נושאים מתקדמים בלמידת מכונה

ד"ר חן חג'ג'

מגישים:

עדן כהן

מרינה נזלסקי

מבוא

הפרויקט שלנו מתמקד בניתוח נתוני קווי נסיעות האוטובוסים במדינת ישראל בשנת 2024.

מערכת התחבורה הציבורית בישראל מורכבת מאלפי קווי אוטובוס הפועלים באזורים גיאוגרפיים שונים עם מאפייני מסלול ותפעול מגוונים. הבנה טובה יותר של נתונים אלה יכולה לשפר את התכנון והניהול של מערכות התחבורה הציבורית במדינה.

מטרות הפרויקט:

1. סיווג קווי אוטובוס למטרופולינים שונים באמצעות למידה מפוקחת כדי לשפר את ניהול התחבורה בכל אזור.
2. זיהוי קבוצות של קווי אוטובוס עם מאפיינים דומים באמצעות למידה לא מפוקחת, כדי לייעל את התפעול ולהפחית בזבוז משאבים.

מלבד טכניקות ואלגוריתמים בלמידת מכונה, ניתן להסיק מסקנות על נסיעות קווי האוטובוס בעזרת שיטות קלאסיות כמו סקרים בקרב הנוסעים ונהגי האוטובוס או תצפיות בפועל.

מערך הנתונים

מקור הנתונים: Data.gov.il

מערך הנתונים: נסועה בקווי אוטובוס שנת 2024

מספר העמודות (תכונות): 53

מספר השורות (רשומות): 10679

שם הקו

מטרופולין

אורך המסלול

מספר נוסעים שבועי

מספר תחנות במסלול

סוג האוטובוס

עלות תפעול לנוסע

סוג שירות

מספק ק"מ שבועי

סוג מסלול

0	RouteID	10678 non-null	int64
1	RouteName	10678 non-null	int64
2	RouteDirection	10678 non-null	int64
3	AgencyName	10678 non-null	object
4	ClusterName	10678 non-null	object
5	Metropolin	10678 non-null	object
6	OriginCityName	10678 non-null	object
7	DestinationCityName	10678 non-null	object
8	RouteType	10678 non-null	object
9	ServiceType	10678 non-null	object
10	RouteParticular	10678 non-null	object
11	BusType	10678 non-null	object
12	BusSize	10678 non-null	object
13	NumOfAlternatives	10678 non-null	int64
14	RouteLength	10678 non-null	float64
15	WeeklyKM	10678 non-null	float64
16	AVGPassengersPerWeek	10667 non-null	float64
17	StationsInRoute	10678 non-null	int64
18	OperationSince	10678 non-null	object
19	UniqueStations	3501 non-null	float64
20	UniqueLocations	718 non-null	object
21	AverageSpeed	10675 non-null	float64
22	AverageTripDuration	10675 non-null	float64
23	OperatingCostPerPassenger	10664 non-null	float64
24	DailyRides(Tuesday)	10678 non-null	int64
25	WeekyRides	10678 non-null	int64
26	DailyPassengers	10667 non-null	float64
27	WeeklyPassengers	10667 non-null	float64
28	AVGCommutersPerRide(Weekly)	10667 non-null	float64
29	WorkDay - 00:00-03:59	998 non-null	float64
30	WorkDay - 04:00-05:59	4161 non-null	float64
31	WorkDay - 06:00-08:59	7890 non-null	float64
32	WorkDay - 09:00-11:59	6620 non-null	float64
33	WorkDay - 12:00-14:59	7666 non-null	float64
34	WorkDay - 15:00-18:59	7577 non-null	float64
35	WorkDay - 19:00-23:59	6397 non-null	float64
36	Friday - 00:00-03:59	38 non-null	float64
37	Friday - 04:00-05:59	2935 non-null	float64
38	Friday - 06:00-08:59	6588 non-null	float64
39	Friday - 09:00-11:59	6354 non-null	float64
40	Friday - 12:00-14:59	6649 non-null	float64
41	Friday - 15:00-18:59	2766 non-null	float64
42	Friday - 19:00-23:59	209 non-null	float64
43	Saturday - 00:00-03:59	1170 non-null	float64
44	Saturday - 04:00-05:59	162 non-null	float64
45	Saturday - 06:00-08:59	418 non-null	float64
46	Saturday - 09:00-11:59	418 non-null	float64
47	Saturday - 12:00-14:59	496 non-null	float64
48	Saturday - 15:00-18:59	1430 non-null	float64
49	Saturday - 19:00-23:59	5127 non-null	float64
50	MaxRidership	10678 non-null	object
51	year	10678 non-null	int64
52	Q	10678 non-null	int64

