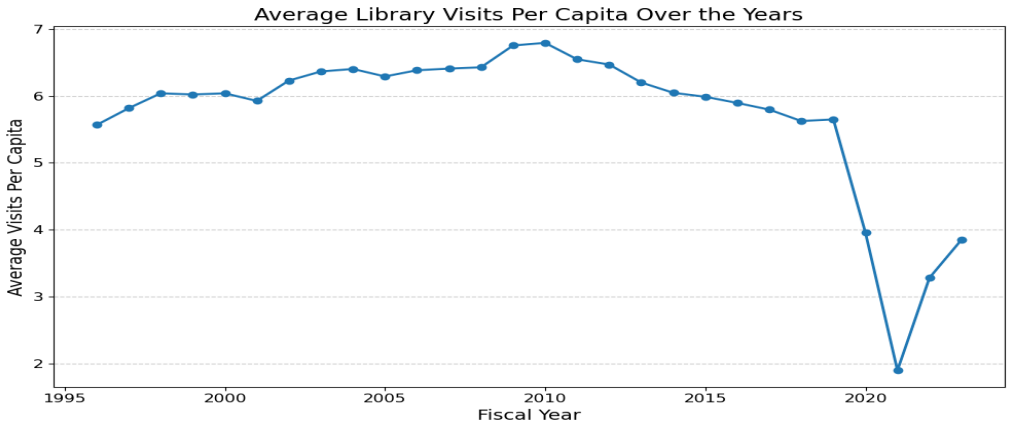
Section 1 - Introduction

הספריות הציבוריות נחשבות כעמודי התווך של החינוך, התרבות והמעורבות הקהילתית. מעבר למתן גישה לספרים ומשאבי למידה, לספריות יש תפקיד חשוב בקידום הישגים אקדמיים.

מחקרים תומכים ברעיון שלקרבה לספריות ציבוריות יש קשר חיובי עם שיפור מיומנויות קריאה והישגים לימודיים בקרב תלמידים, תוך הדגשת ערכן החברתי. לדוגמה, מחקר מצביע על קשר ישיר בין גישה למשאבי ספריה לבין שיפור ברמות האוריינות בקרב ילדים בגיל בית הספר (מקור)

עם זאת, בשנים האחרונות ניכרת ירידה בשימוש בספריות הציבוריות. ניתוח מגמות במדינת קונטיקט מראה ירידה מתמשכת במספר הביקורים לנפש החל משנת 2010. מגפת הקורונה בשנת 2020 האיצה את הירידה, אך הנתונים מצביעים על כך שהמגמה החלה עוד קודם לכן. אמנם נרשמה התאוששות חלקית לאחר הקורונה, אך היא אינה מספיקה כדי לשחזר את הרלוונטיות שאפיינה את הספריות בעבר.



מתוך ההבנה לחשיבותן החברתית, התרבותית והחינוכית של הספריות הציבוריות, ומתוך הדאגה להמשך תפקידן כמוקדים קהילתיים, נולד הרעיון לפרויקט זה. כחלק מהמוטיבציה האישית, אנחנו, כסטודנטים המשתמשים בשירותי הספרייה באופן תדיר, החלטנו לחקור את הגורמים המשפיעים על המעורבות החברתית בספריות. מטרתנו הייתה לסווג ספריות לפי מדד מעורבות חברתית, אשר יאפשר זיהוי ספריות בעלות מעורבות גבוהה לעומת ספריות בעלות מעורבות נמוכה. באמצעות שילוב של מודלים מפוקחים ולא מפוקחים, ניסינו להבין את המאפיינים של ספריות מצליחות בהשוואה לאחרות.

Section 2 - Dataset and Features

בחרנו לעבוד על נתונים מתוך קובץ csv מאתר data.gov של ארצות הברית, המכיל מידע מקיף על ספריות ציבוריות במדינת קונטיקט. הקובץ כולל נתונים היסטוריים על 208 ספריות שונות, שנאספו בין השנים 1996 ל-2023.

המידע בדאטה מתחלק לשלוש קטגוריות עיקריות:

- נתונים דמוגרפיים - דירוג כלכלי של האזור וכמות האוכלוסייה שהספרייה משרתת.
- נתונים חברתיים - כמות הביקורים בספרייה, כמות המשאילים הרשומים, מספר ההשאלות ומספר הצפיות בתוכניות סינכרוניות ומוקלטות.
- נתונים כלכליים - הכנסות תפעוליות, הקצאות מיסים מהעירייה, הוצאות תפעוליות והוצאות על משכורות וחומרים.

