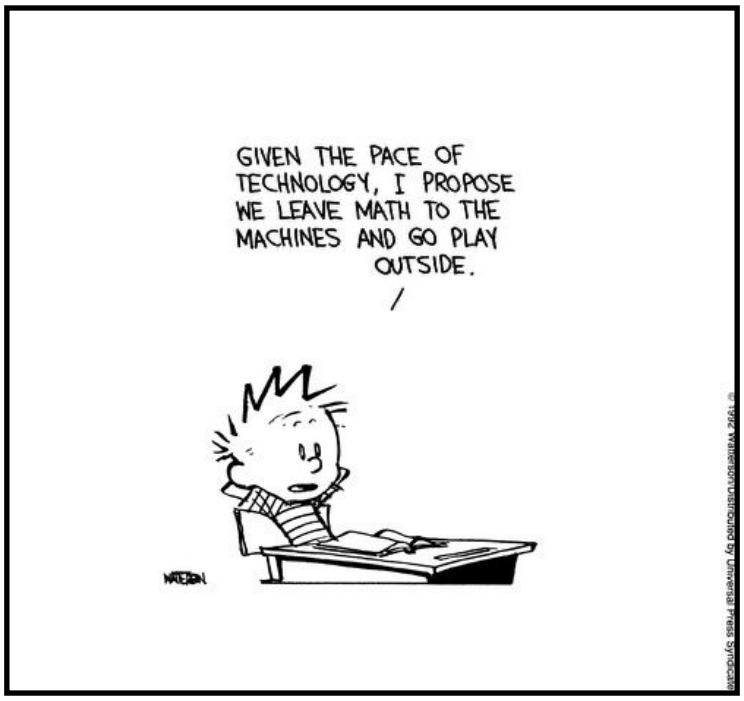
Advanced Topics in Machine Learning

Dr. Chen Hajaj

Course Project



Team members
Dima Levin
314144148

Section 1 - Introduction

- Introduce the problem and its significance, underlining why it was chosen.
- Highlight your objectives of this work (i.e., what you aim to achieve)

Section 2 - Dataset and Features

- Detail the API, crawling procedure, and dataset.
- Discuss the preprocessing steps, feature selection and explain the rationale behind these choices.

Section 3- Methodology

- Describe the algorithms and techniques employed, emphasizing the rationale for their
- selection.
- This section should not only describe what was done but also why, showcasing your
- analytical thinking and the strategic choices made during the project.

Section 4- Experiments and Results

Outline your experimental design, including:

- Parameter Choices: justify the parameters selected, reflecting on how they influence the outcomes.
- Discuss the metrics used for evaluation, such as accuracy, precision, recall, F1 score, etc., and explain why they are appropriate for your problem.
- Explain the significance of these metrics is crucial for demonstrating the depth of your analytical skills.
- Provide a comprehensive explanation of your findings. Analyze the success, limitations, and unexpected outcomes of your experiments.
- Algorithm Performance: evaluate how different algorithms you've utilized performed and hypothesize why some were more effective than others.

Section 5 - Conclusion and Discussion

- Summarize your contributions and reflect on the project's implications.
- Detail the roles and contributions of each team member.
- Discuss potential future directions you or others might explore, considering the lessons learned and new questions that have emerged.

חלק 1 - הקדמה

תחזוקה ומצב של מתקני ספורט ממלאים תפקיד מרכזי בקידום רווחת הקהילה, עידוד השתתפות בספורט ותמיכה במעורבות של בני נוער.
הרשויות המקומיות מתמודדות עם האתגר לנהל מתקנים אלו בצורה אפקטיבית, תוך הבטחתם להיות בטוחים, נגישים ומתאימים למטרותיהם.
עם זאת, מגבלות משאבים מובילות לעיתים קרובות לתחזוקה ושימוש לא אחידים במתקנים אלה, מה שמדגיש את הצורך בתובנות מבוססות נתונים לשיפור הניהול שלהם.