מתודולוגיה- למידה מפוקחת

1. ניקוי הנתונים והשלמת נתונים חסרים
2. הנדסת תכנות (Feature Engineering)
3. קידוד (Encoding)
4. סינון תכונות בעזרת Feature importance ובעזרת Pearson Corolation.
5. התאמת סקלארים בהתאם לסוג המודל
6. בחינת מספר מודלי Classification:

Random Forest

מבוסס על עצי החלטה מרובים, התוצאה הסופית היא החלטת רוב. עמיד בפני Overfitting ומבצע סיווג מהיר.

Gradient Boosting

מייצר עץ החלטה אחד בכל פעם, מתקנים טעויות שנעשו בעצים קודמים. משפר תוצאות משמעותית על ידי תיקון טעויות.

Logistic Regression

מודל לינארי הממיר תוצאה להסתברות עם פונקציית סיגמואיד. מתאים לפירוש קל של תוצאות.

SVM

מתמקד ביצירת גבול החלטה אופטימלי בין קטגוריות. מתאים לבעיות עם גבול החלטה ברור ומימד גבוה.

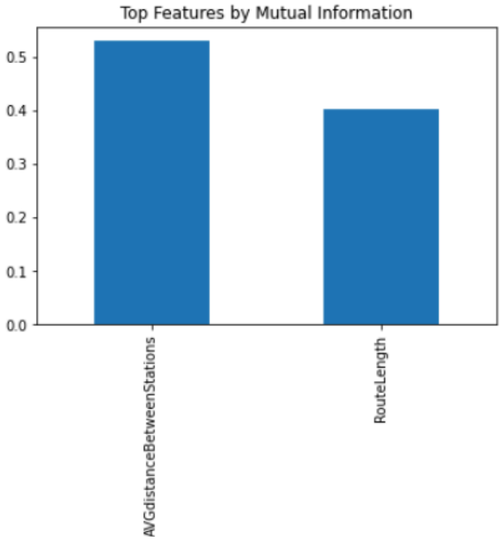
KNN

מבוסס על חישוב מרחקים ומסווג לפי הקטגוריה של השכנים הקרובים ביותר. גמיש ומתאים לבעיות סיווג מרובות קטגוריות.

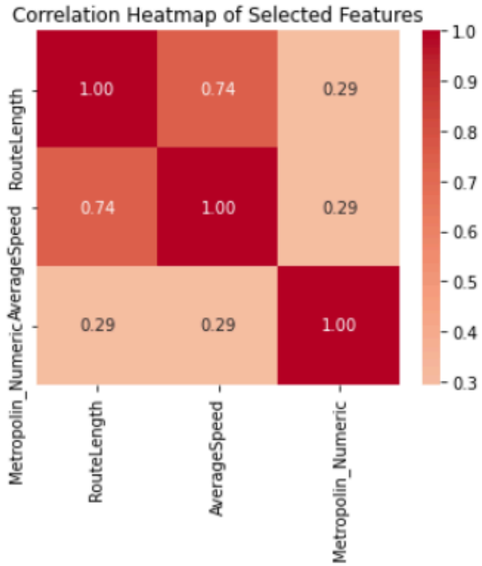
7. בחירת המודל לפי מדדי הערכה: Accuracy, Precision, Recall, F1 Score.