השלב הראשון בעבודתנו היה לטפל בערכים החסרים בדאטה. תחילה, סיננו ספריות שלא הכילו מידע עבור כל השנים, מה שהוביל לירידה במספר הספריות מ-208 ל-170. לאחר מכן, זיהינו באמצעות ניתוח סטטיסטי

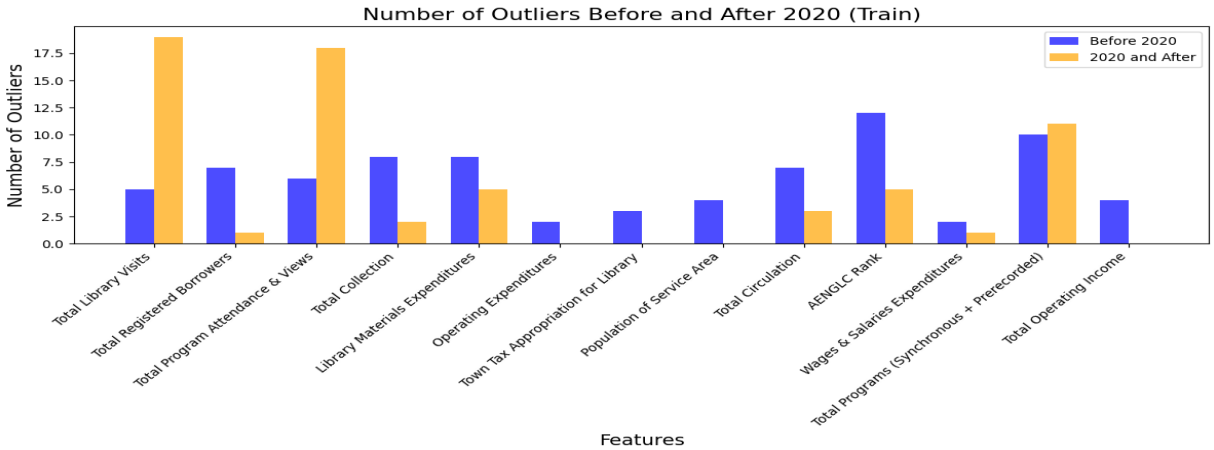
ובחינה ויזואלית כי עשר ספריות נוספות אחראיות לרוב הערכים החסרים. החלטנו להסיר גם אותן, וכך נותרנו עם 160 ספריות, המייצגות כ-85% מהדאטה המקורי.

בהמשך, השלמנו את הערכים החסרים בהתאם לאופי הנתונים של כל עמודה. לדוגמה, בעמודת סך ההכנסות התפעוליות זיהינו מגמת עלייה ברורה לאורך השנים (ראו גרף מצורף). בהסתמך על מגמה זו, החלטנו להשלים את הערכים החסרים בעמודה זו באמצעות רגרסיה לינארית. גישה זו נבחרה בשל התאמתה לנתונים המראים דפוס לינארי לאורך זמן.

כמו כן, בתהליך העבודה השתמשנו במגוון שיטות להשלמת נתונים חסרים, כגון אינטרפולציה לינארית, השלמה לאחור, השלמה לפנים, שימוש בממוצע פשוט ועוד. השימוש בכלי השלמה אלה התבצע תוך התאמה לכל משתנה ולפי אופי הנתונים והמגמות שזיהינו.

בנוסף, הקדשנו תשומת לב לזיהוי וטיפול ברעשים וערכים חריגים. מטרתנו הייתה להבטיח שמידע חריג לא ישפיע באופן בלתי פרופורציונלי על תוצאות המודלים שלנו. במסגרת זו, נביא לדוגמא את בחינת נתונים חריגים לשתי תקופות: לפני שנת 2020 ולאחריה.

בגרף המצורף ניתן לראות את מספר הערכים החריגים לפי 3 סטיות תקן בעמודות נבחרות בדאטה, כשהם מחולקים לפי שתי התקופות הנ"ל. הממצאים שלנו הראו כי הייתה עלייה משמעותית בערכים החריגים בנתונים הקשורים לביקורים בספריות והשתתפות בפעילויות במהלך שנת 2020 ולאחר מכן. העלייה מוסברת בעיקר בשל מגפת הקורונה, שגרמה לירידה חדה בפעילות הפיזית בספריות.



