הצעת מחקר – ספריות ציבוריות בארצות הברית

רותם בן עטר 318299914

ברק וירצברגר 316597541

בחירת מערך נתונים

נושא : ספריות ציבוריות בארצות הברית.

מקור הנתונים : https://catalog.data.gov/dataset/public-libraries-b1aaf

תיאור קצר : המידע במאגר מאפשר לבצע ניתוחים מעמיקים לזיהוי דפוסי ביצוע של הספריות, סיווג ספריות לפי רמת הפעילות, זיהוי ספריות חריגות, וניתוח מגמות לאורך זמן.

מוטיבציה

בעיות מרכזיות

1. הערכת ביצועים- קשה להעריך בצורה אחידה את רמת הביצועים של ספריות, בהתחשב בהבדלים בגודל האוכלוסייה, התקציבים והפעילות של כל ספרייה.
2. זיהוי ספריות חריגות- ספריות מסוימות עשויות להיות חריגות מבחינת ההוצאות או הפעילות שלהן, מה שמקשה על ניהול משאבים אפקטיבי.

מדוע הבעיות משמעותיות ואיזה ערך ניתן לתת להמשך?

1. שיפור מדדי הביצועים של הספריות יאפשר תיעדוף משאבים לשיפור תפעול הספריות.
2. זיהוי ספריות מצטיינות או ספריות שדורשות תשומת לב יוכל לשפר את איכות השירותים לציבור.

שיטה

למידה לא מפוקחת (Unsupervised Learning)

1. **K-Means Clustering -** שימוש באלגוריתם לחלוקת הספריות לאשכולות על פי משתנים כמו ביקורים לנפש, השאלות, ההוצאות, וגודל האוכלוסייה במטרה לזהות קבוצות בעלות מאפיינים דומים.
2. **Isolation Forest -** זיהוי חריגות על ידי בידוד ספריות עם פעילות או תקציב חריגים.
3. **DBSCAN -** חלוקה של ספריות על בסיס מבנה צפיפות הנתונים, כולל זיהוי ספריות שאינן מתאימות לאף קבוצה ברורה.

למידה מונחית (Supervised Learning)

1. **XGBoost** - תחזית משתנים רציפים כמו ביקורים או הכנסות, או סיווג הספריות לפי קריטריונים.
2. **Random Forest –** סיווג ספריות (מצוינות, ממוצע, שיפור נדרש).
3. **Decision Tree –** מודל נוסף לסיווג.

ניסויים מתוכננים

1. חלוקת כל המודלים ל- Train&Test
2. **Isolation Forest** - Percentage of Anomalies Detected
3. **Clustering** – Silhouette Score
4. **Decision Tree** - Accuracy, Precision, Recall, F1 Score
5. **Random Forest** - Accuracy, AUC-ROC
6. **DBSCAN -** Noise Ratio ,Silhouette Score