למידה מפוקחת- ניסויים

```
Top 10 features based on mutual information:
AVGdistanceBetweenStations    0.529447
RouteLength                   0.402145
dtype: float64
```




```
Features with correlation higher than 0.2 with 'Metropolin':
['RouteLength', 'AverageSpeed']
```



ניסויים בבחירת תכונות

- סיננו תכונות דומות לעמודת ה"מטרופולין" לפי Cramer's V > 0.5, מדד המעריך קשר בין משתנים קטגוריאליים.
- בחרנו תכונות עם ערך Mutual Information מעל 0.4, המצביע על תלות בין משתנים.
- בחרנו תכונות עם Pearson Correlation מעל 0.2 (קשר חיובי) או מתחת ל- 0.2- (קשר שלילי) עם Metropolin_Numeric.

ניסויים במודל

- חלוקת לסט אימון וסט בדיקה ביחס של 80:20 באמצעות הפונקציה train_test_split.
- נעשה שימוש ב-Cross-Validation עם 5 קיפולים (fold=5), המבטיח הערכה יציבה על פני קיפולים שונים של הנתונים.

למידה מפוקחת- תוצאות

המודל שהשיג את הביצועים הטובים ביותר היה **RandomForest** עם דיוק ממוצע של 71.8% לפי הצלבה (Cross-Validation)

המודל הציג את התוצאות הבאות:

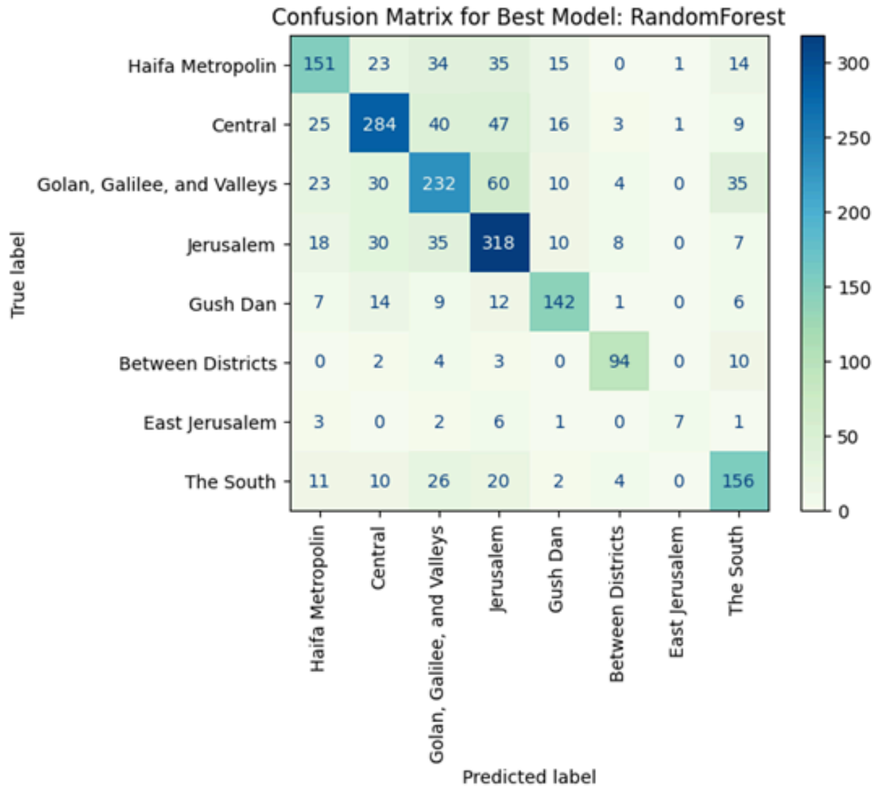
- **דיוק (Presicion):** 80% ב"גוש דן" ו-74% ב" בין מחוזי".
- **רגישות (Recall):** גבוהה ב"בין מחוזי" (85%), אך נמוכה מאוד ב"מזרח ירושלים" (30%).
- **מדד F1:** הראה 68% ב"ירושלים", מה שמעיד על איזון סביר בין דיוק ורגישות.
- **דיוק כולל(Accuracy):** 67% על סט הבדיקה, עם הבדל בין המטרופולינים.

```
Model: RandomForest
              precision    recall  f1-score   support

           1       0.66      0.61      0.63       254
           2       0.67      0.67      0.67       413
           3       0.61      0.61      0.61       413
           4       0.68      0.71      0.70       463
           5       0.75      0.73      0.74       235
           6       0.78      0.77      0.78       102
           7       0.83      0.67      0.74        15
           8       0.68      0.70      0.69       232

    accuracy                           0.68      2127
   macro avg       0.71      0.68      0.69      2127
weighted avg       0.68      0.68      0.68      2127

Cross-Validation Accuracy: 0.721
```