לאור זאת, קיבלנו החלטה לגבי הטיפול בערכים החריגים. נתוני הביקורים הפיזיים וההשתתפות בפעילויות לאחר שנת 2020 לא הוגדרו כערכים חריגים, מכיוון שהם שיקפו באופן נאמן את ההשפעות הישירות של מגפת הקורונה ומהווים נתוני אמת לתקופה ייחודית זו. לעומת זאת, בעמודות אחרות שבהן הערכים החריגים לא נבעו ממגמות ברורות, נקטנו בגישות מובנות לטיפול, כגון הסרה או תיקון מבוסס ממוצע וסטיות תקן.

Section 3 - Methodology

המתודולוגיה שבה השתמשנו התמקדה בחלוקה מושכלת של שלבי העבודה ובשימוש בשתי מחברות נפרדות. מחברת אחת הוקדשה ללמידה המפוקחת (Supervised Learning), ואילו השנייה הוקדשה ללמידה הלא מפוקחת (Unsupervised Learning). חלוקה זו אפשרה לנו לשמור על סדר, לאפשר בחינה מדויקת של כל שיטה, ולהימנע מתלות בין התהליכים.

בגישה של למידה לא מפוקחת, התמקדנו בזיהוי דפוסים ותבניות בדאטה ובסיווג ספריות. משמעות הדבר היא שכל עיבוד נתונים יכול להתבצע על הדאטה במלואו, ללא חלוקה לסטי אימון ובדיקה, מתוך הבנה שאין חשש לזליגת מידע (Data Leakage). בנוסף, בתהליך הלמידה לא מפוקחת, כמות הדאטה שנשמרה הייתה מקסימלית, מה שהגביר את איכות התוצאות ואת היכולת לייצר חלוקה משמעותית לקבוצות.

לעומת זאת, בלמידה מפוקחת, המטרה הייתה לבנות מודלים המסוגלים לסווג. לצורך כך, חילקנו את הדאטה לסט אימון (Training Set) ולסט בדיקה (Test Set) כבר בתחילת התהליך. כל שלב של עיבוד נתונים או השלמת ערכים חסרים בוצע בנפרד על כל אחד מהסטים, על מנת למנוע זליגה של מידע בין הסטים ולשמור על אמינות המודלים שנבנו.

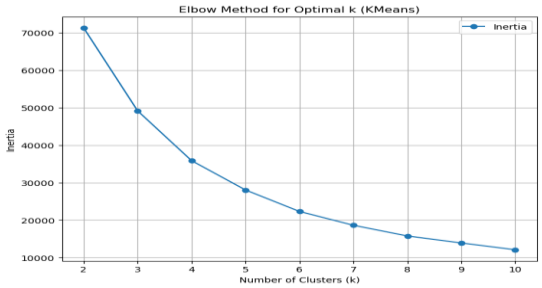
למידה לא מפוקחת

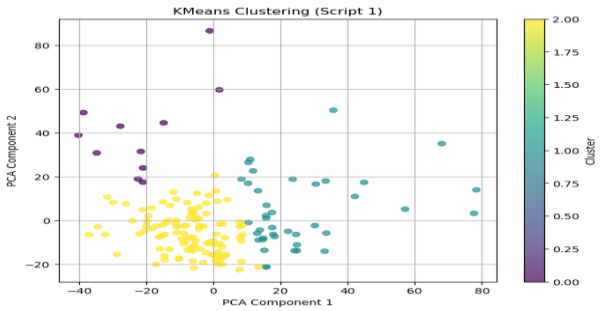
המטרה המרכזית של הלמידה הלא מפוקחת הייתה לחלק את הספריות הציבוריות לקבוצות (Clusters) על סמך דפוסים ותכונות משותפות. החלוקה נועדה לאפשר זיהוי של מאפיינים ייחודיים לספריות מצליחות מול ספריות אחרות.
במהלך הקלסטרינג, קיבצנו כל ספריה כיחידה נפרדת לאורך הזמן. לשם כך, יצרנו חמישה פיצ'רים סטטיסטיים חדשים עבור כל עמודה: ממוצע, חציון, ערכים מינימליים ומקסימליים, וסטיית תקן. גישה זו איפשרה לנו לשמר את האופי הדינמי של הנתונים ולספק ייצוג עשיר לכל ספריה בלי לאבד מידע רב אם היינו לוקחים רק את הממוצע. על מנת לצמצם את מורכבות הנתונים ולהקל על המודלים בתהליך הקלסטרינג, השתמשנו ב-PCA. האלגוריתם שמר על 78% מהשונות בנתונים, תוך הפחתת מספר המימדים ושמירה על המידע הקריטי. בנוסף, השימוש ב-PCA סייע ביצירת ויזואליזציות ברורות של הקלסטרים.