פרויקט זה יתמודד עם הבעיה באמצעות ניתוח נתונים להערכת תחזוקת מתקני ספורט ולחיזוי מצבם בערים שונות בישראל.
באמצעות ניתוח מאגר נתונים המכיל מאפיינים כמו גיל המתקן, זמינות תשתיות (לדוגמה, תאורה, גידור) ודפוסי שימוש, אנו שואפים להפיק תובנות יישומיות שיתמכו בקבלת החלטות טובה יותר על ידי הרשויות המקומיות.

המטרות המרכזיות של העבודה:

- פיתוח מודלים חיזויים המסוגלים לקבוע את מצב המתקנים בהתבסס על מאפייניהם, מה שיאפשר תכנון תחזוקה יזום.
- ביצוע clustering כדי לקבץ מתקנים לקטגוריות משמעותיות, זיהוי תבניות כמו מתקנים לא מנוצלים או מוזנחים.
- זיהוי חריגות במאגר הנתונים, תוך הדגשה של מתקנים שסוטים משמעותית מהמצופה, כגון מתקנים חדשים עם תשתיות חסרות.
- הפקת המלצות יישומיות לשיפור ניהול המתקנים והקצאת המשאבים.

📈 חלק 2 - מאגר נתונים ותכונות

מאגר הנתונים ותהליך ה-crawling

מאגר הנתונים עבור פרויקט זה נלקח מפורטל הנתונים הפתוחים של ישראל, המספק נתונים ציבוריים בתחומים שונים. המאגר מכיל מידע מפורט על מתקני ספורט בערים שונות, כולל סוג המתקן, זמינות תשתיות, דפוסי שימוש ומאפיינים פיזיים.
הנתונים הורדו בפורמט CSV ויובאו לסביבת ניתוח מובנית עבור קדם-עיבוד ומידול.

שלבי קדם-עיבוד

1. **טיפול בערכים חסרים:**
עמודות עם נתונים חסרים משמעותית או מולאו או הוסרו על בסיס הרלוונטיות שלהן לניתוח.
   - ערכים חסרים מספריים (למשל, גיל המתקן) הוסרו.
2. **הנדסת מאפיינים (Feature Engineering):**
מאפיין חדש, "גיל המתקן", חושב על ידי חיסור שנת הבנייה מהשנה הנוכחית (2025).
משתנים קטגוריאליים כמו "זמינות" ו"מצב המתקן" קודדו לייצוגים מספריים לשימוש במודלים של למידת מכונה.
3. **בחירת תכונות (Feature Selection)**:
המאפיינים המרכזיים כללו:
   - גיל המתקן: משקף את גיל המתקן, משתנה קריטי במצבו.
   - זמינות גידור ותאורה: אינדיקטורים להשלמת תשתיות.
   - קטגוריית זמינות: משקפת את הנגישות ודפוסי השימוש במתקן.
   - עמודות לא רלוונטיות, כמו מזהים (ID, רחוב), הוסרו להפחתת רעשים בנתונים.

רציונל מאחורי הבחירות

- Imputation: הבטחת שלמות הנתונים מבלי לשנות משמעותית את שלמותם.
- הנדסת מאפיינים: מאפיינים כמו "גיל המתקן" ותשתיות מספקים תובנות חשובות לצרכי תחזוקה.
- קידוד משתנים קטגוריאליים: מאפשר שילוב קל במודלים הדורשים נתונים מספריים.
- בחירת מאפיינים: מיקוד במשתנים בעלי השפעה ישירה על מצב המתקן.

🎯 חלק 3 - מתודולוגיה

מודלים חיזויים

Decision Tree Classifier:

- תיאור: מודל מבוסס עצים שמחלק את הנתונים לתת-קבוצות לפי ספים של מאפיינים.
- רציונל: מספק מבנה ברור ואינטרפרטבילי, מתאים להבנת גורמים משפיעים על מצב המתקן.