השוואה בין מודלים ומסקנות

KNN
- Accuracy : 41%
- Cross -Validation Accuracy : 39.3%

SVM
- Accuracy: 34%
- Cross-Validation Accuracy 30.8%.

Logistic Regression
- Accuracy: 30%
- Cross-Validation Accuracy 28.1%.

Gradient Boosting
- Accuracy: 45%
- Cross-Validation Accuracy 42.9%.

Random Forest
- Accuracy: 67%
- Cross-Validation Accuracy : 72.2%.

מסקנות סופיות:

- Random Forest הוא המודל הטוב ביותר עבור הסיווג לפי המטרופולין מכל המודלים שנבחנו, בזכות היכולות של עבודה עם מידע מורכב ולא לינארי, וצרכים של שיפור איכות המודלים.
- Logistic Regression ו-SVM אינם מתאימים, בגלל שהנחת מודלים לינאריים על בעיה לא לינארית.
- Gradient Boosting ו-KNN מציגים ביצועים בינוניים ויכולים להשתפר עם כיוונון פרמטרים טוב יותר, והרצת Hyperparameter Tuning.

מתודולוגיה- למידה לא מפוקחת

1. שימוש בנתונים המקודדים
2. סינון הנתונים למטרופולין מסוג "גוש דן".
3. סטנדרטיזציה.
4. הורדת ממדי הנתונים בעזרת PCA, סינון המשתנים החלשים והגדלת השונות המוסברת.
5. בחינת מודלי Clustering:

GMM

מודל זה מאפשר קבוץ של נתונים במבנים לא לינאריים ומסתמך על הנחת עבודה לפיה הנתונים מגיעים מתערובת של כמה התפלגויות גאוסיאניות.

בחירת מספר האשכולות האופטימלי לפי מדדי:

- AIC
- BIC
- Silhouette Score
- Elbow Method (log likelihood)

K-means

מודל זה מאפשר קיבוץ הנתונים לאשכולות והוא משתמש באלגוריתם של חישוב מרכזי האשכולות וממקסמת את המרחק בין המרכזים.

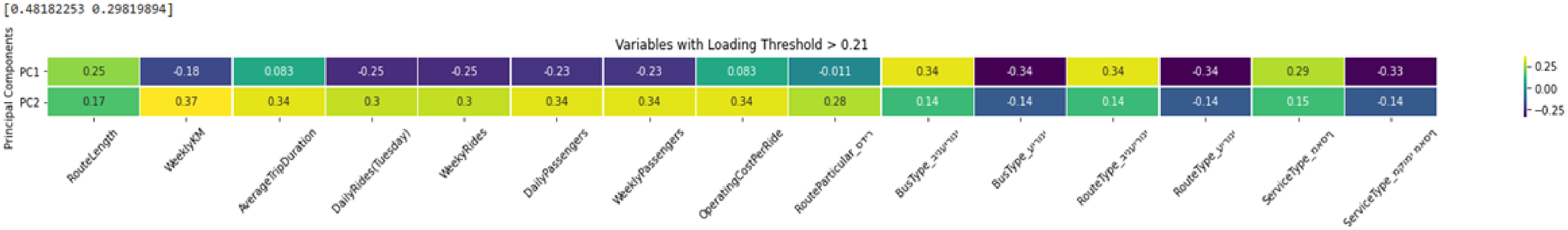
בחירת מספר האשכולות האופטימלים לפי מדדי:

- Silhouette Score
- Elbow Method

למידה בלתי מפוקחת - ניסויים

PCA

על מנת להגדיל את השונות המוסברת בין 2 רכיבים העיקריים מעל ל-70%, הגדרנו סף השפעה של 0.2, אשר מסנן את כל התכונות שמשפיעות פחות מ0.2 על הרכיבים.