1. מודל **K-Means**

לאחר עיבוד הנתונים, השתמשנו במודל K-Means, שמטרתו לחלק את הנתונים למספר מוגדר מראש של קבוצות. כל קבוצה מתוארת כמעגל או כדור סביב מרכז מסוים (Centroid). במהלך תהליך האיטרציות של המודל, ה-Centroids מתעדכנים כל הזמן עד שהמודל מגיע לאיזון שבו הנקודות אינן משנות את שיוכן לקלאסטרים.

לבחירת מספר הקלסטרים האופטימלי, השתמשנו בשיטת Elbow, שמטרתה לזהות את הנקודה שבה תוספת קלסטרים אינה משפרת משמעותית את דיוק החלוקה. שיטה זו מתבססת על ניתוח השגיאות הכוללות (Inertia) כאשר הנקודה האופטימלית היא זו שבה ערך השגיאות מתחיל להתמתן.



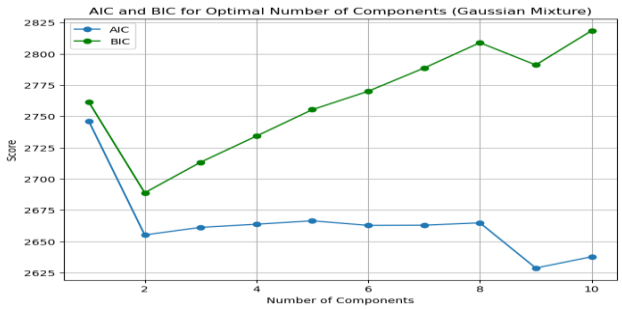


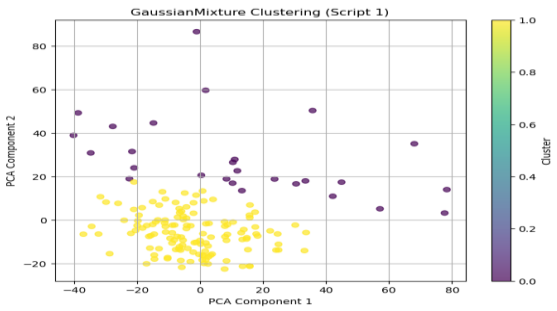
2. מודל **Gaussian Mixture**

מודל זה מניח שהנתונים מתפלגים למספר התפלגויות נורמליות, ומחלק אותם לקבוצות בהתאם. המודל מאפשר חפיפה בין הקבוצות ומייצג את הנתונים כהרכב של התפלגויות גאוסיאניות. לבחירת מספר הקבוצות האופטימלי, נעזרנו במדדי **AIC** ו-**BIC**:

- **AIC** בוחן את איכות התאמת המודל לנתונים, כאשר ערכים נמוכים יותר מעידים על התאמה טובה.
- **BIC** מעניש על מורכבות המודל, על מנת למנוע התאמת יתר.

הגרף של AIC ו-BIC סייע לנו לזהות את מספר הרכיבים האופטימלי בנקודת המינימום שלהם.