Random Forest Classifier:

- תיאור: שיטה אנסמבלית המשלבת עצים רבים לשיפור ביצועים.
- רציונל: מונע overfitting ומספק תחזיות מדויקות יותר באמצעות ממוצע תוצאות.

Gradient Boosting Classifier:

- תיאור: שיטה אנסמבלית הבונה עצים עוקבים המתקנים טעויות קודמות.
- רציונל: מצטיין בטיפול בדאטה לא מאוזן ובקשרים מורכבים.

ניתוח clustering

K-Means Clustering:

- תיאור: אלגוריתם המפריד נתונים לקבוצות לפי קרבתם למרכזי הקבוצות.
- רציונל: מסייע בזיהוי תבניות במאפייני המתקנים.

זיהוי חריגות (Anomaly Detection)

Isolation Forest:

- תיאור: אלגוריתם זיהוי חריגות לא מפוקח, היוצר חלוקות אקראיות בנתונים כדי לבודד חריגות.
- רציונל: Isolation Forest יעיל לטיפול בנתונים בעלי מימד גבוה ומזהה מתקנים שסוטים משמעותית מהנורמה, כגון מתקנים חדשים ללא תשתיות חיוניות.
- הערכה: החריגות שזוהו נבדקו באופן ידני כדי לוודא את חשיבותן.

קדם-עיבוד וסטנדרטיזציה

סטנדרטיזציה (Standardization):

- תיאור: תכונות הותאמו כך שיהיו בעלות ממוצע 0 וסטיית תקן 1.
- רציונל: סטנדרטיזציה מבטיחה שכל התכונות יתרמו במידה שווה בשיטות מבוססות מרחק כמו clustering וזיהוי חריגות.

בחירות אסטרטגיות

- גיוון מודלים: השימוש במודלים מגוונים (Decision Trees, Random Forest, Gradient Boosting) מבטיח תחזיות מדויקות ומספק תובנות על היבטים שונים של הנתונים.
- מיקוד בתכונות: תכונות נבחרו לפי הרלוונטיות הישירה שלהן למצב המתקנים, כדי לאפשר למודלים להתמקד בגורמים משמעותיים.
- שיטות לא מפוקחות: clustering וזיהוי חריגות משלימים את המידול החיזוי בכך שהם חושפים תבניות נסתרות ונתונים חורגים.