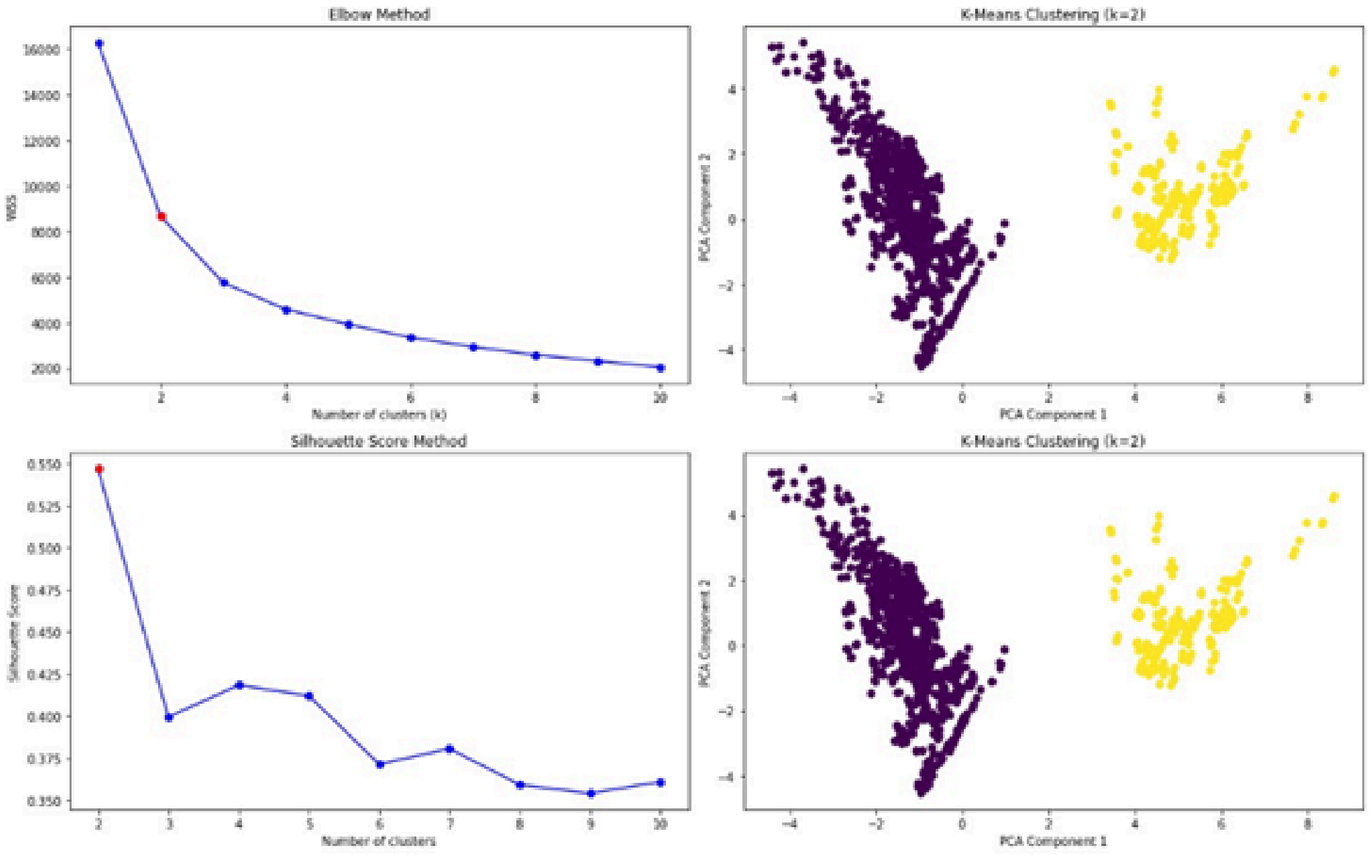
K-Means

כדי למצוא את מספר האשכולות האופטימלי, רשמנו אלגוריתם שמציג את מדדי ה Elbow, Silhouette בצורה גרפית, ובוחר באופן אוטומטי את התוצאה הטובה ביותר.