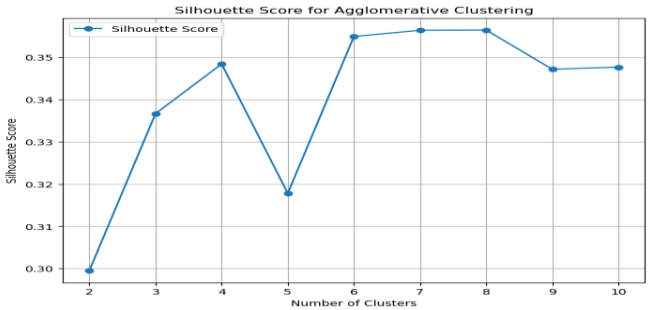


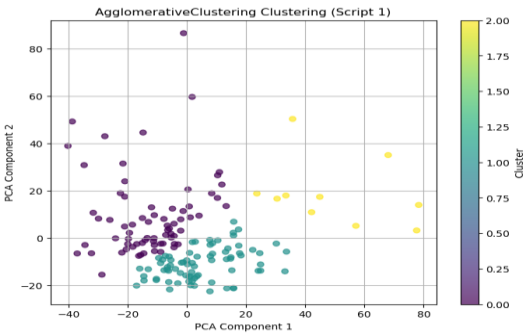


3. מודל **Agglomerative Clustering**

מודל היררכי שמתחיל בחיבור של כל נקודה לקבוצה עצמאית, וממשיך באיחוד הקבוצות הקרובות ביותר על פי מרחק מוגדר, עד שמגיעים למספר הקבוצות הרצוי.

המודל מתאים במיוחד למצבים שבהם יש חשיבות להבנה היררכית של הקשרים בין נקודות. כדי לבחור את מספר הקלאסטרים האופטימלי, השתמשנו במדד **Silhouette Score**. מדד זה מעריך את איכות הקלסטרים על ידי חישוב כמה טוב כל נקודה מותאמת לקבוצה שלה בהשוואה לקבוצות אחרות. ערך גבוה יותר וקרוב יותר ל1 מעיד על חלוקה טובה יותר של הקבוצות ואילו ערך קרוב יותר ל1- מעיד על חלוקה גרועה.





לבסוף כדי לבחור את המודל הטוב ביותר, יצרנו טבלה המציגה את מדד ה-**Silhouette Score** עבור שלושת המודלים שנבחנו. מודל **GaussianMixture** משיג את הציון הגבוה ביותר, מה שהוביל לבחירתו כמועדף לניתוח הקבוצות.

	Silhouette Score
KMeans	0.430529
GaussianMixture	0.495432
AgglomerativeClustering	0.336652

Metric	0	1
AENGLC Rank	36.461640	94.937419
Population of Service Area	20681.462963	21691.721101
Library Visits Per Capita Served	9.018675	5.062477
Percent of Residents with Library Cards	0.579348	0.497394
Circulation Per Capita Served	12.621043	7.541689
Total Program Attendance & Views Per Capita Se...	1.077724	0.434007
Collection Per Capita Served	8.854630	4.996970
Operating Income Per Capita	89.591032	37.874700
Tax Appropriation Per Capita Served	56.895040	33.051667
Operating Expenditures Per Capita	84.451455	37.572059
Wages & Salaries Expenditures Per Capita	47.195150	22.692152
Library Materials Expenditures Per Capita	8.503851	4.403882
Profit Per Capita	5.864093	6.268699

המודל חילק את הספריות לשתי קבוצות עיקריות, עם הבדלים ברורים במדדים הסוציו-אקונומיים ובמדדי המעורבות החברתית.

הממצאים מצביעים על כך שקלסטר 0 מתאפיין במעורבות חברתית גבוהה יותר: אחוז גבוה של משתמשים רשומים, יותר ביקורים לנפש, יותר השאלות, וצפיות בתוכניות. הספריות בקלסטר זה ממוקמות באזורים סוציו-אקונומיים גבוהים יותר ומשקיעות יותר במשאבים ובתשתיות, מה שמסביר את הפעילות החברתית הערה ואת המשאבים הכלכליים הגבוהים. לעומת זאת, קלסטר 1 מתאפיין במעורבות חברתית נמוכה יותר, עם פחות ביקורים והשאלות לנפש. הספריות בקבוצה זו ממוקמות באזורים עם דירוג סוציו-אקונומי נמוך יותר, מה שעשוי להשפיע על היכולת של הקהילה להתחבר לשירותי הספרייה.

<u>למידה מפוקחת</u>

עבור משימת הסיווג שלנו, פיתחנו עמודת מטרה שנועדה למדוד את רמת המעורבות הקהילתית בכל ספרייה בהתבסס על שני פיצ'רים עיקריים:

1. **Library Visits Per Capita Served** - כמות הביקורים לנפש.
2. **Total Registered Borrowers per Capita** - כמות המשאילים הרשומים לנפש.