📊 חלק 4 - ניסויים ותוצאות

עיצוב ניסויים

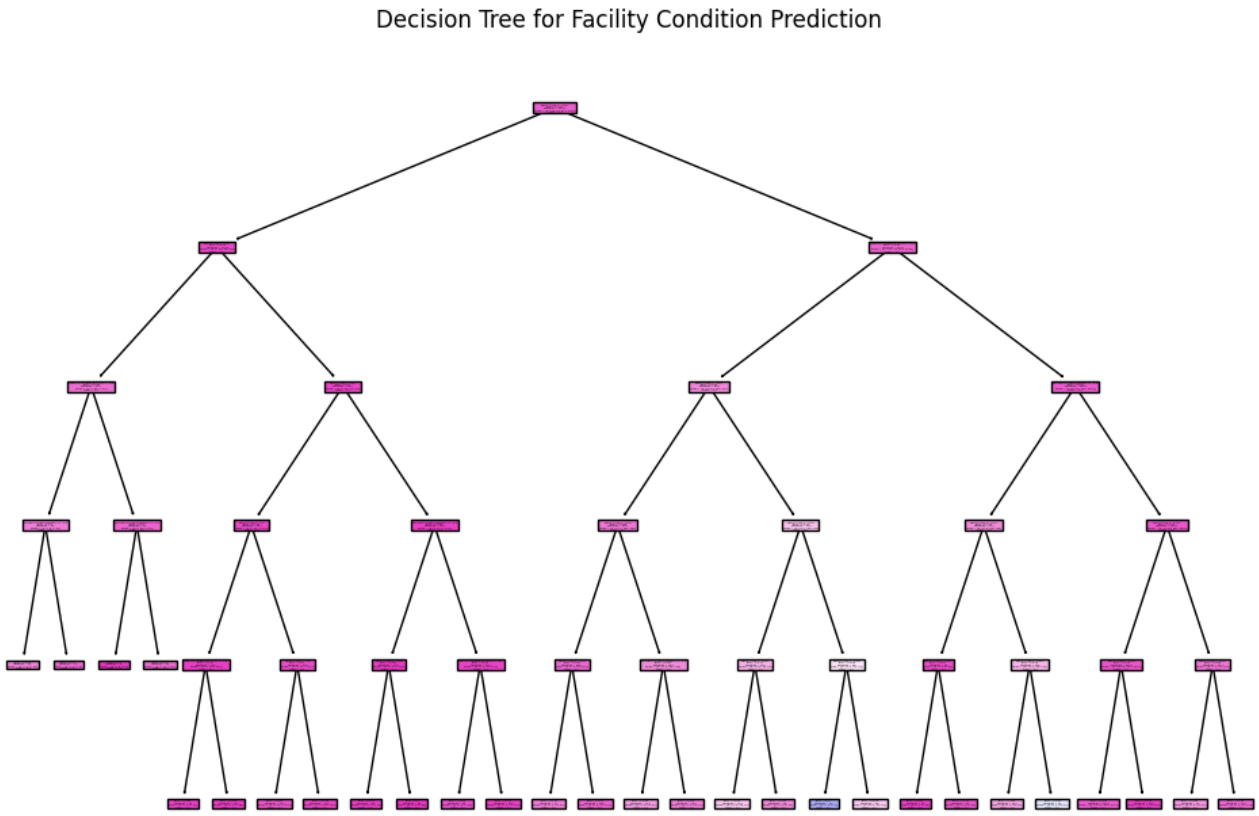
העיצוב התמקד בהערכת האפקטיביות של האלגוריתמים והטכניקות השונות על פני מאגר הנתונים. המודלים אומנו ונבחנו באמצעות חלוקה של 80%-20% לסטים של אימון ובדיקה, על מנת להבטיח כמות נתונים מספקת לאימות.

בחירות פרמטרים:

1. Decision Tree:
   עומק מקסימלי: 5 (לאיזון בין אינטרפרטביליות ודיוק).random_state: 42 (לשחזוריות).
2. Random Forest:
   מספר מעריכים: 100 (ליצירת תחזיות יציבות). עומק מקסימלי: 10 (למניעת overfitting).
   random_state: 42.
3. Gradient Boosting:
   learning_rate: 0.1 (לאיזון מהירות התכנסות ודיוק).ממספר estimators: 100. עומק מקסימלי: 5.

מדדי הערכה:

- Accuracy: מודד את נכונות התחזיות הכוללת.
- Precision: מציין את שיעור התחזיות הנכונות עבור כל קטגוריה.
- Recall: מדגיש את יכולת המודל לזהות את כל המקרים בפועל של קטגוריה מסוימת.
- F1 Score: משלב precision ו-recall למטריקה יחידה.
- Silhouette Score (ל-clustering): מעריך עד כמה הקבוצות נפרדות היטב.



ממצאים

מודלים חיזויים:

1. Decision Tree:
דיוק: (83.23%).
יתרונות: קל לאינטרפרטציה ולהדמיה.
חסרונות: נוטה ל-overfitting בעצים עמוקים יותר.
2. Random Forest:
דיוק: (83.23%) עקב שיטת האנסמבל.
יתרונות: חסין ל-overfitting ומספק תובנות על חשיבות התכונות.
חסרונות: פחות אינטרפרטבילי בהשוואה ל-Decision Tree.
3. Gradient Boosting:
דיוק: (83.23%) עם איזון טוב בין המטריקות.
יתרונות: מטפל היטב בקשרים מורכבים ואי-איזונים.
חסרונות: תובעני מבחינה חישובית.