GMM

כדי למצוא את מספר האשכולות האופטימלי, רשמנו אלגוריתם שמציג את מדדי ה Elbow, Silhouette, AIC, BIC בצורה גרפית, ובוחר באופן אוטומטי את התוצאה הטובה ביותר.

למידה לא מפוקחת- תוצאות



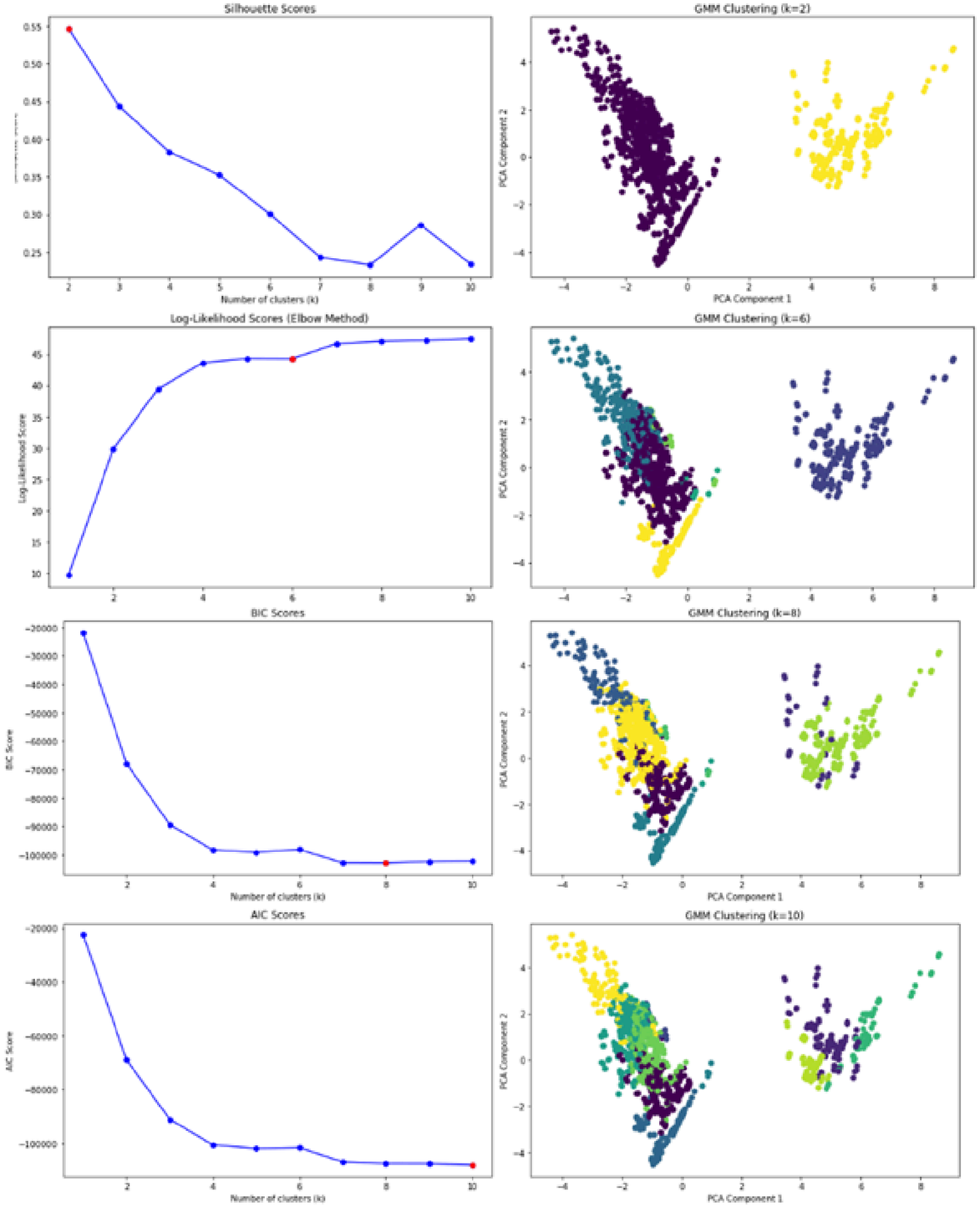
K-Means

גם בשיטת Elbow וגם בשיטת

Silheouette מספר האשכולות

האופטימלי שהתקבל הוא 2.