המדד, כפי שמוצג במשוואה למטה, חושב על בסיס משוקלל, כך שנתנו דגש גבוה יותר (70%) לכמות הביקורים, מתוך הבנה שהיא מייצגת בצורה מדויקת יותר את הקשר בין הקהילה לספרייה, בעוד כמות המשאילים (30%) שימשה כתוספת המחזקת את הערכת המעורבות:

$$Performance\ Score = (Library\ Visits\ Per\ Capita\ Served \times 0.7) + (Total\ Registered\ Borrowers\ per\ Capita \times 0.3)$$

המדד נועד לשקף את חיבור הקהילה לספרייה, הן מבחינת שימוש פעיל בשירותים והן מבחינת הרשמה.

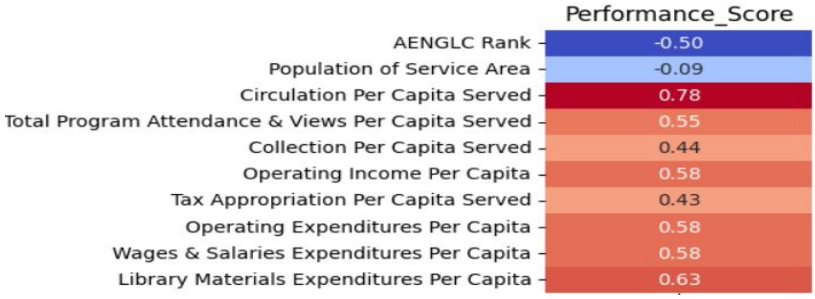
תוצאות לאחר אפיון על הנתונים:

```
Performance_Score Statistics:
count     3581.000000
mean         4.190322
std          2.702202
min          0.096080
25%          2.315332
50%          3.635880
75%          5.416210
90%          7.616313
max         18.232688
```

ולפי התפלגות זו חילקנו ל4 קטגוריות:

- **Excellent**: העשירון העליון, ערך גבוה מ7.7 המייצגים מעורבות קהילתית גבוהה במיוחד.
- **Good**: בין החציון לעשירון העליון, עם מעורבות חיובית.
- **Average**: בין הרבעון הראשון לחציון, מעורבות ממוצעת.
- **Failing**: האחוזונים התחתונים בין האפס לרבעון הראשון, דורשות שיפור משמעותי.

יצרנו מפת חום המציגה את הקורלציות בין עמודת המטרה (Performance_Score) לפיצ'רים האחרים בדאטה בעזרתם המודלים יסווגו.



עבור משימת הסיווג שלנו בחרנו
להשתמש במודלי סיווג שונים ובחנו את ביצועיהם על ידי שימוש ב-5 קרוסים. עשינו זאת כדי להבטיח שהתוצאות אינן אקראיות ומשקפות באופן עקבי את ביצועי המודלים, גם בהשוואה לסט הבדיקה.

לכל המודלים השתמשנו באותן מטריקות הערכה: **F1-Score**, **Recall**, **Precision**, ו-**Accuracy**.

בחרנו להעריך חמישה מודלים מרכזיים:

1. **Decision Tree**
2. **Random Forest**
3. **XGBoost**
4. **Gradient Boosting**
5. **K-Nearest Neighbors (KNN)**

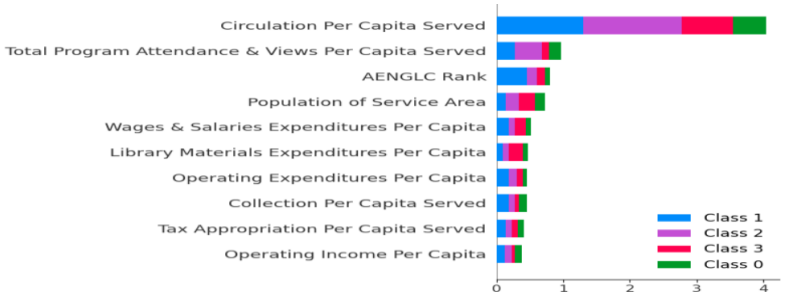
בנוסף, במהלך העבודה בחנו גם מודלים אחרים כמו **Logistic Regression**, **SVM**, ו- **Neural Networks**. עם זאת, ביצועיהם היו נמוכים מאוד ביחס לנתונים שלנו, מה שהוביל להחלטה להתמקד בעיקר במודלים מבוססי עצים (Tree-Based Models). מודלים אלו התגלו כיעילים במיוחד בהתמודדות עם המורכבות והדפוסים של הדאטה, והציגו ביצועים טובים בהרבה בהשוואה לאלטרנטיבות.

