה.

- איזון הנתונים - עמודת המטרה הראתה חוסר איזון בין הקטגוריות, ולכן נעשה שימוש בטכניקות לאיזון הנתונים, על מנת למנוע הטיה של המודלים לטובת קטגוריות דומיננטיות.

- בחירת המודלים - נבחרו מודלים שונים המותאמים לנתונים מגוונים (קטגוריאליים ומספריים) ולבעיות סיווג, כדי לבצע השוואת ביצועים ולמצוא את המודל המתאים ביותר.

הנדסת תכונות

במהלך תהליך הנדסת התכונות, הוספנו עמודות חדשות בעלות ערך לחיזוי:

- project_days_since_update: מספר הימים שעברו מאז העדכון האחרון של המידע על הפרויקט. תכונה זו נועדה להכניס מידע כרונולוגי.שמשקף את עדכניות המידע על הפרויקט. ייתכן שפרויקטים מתקדמים יותר יעמדו ביתר קלות בתקני ההסמכה האנרגטיים.
- density: צפיפות, המחושבת כשטח הבנייה (building_area) חלקי מספר יחידות הדיור (residential_units). תכונה זו מוסיפה פרספקטיבה על יעילות הניצול של השטח.ומהווה רכיב חשוב עבור עמודת המטרה.
- region: האזור הגיאוגרפי שבו נמצא המבנה (כגון צפון, מרכז, דרום). תכונה זו נועדה להוסיף מידע גיאוגרפי שעשוי להשפיע על הביצועים הסביבתיים של המבנה. ייתכן שאזורי אקלים שונים או צפיפות אוכלוסין משתנה משפיעים על עמידה בתקני ההסמכה הירוקים.

איזון נתונים

עמודת המטרה certificate_energy_pre התגלתה כלא מאוזנת, עם ייצוג יתר של קטגוריות מסוימות וחוסר ייצוג של קטגוריה נדירה A+. איזון הנתונים בוצע באמצעות שיטת SMOTE (Synthetic Minority Oversampling Technique), אשר יוצרת דוגמאות סינתטיות לקטגוריות המיוצגות פחות, תוך שמירה על המבנה הכללי של הדאטה סט.כך ניתן להשיג מודל מאוזן שמייצג היטב את כלל הקטגוריות.

בחירת מודלים - מפוקחים

נבחרו מודלים שונים המותאמים לנתונים מגוונים (קטגוריאליים ומספריים) ולבעיות סיווג, כדי לבצע השוואת ביצועים ולמצוא את המודל המתאים ביותר.כל המודלים הוערכו על פי דיוק Accuracy, Precision, Recall, ו-F1-Score.

- **Decision Tree**: עץ החלטה מתאים במיוחד לנתונים עם קשרים פשוטים וברורים בין משתנים. הדאטה-סט שלנו מכיל משתנים גיאוגרפיים, מספריים וקטגוריאליים, שעץ החלטה יכול לטפל בהם בצורה טבעית, ללא צורך בעיבוד מורכב.עץ החלטה מספק פרשנות ברורה ותובנות על החשיבות היחסית של כל משתנה בתהליך החיזוי.
- **Random Forest**: מתמודד היטב עם נתונים לא מאוזנים, כמו במקרה של עמודת המטרה. הוא עמיד בפני רעש (Noise) בערכים ועוזר למנוע overfitting, בעיה נפוצה במודלים פשוטים יותר.שילוב של מספר עצי החלטה יוצר מודל יציב וחזק יותר, שמתאים במיוחד לדאטהסט עם משתנים מגוונים כמו שלנו.
- **Gradient Boosting:** מתאים לבעיות חיזוי מורכבות, בהן הקשרים בין המשתנים אינם ליניאריים. הדאטהסט כולל משתנים כמו density ו-floors_above_ground, שייתכן שיש להם קשרים מורכבים עם עמודת המטרה. השיטה מייעלת את תהליך הלמידה באמצעות תיקון שגיאות של מודלים קודמים, מה שמשפר את ביצועי המודל.
- **XGBoost**: גרסה מתקדמת של Gradient Boosting, מותאם במיוחד לדאטה-סטים גדולים יחסית עם חוסר איזון מסוים בין הקטגוריות. היכולת של XGBoost להתמודד עם ערכים חסרים ויעילותו הגבוהה הופכות אותו לבחירה מתאימה. מספק דיוק גבוה עם זמן חישוב סביר, ומאפשר כיוונון פרמטרים מדויק לשיפור נוסף.
- **SVM :** מתאים לנתונים שאינם ליניאריים. מכיוון שעמודת המטרה מכילה קטגוריות שייתכן שהן אינן ניתנות להפרדה ברורה במרחב הפיצ'רים, SVM מסוגל להפריד ביניהן על ידי מיפוי למרחב מימדים גבוה יותר וחזק במיוחד במקרים בהם הגבול בין הקטגוריות אינו ברור. בפועל המודל לא הצליח לחזות את הערך הנדיר בעמודת המטרה , אז הוחלף במודל KNN שהביא חיזוי מדויק יותר.