Clustering:

- מספר אופטימלי של קבוצות: 3 (בהתבסס על elbow method).
- קטגוריות שזוהו: מתקנים מתוחזקים היטב, מתקנים שאינם בשימוש ומתקנים מוזנחים.
- Silhouette Score: 0.72, מעיד על קבוצות מופרדות היטב.

זיהוי חריגות:

- אחוז חריגות: 10% מהמתקנים.
- דוגמאות: מתקנים חדשים ללא תשתיות או מתקנים ישנים במצב מצוין.
- תובנות: הדגישו אזורים הזקוקים לתשומת לב.

ניתוח

- הצלחות: המודלים סיפקו תובנות יישומיות ותחזיות מדויקות שתומכות בקבלת החלטות מבוססות נתונים.
- מגבלות: קטגוריות לא מאוזנות (למשל, מעט מתקנים מוזנחים) השפיעו מעט על ה-recall.
- תוצאות בלתי צפויות: זיהוי חריגות חשף מתקנים שהתעלמו מהם בהנחות הראשוניות, כמו מתקנים ישנים במצב טוב.

📚 חלק 5 - מסקנות ודיון

סיכום התרומות

פרויקט זה ניצל בהצלחה ניתוח נתונים להערכת וחיזוי מצבם של מתקני ספורט בערים שונות בישראל. באמצעות שימוש במודלים חיזויים, clustering, וזיהוי חריגות, הניתוח סיפק תובנות יישומיות על מצב המתקנים. התרומות המרכזיות כוללות:

- פיתוח מודלים חיזויים חזקים (כגון Gradient Boosting) שהשיגו דיוק גבוה בתחזיות מצב המתקנים.
- זיהוי קבוצות משמעותיות של מתקנים, תוך חלוקה לקטגוריות של מתקנים מתוחזקים היטב, מתקנים שאינם בשימוש, ומתקנים מוזנחים.
- זיהוי חריגות, תוך הדגשת מתקנים שסוטים באופן משמעותי מהתבניות הצפויות ודורשים תשומת לב מיוחדת.
- מתודולוגיה מקיפה המשלבת טכניקות מפוקחות ולא מפוקחות, הממחישה את הפוטנציאל של גישות מבוססות נתונים בקבלת החלטות עירוניות.

שיקוף השלכות של הפרויקט

הממצאים מפרויקט זה נושאים השלכות מעשיות על הקצאת משאבים ותכנון תחזוקה על ידי הרשויות המקומיות. לדוגמה, זיהוי מתקנים מוזנחים יכול לסייע בתעדוף שיפוצים, בעוד שזיהוי חריגות עשוי לחשוף אי-יעילות נסתרת או מקרים יוצאי דופן. ניתוח ה-clustering מספק מסגרת לסיווג מתקנים בהתבסס על מצבם ודפוסי השימוש שלהם, המאפשרת התערבויות ממוקדות.

כיוונים עתידיים אפשריים

- שילוב נתונים נוספים: הרחבת מאגר הנתונים לכלול מאפיינים נוספים, כגון השקעות כספיות ומשוב משתמשים, עשויה לשפר את הדיוק החיזוי ואת התובנות.
- מעקב בזמן אמת: שילוב חיישני IoT במתקנים יכול לספק נתונים בזמן אמת על השימוש בתשתיות ומצבם, ולאפשר קבלת החלטות דינמית.
- מודלים מתקדמים: חקירת טכניקות למידה עמוקה או שיטות אנסמבל המשלבות clustering וחיזוי עשויה לשפר את הדיוק והעמידות.
- ניתוח מרחבי: הוספת נתונים גיאוגרפיים יכולה לסייע בניתוח פריסת המתקנים ונגישותם, ולהבטיח חלוקה שוויונית של משאבים.