Sorted Results by Test F1 Score:

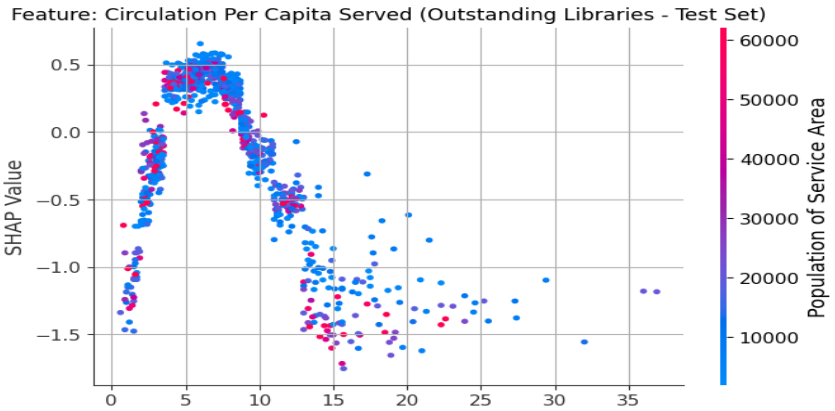
	Model	Mean CV Accuracy	Test Accuracy	Test F1 Score
3	Gradient Boosting	0.753697	0.750000	0.751272
2	XGBoost	0.736387	0.748884	0.750387
4	KNN	0.616029	0.631696	0.632773
1	Random Forest	0.640322	0.620536	0.629219
0	Decision Tree	0.602901	0.584821	0.598673

תוצאות המודלים מצביעות על כך שמודלי ה-Boosting, השיגו את הביצועים הטובים ביותר במשימת הסיווג שלנו. היתרון שלהם טמון בכך שהם בונים עצים בצורה סדרתית, כאשר כל עץ מתקן את השגיאות של העץ הקודם, מה שמשפר את הדיוק. לעומתם, Random Forest משתמש בממוצע של מספר עצים אקראיים, בעוד Decision Tree בונה עץ יחיד ופשוט יותר. גישה זו מסבירה את היתרון המשמעותי של Boosting בהתמודדות עם דאטה מורכב.

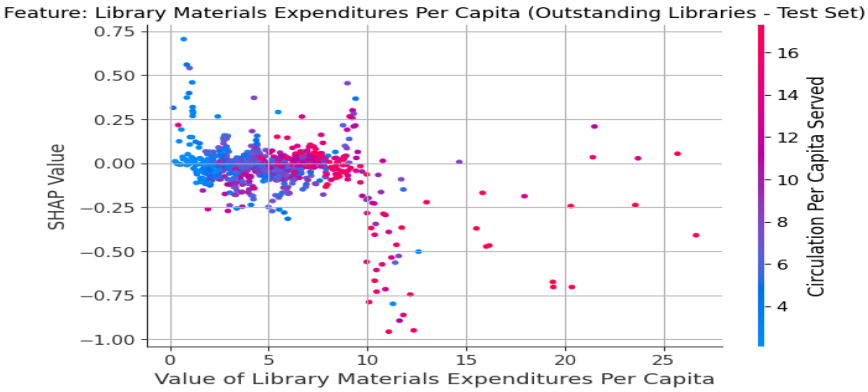
לאחר הסיווג, השתמשנו בניתוח SHAP כדי להבין את התרומה של כל פיצ'ר לתוצאות המודל ולשפר את פרשנותו. הגרף מתאר את מידת ההשפעה של הפיצרים לפי הסדר.