בחירת מודלים - לא מפוקחים

- **K-Means :**הוא אלגוריתם המבצע חלוקה של נתונים לקבוצות על בסיס קרבה במרחב התכונות, באמצעות חישוב מרחק אוקלידי.הנתונים שלנו כוללים משתנים מספריים מרכזיים, כמו building_area, floors_above_ground, ו-residential_units, שמתארים מאפיינים פיזיים של מבנים. לאחר סקלת נתונים, משתנים אלה מתאימים במיוחד לניתוח מבוסס מרחק, שהוא עיקרון מרכזי ב-K-Means. השתמשנו ב-ELBOW METHOD לבחירת מספר האשכולות K בניתוח זה.

- **DBSCAN:** הוא אלגוריתם המבוסס על צפיפות נתונים. הוא מזהה קלאסטרים גם בצורות לא ליניאריות, מסוגל להתמודד עם נקודות חריגות (Outliers).הנתונים הגיאוגרפיים שלנו (X, Y, gush, helka) מייצגים מיקומים פיזיים של מבנים, שייתכן ויש להם קשר צפיפותי מובחן. מכיוון שאין צורך להגדיר מראש את מספר הקלאסטרים, האלגוריתם מתאים לניתוח ראשוני במקרים בהם לא ברור כיצד לחלק את הנתונים, מה "שיושב" על הפרוקיט שלנו בצורה די טובה.

ניסויים ותוצאות

הנתונים חולקו לסט אימון (80%) ולסט בדיקה (20%) באמצעות train_test_split. הפיצול בוצע בצורה אקראית תוך שימוש בפרמטר random_state=42, כדי להבטיח שחזוריות של הניסויים.

בחירת פרמטרים

Random Forest: השתמשנו ב-n_estimators=100 למספר עצים גבוה ויציב, ו-max_depth=None ללמידת דפוסים מורכבים.
XGBoost: נבחרו n_estimators=100 לאיזון ביצועים וזמן עיבוד, learning_rate=0.1 ללמידה הדרגתית, ו-max_depth=6 למניעת התאמת יתר.
Decision Tree: הוגבל ל-max_depth=5 כדי למנוע התאמת יתר, תוך שימוש ב-Gini Index כקריטריון פיצול.
K-Means: נבחר על בסיס Silhouette Score, ו-k-means++init="" לאתחול יעיל.
SMOTE: הוגדר k_neighbors=5 לגיוון דוגמאות סינתטיות ו-sampling_strategy="minority" לאיזון הקטגוריה המועטה.
KNN:
n_neighbors=5 , מספר שכנים לאיזון בין רעש לדיוק. מטריקת Minkowski עם ברירת המחדל (p=2)
Gradient Boosting:
n_estimators=100: מספר איטרציות לשיפור הדרגתי.learning_rate=0.1: קצב למידה המווסת את תרומת כל עץ.max_depth=3: עומק מרבי למניעת התאמת יתר.learning_rate=0.1: קצב למידה המווסת את תרומת כל עץ.