למידה לא מפוקחת- תוצאות



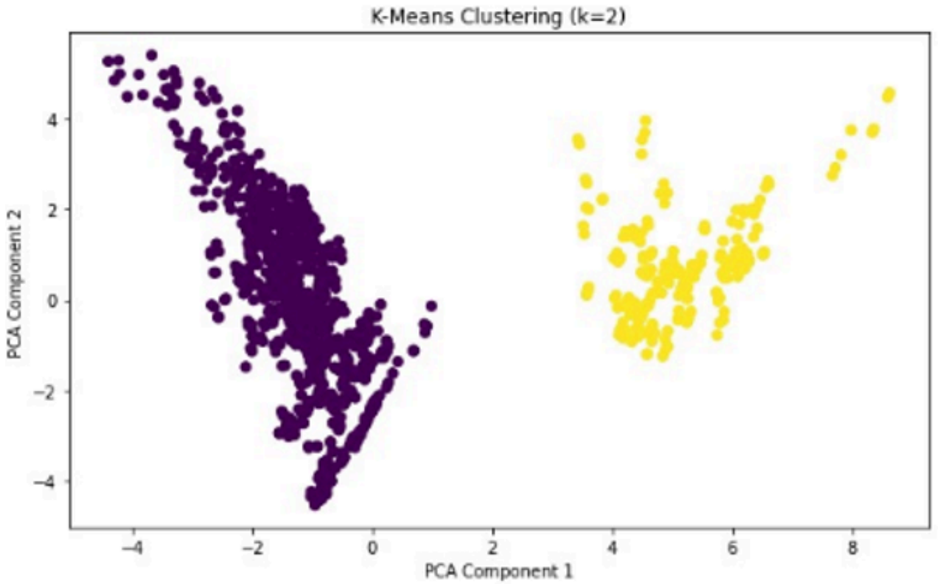
- Silhouette Score מראה תוצאה של 2 אשכולות עם הפרדה ברורה.
- Elbow Method (Log-Likelihood)- מצביע על 6 אשכולות עם חפיפה.
- AIC- התוצאה היא 10 אשכולות עם חפיפה.
- BIC- התוצאה היא 8 אשכולות עם חפיפה.

השוואה בין מודלים ומסקנות

בשני המודלים התקבלו תוצאות שמצביעות על חלוקה ל-2 אשכולות. לכן נבחר בK=2 מכיוון שחלוקת האשכולות ברורה וחד משמעית, מה שמקל על איפיון האשכולות.

איפיון הרכיבים:

- **PC1**- מורכב מהמשתנים חזקים כמו: Bus Type, Routh length בינעירוני, Routh type בינעירוני.
  ערכים גבוהים מצביעים על קווים ארוכים ובינעירוניים
- **PC2**-מורכב מהמשתנים חזקים כמו: Operation cost per ride, Weekly KM, Daily Passengers.
  ערכים גבוהים מצביעים על קווים עם קילומטרג' גבוהה שיכולים להיות יקרים יותר עם עלויות תפעול גבוהות.



איפיון האשכולות:

- **אשכול 1 (סגול)-** מצביע על קווים עירוניים בגוש דן (בדרך כלל אורך מסלול קצר יותר) וייתכן גם עלויות תפעול נמוכות יותר ותחבורה עם תדירות גבוהה יותר.
- **אשכול 2 (צהוב)**- מצביע על קווים בינעירוניים בגוש דן (כיוון שיש כאן אורך מסלול ארוך יותר וייתכן שעלוית התפעול גבוהות יותר.

שאלות?