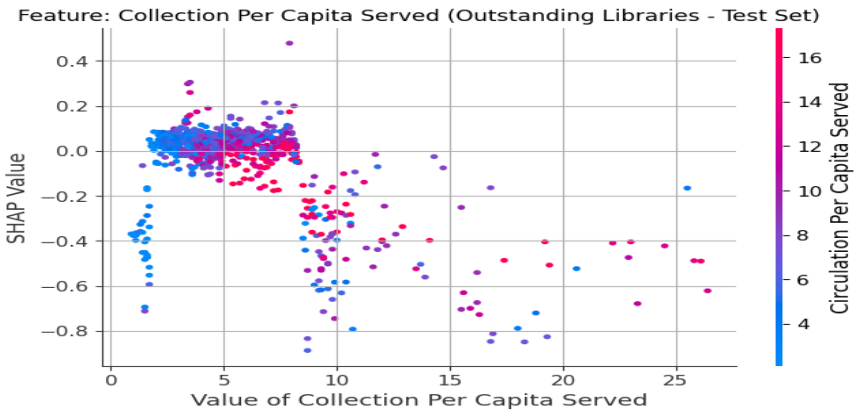
מכאן אנחנו רצינו להתמקד בספריות מצטיינות ומה היו הפיצ'רים שעזרו להם להגיע למעמד זה.



ניתן לראות שערך אידיאלי של השאלות לנפש הוא בין 5 ל-10, שבו ההשפעה החיובית על הסיווג היא הגבוהה ביותר. מעבר לטווח זה, ההשפעה יורדת, מה שמעיד על כך שמספר גבוה מדי של השאלות עשוי לשקף שחיקה או כמות מוגבלת מאוד של מעורבים, ולכן רצוי להתמקד בעידוד השאלות בטווח האופטימלי לשימור מעורבות קהילה גדולה וחזקה.



ניתן לראות כי עד לרמה מסוימת של הוצאות (בערך 10 לנפש), ישנה השפעה חיובית על הסיווג. מעבר לרמה זו, ההשפעה מתייצבת או אף פוחתת, מה שמעיד על כך שהשקעה עודפת בציוד אינה תורמת משמעותית למעורבות הקהילתית וייתכן שכדאי לנתב את התקציב לאפיקים אחרים שמגבירים מעורבות.



נראה כי מעבר לרמה מסוימת של אוספים לנפש (בסביבות 5–10), אין שיפור משמעותי במעורבות הקהילתית. הדבר מצביע על כך שמגוון וכמות ספרים מאוזנים עשויים להספיק לצורכי הקהילה, בעוד שהשקעה עודפת באוספים עלולה לייצר עלויות תחזוקה מיותרות ולא לתרום משמעותית למעמד הספרייה.

סיכום

העבודה על הפרויקט הייתה עבורנו חוויה מעשירה ומלמדת. אנו מודעים לכך שספריות ציבוריות אינן גופים רווחיים, ולעיתים נתפסות כנטל כלכלי מצד גופים עירוניים וממשלתיים. עם זאת, אנו מקווים כי הצלחנו לזהות נקודות מפתח שעשויות, מצד אחד, להפחית במעט את העול הכלכלי, ומצד שני, לשפר את המעורבות הקהילתית בכל ספרייה. הממצאים שהצגנו בפרויקט זה יכולים, לדעתנו, לשמש בסיס לשיפור השירותים ולהעלאת רמת המעורבות הקהילתית בספריות הציבוריות.

היינו שמחים לראות את ההמלצות שלנו מיושמות בעתיד לטובת הקהילות שהספריות משרתות. במהלך העבודה על הפרויקט, יצרנו קשר עם מהנדסת הנתונים שחתומה על הדאטה שבחרנו מהאינטרנט, במטרה לקבל הבהרות על הנתונים. לצערנו, היא הודיעה כי היא עוזבת את תפקידה ולא תוכל לספק מענה לשאלותינו, מה שהוביל אותנו לעבוד בצורה עצמאית לחלוטין.

עבדנו על הפרויקט בשיתוף פעולה מלא, כאשר כל אחד מאיתנו תרם את חלקו: רותם התמקד בלמידה הלא מפוקחת, ואילו ברק התרכז בלמידה המפוקחת. חלוקה זו אפשרה לנו לנצל את הזמן ביעילות תוך שמירה על שיתוף פעולה וידע בכל המהלכים שביצענו לאורך הפרויקט.

תודה רבה על ההזדמנות לקחת חלק בפרויקט חשוב ומשמעותי זה.

לינק לגיטהאב של העבודה:https://github.com/ROTEM0805/CT-libraries-Analysis