מדדי הערכה וחשיבותם

- Accuracy -הדיוק מספק אינדיקציה בסיסית של אחוז התחזיות הנכונות, אך הוא אינו מתאים כמדד עיקרי בפרויקט זה. במקרה של חוסר איזון בנתונים, דיוק גבוה עלול לשקף מודל שמנבא תמיד את הקטגוריה הנפוצה ביותר, כמו קטגוריה "A" או "B", מבלי להתחשב בקטגוריות אחרות, כמו "C". זה עלול לגרום לכך שהמודל לא יתפקד היטב בקטגוריות החשובות בפרויקט.
- Precision -.Precision חשוב במיוחד לזיהוי מבנים שמסומנים כמתאימים להסמכה אנרגטית. לדוגמה, תחזית שגויה שמסמנת מבנה כמתאים (כאשר הוא לא באמת מתאים) עלולה להוביל להשלכות תכנוניות חמורות או לאישור מבנה שלא עומד בתקנים. מדד זה מבטיח שכאשר המודל מזהה מבנה כמתאים, הוא עושה זאת ברמת ודאות גבוהה.
- Recall -Recall קריטי בפרויקט זה כדי לוודא שאף מבנה חשוב לא "ייפול בין הכיסאות". לדוגמה, אם מבנה שאינו עומד בתקנים לא יזוהה על ידי המודל (False Negative), זה עלול לגרום להמשך בעיות תכנוניות או סביבתיות. Recall מבטיח שהמודל מזהה את כל המבנים הלא מתאימים, גם במחיר של כמה False Positives.
- F1-Score הוא מדד מרכזי בפרויקט זה, שכן הוא מספק תמונה מאוזנת של ביצועי המודל במצבים שבהם יש קטגוריות נדירות.F1-Score משלב בין Precision ו-Recall ומהווה מדד מאוזן להערכת ביצועי המודל. זהו המדד החשוב ביותר בפרויקט שלנו, שכן הוא מאפשר התייחסות גם לדיוק וגם לזכירה. בעבודה עם עמודת מטרה לא מאוזנת כמו certificate_energy_pre, שבה חשוב לזהות גם קטגוריות נדירות, מדד זה משקף את הביצועים הכלליים בצורה המיטבית.
- Silhouette Score (במודלים לא מפוקחים)
  מדד להערכת איכות אשכולות בניתוח K-Means.
  ערכים גבוהים מצביעים על הפרדה טובה בין אשכולות.

הסבר על הממצאים

הצלחות

Random Forest: השיג את הביצועים הטובים ביותר (דיוק 83%, F1 Score 0.82), עם איזון טוב בזיהוי קטגוריות.
XGBoost: תוצאות קרובות ל-Random Forest (דיוק 82%, F1 Score 0.81), מתאים לנתונים מורכבים.

Gradient Boosting ו-Decision Tree: סיפקו ביצועים טובים (דיוק 77%), מתאימים לדפוסים פשוטים יותר.

מגבלות

kNN: דיוק נמוך (63%, F1 Score 0.65), רגיש לרעש ולחוסר איזון בנתונים.

K-Means: Silhouette Score נמוך (0.15), מעיד על הפרדה מוגבלת.

DBSCAN: Silhouette Score שלילי (-0.01), התקשה בזיהוי אשכולות וזיהה הרבה רעש.

תוצאות בלתי צפויות

הקטגוריה A+ לא זוהתה היטב בכל המודלים בשל מיעוט דוגמאות.

kNN הצליח ב-Recall בקטגוריה זו (1.00), אך ביצועיו בקטגוריות אחרות היו חלשים.

DBSCAN נכשל בזיהוי אשכולות משמעותיים, כנראה עקב מבנה הנתונים.

מסקנה

אלגוריתמים מבוססי עצים (Random Forest ו-XGBoost) הראו יתרון ברור בנתונים אלו, בעוד ש-kNN ו-K-Means התקשו להתמודד עם חוסר איזון ומורכבות. בחירה נכונה של טכניקות איזון ומודלים מתקדמים מאפשרת שיפור משמעותי בביצועים ובזיהוי דפוסים.

ביצועי אלגוריתמים

Random Forest: ביצועים גבוהים בזכות שילוב עצים מרובים, המסייע בצמצום השפעת רעש וחוסר איזון בנתונים.

XGBoost: דומה ל-Random Forest, אך מהיר יותר ומתאים במיוחד לנתונים מורכבים הודות לאופטימיזציות מתקדמות.

Gradient Boosting: סיפק ביצועים טובים, אך היה פחות מדויק מהאלגוריתמים הקודמים בשל רגישות להתאמת יתר.

Decision Tree: מתאים לדפוסים פשוטים אך פחות יעיל במקרים של מורכבות נתונים.

kNN: רגישות גבוהה לרעש ולחוסר איזון בנתונים פגעה בביצועיו.

K-Means ו-**DBSCAN**: התקשו ליצור אשכולות משמעותיים בשל מבנה הנתונים ורגישות לפרמטרים.

מסקנה: אלגוריתמים מבוססי עצים הוכיחו את יעילותם בניתוח נתונים מורכבים לעומת אלגוריתמים רגישים יותר כמו kNN ו-K-Means.

מסקנות ודיון

בפרויקט זה התמקדנו בניתוח מבנים ירוקים בישראל באמצעות למידת מכונה, תוך התמודדות עם נתונים מורכבים ורבי אתגרים. אחד האתגרים המשמעותיים היה טיפול בערכים חסרים, במיוחד בעמודות קריטיות כמו תקני אנרגיה (certificate_energy_pre) ונתוני עדכון. נאלצנו לקבל החלטות מושכלות לגבי הסרת עמודות עם ערכים חסרים רבים או שימוש בטכניקות למילוי נתונים, כדי לשמר את איכות המידע מבלי להטות את המודלים. בנוסף, בדקנו מאפיינים כמו התאמת מבנים לאקלים האזורי ושימוש באשכולות לזיהוי מגמות ייחודיות, אך הם לא הוכנסו למודלים בשל מגבלות נתונים או אי-התאמה לתוצאות הניתוח.

בשלב בניית המודלים השתמשנו במגוון אלגוריתמים כדי להתמודד עם מורכבות הנתונים. המודלים Random Forest ו-XGBoost הצטיינו במיוחד, עם דיוק של מעל 80% ויכולת לזהות מבנים ירוקים תוך שמירה על איזון בין Precision ו-Recall. מודלים אלו הראו עמידות לחוסר איזון בנתונים והפגינו ביצועים יציבים הודות ליכולתם לשלב תובנות ממספר רב של עצים. לעומתם, Gradient Boosting ו-Decision Tree סיפקו ביצועים טובים יותר עבור דפוסים פשוטים, אך היו מוגבלים בהבנת דפוסים מורכבים.

המודלים הלא מפוקחים, K-Means ו-DBSCAN, התקשו להציג אשכולות משמעותיים, כפי שנראה מה-Silhouette Score הנמוך. עם זאת, ניתוח זה סיפק תובנות כלליות על מבנים חריגים שאינם עומדים בתקן.
תהליך Feature Engineering תרם לשיפור ביצועי המודלים על ידי הוספת מאפיינים כמו חלוקה לאזורים גיאוגרפיים ונתוני עדכון כרונולוגיים. יחד עם זאת, קטגוריות נדירות כמו A+ נותרו מאתגרות לזיהוי, עקב מיעוט דוגמאות והשפעתן המוגבלת על תהליך האימון.
הפרויקט מדגיש את הפוטנציאל של למידת מכונה לניתוח קיימות בישראל ומציב בסיס להמשך פיתוח, כולל שילוב נתונים סביבתיים נוספים, כמו תנאי אקלים וצריכת אנרגיה, ושימוש במודלים מתקדמים להעמקת הניתוח. העבודה בצוות כללה שילוב של ניקוי נתונים, בניית מודלים וניתוח תוצאות, והובילה להפקת תובנות משמעותיות שעשויות לתרום ליישום פתרונות חדשניים בתחום הקיימות.