



החוג למדעי המחשב

מוגש במסגרת הקורס: ביג דאטה

מטלה: פרויקט סוף קורס

מרצה: דר' איתי שרון

מגישים:

עומרי בר-חיים: 315696161

איתמר דרור: 315820134

סמסטר ב' - תשפ"ג

תאריך הגשה: 31.08.23

1. רקע:

בחרנו לערוך מחקר המבוסס על נתונים סטטיסטיים מתחרות האירוויזיון בין השנים 1975-2002. מניע המחקר הוא ניסיון לאתר דפוסי התנהגות בהצבעות בין מדינות באירוויזיון. קישור לפרויקט: <https://github.com/omrim12/EurovisionMatches>

2. שאלת המחקר:

האם קיימות "בריתות" הצבעה בין מדינות ספציפיות באירוויזיון?
כלומר, נשאלת השאלה האם קיימת הדדיות בדפוסי ההצבעה בין מדינות מסוימות ביחס להצבעה ההמוצעת של כל מדינה - לכל שנה באירוויזיון.
יש לציין כי שאלת המחקר הנוכחית שונה מהצעת השאלה המקורית עקב דיוק ודיון עם המרצה בכיתה.

3. הדאטה ומאפייניו:

3.1. את נתוני האירוויזיון עיבדנו מתוך דאטה-סט של נתוני ההצבעות באירוויזיון לאורך השנים 1975-2019, כאשר נתוני ההצבעות שנלקחו בחשבון הם עבור תחרויות גמר ושל חבר השופטים בלבד.
הנתונים אותם עיבדנו נלקחו מדאטה-סט בשם Eurovision Song Contest scores 1975-2019 מאתר Kaggle, לינק למידע:
<https://www.kaggle.com/datasets/datagraver/eurovision-song-contest-scores-19752019>

3.2. כל נתון הצבעה מתוך דאטה-סט המקור נותח על סמך המאפיינים הבאים:

- שנה בה התבצעה ההצבעה
- מסגרת התחרות בה התבצעה ההצבעה (גמר/חצי גמר)
- אופי ההצבעה (שופטים/קהל)
- המדינה ממנה התבצעה ההצבעה
- המדינה אליה התבצעה ההצבעה
- מספר הנקודות שניתנו בהצבעה

3.3. את דאטה-סט המטרה בנינו על סמך מדד אותו הגדרנו בתור FVR – Fair Voting Ratio המתאר את מדד ההדדיות בהצבעות בין כל שתי מדינות, שאת אופן חישובו נפרט בהמשך.
מאפייני דאטה-סט המטרה אותם זיקקנו על סמך נתונים דאטה-סט המקור:
- מדינה A
- מדינה B
- סכום מדדי ה FVR מ A ל-B לאורך השנים
- סכום מדדי ה FVR מ B ל-A לאורך השנים

כמו כן, נפרט בהמשך מאפיינים נוספים אותם חילצנו מהמידע בשלבי ניתוח מתקדמים יותר של הפרויקט.

4. שיטות, ושלבים:

4.1. ניקוי נתוני המקור ובניית דאטה-סט המטרה:

- 4.1.1. סינון הצבעות של מדינות לעצמן (לא רלוונטי לניתוח מדד הדדיות)
 - 4.1.2. סינון תוצאות הצבעה עבור תחרויות חצי-גמר
 - 4.1.3. סינון הצבעות מטעם קהל (Televoting)
 - 4.1.4. סינון מדינות שלא קיבלו נקודות כלל בתחרויות מסוימות
 - 4.1.5. חישוב FVR בין כל זוג מדינות באירופיזיון – עבור הצבעה מסוימת ממדינה A למדינה B בשנה מסוימת, ננרמל את ערך ההצבעה ע"י חיסור ההצבעה הממוצעת שמדינה B קיבלה באותה שנה, זאת על מנת "לנטרל" את איכות השיר.
- את מדד FVR סכמנו לאורך השנים על מנת לזהות דפוסי הצבעה בין המדינות A ו-B

4.2. הצגת התפלגות השוני בין ערכי FVR של כל זוג מדינות

להלן המודלים השונים אותם חישבנו עבור ערכי FVR בין כל זוג מדינות באירופיזיון בשנים 1975-2002:

4.3. ניתוח מודל רגרסיה לינארית

4.4. ניתוח מודל Heatmap

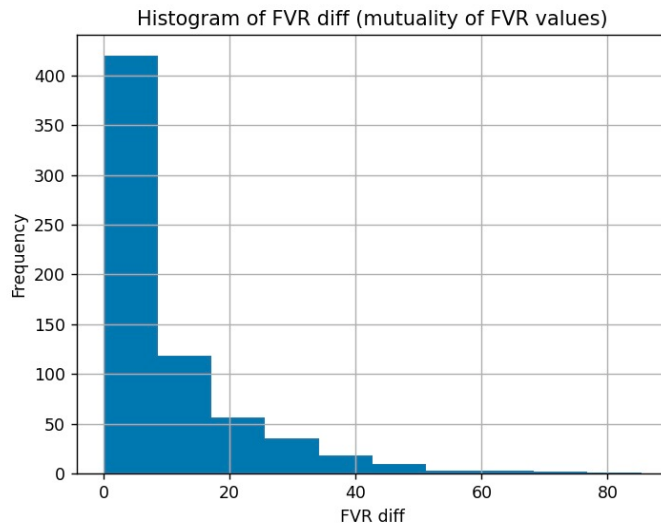
4.5. ניתוח קורלציה

4.6. ניתוח מודל Clustering בשיטת K-means

4.7. ניתוח מודל Random Forest (ולידציה של סיווג שיטת ה-Clustering)

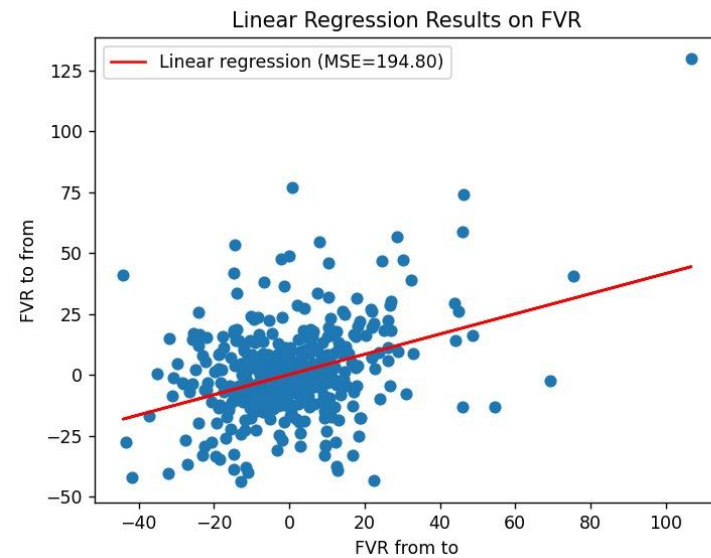
5. תוצאות:

5.1. הצגת התפלגות השוני בין ערכי FVR של כל זוג מדינות



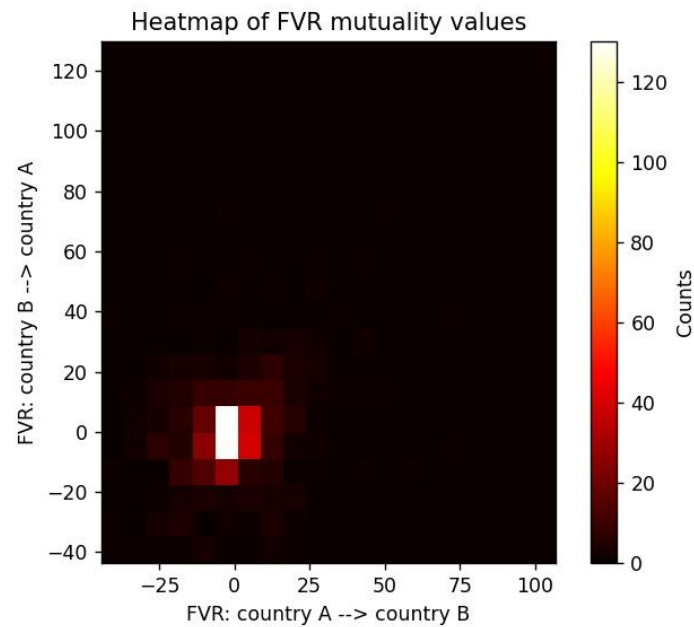
את הגרף הנ"ל חישבנו על סמך שוני ערכי FVR בין כל זוג מדינות באירופיזיון בשנים 1975-2002. כלומר: לכל זוג מדינות A.B בין השנים 1975-2002 - נסתכל על ערכי $|FVR:A \rightarrow B - FVR:B \rightarrow A|$

Linear Regression .5.2



בגרף הנ"ל הצגנו את קוארדינטות ערכי הFVR במרחב אוקלידי (R^2), זאת על מנת להעריך את התפלגות היחסים בין ערכי הFVR בין כל זוג מדינות לאורך השנים.

Heatmap .5.3



בגרף הנ"ל הצגנו את שכיחות יחסי הערכים של FVR בין כל זוג מדינות באירוויזיון לכל השנים.

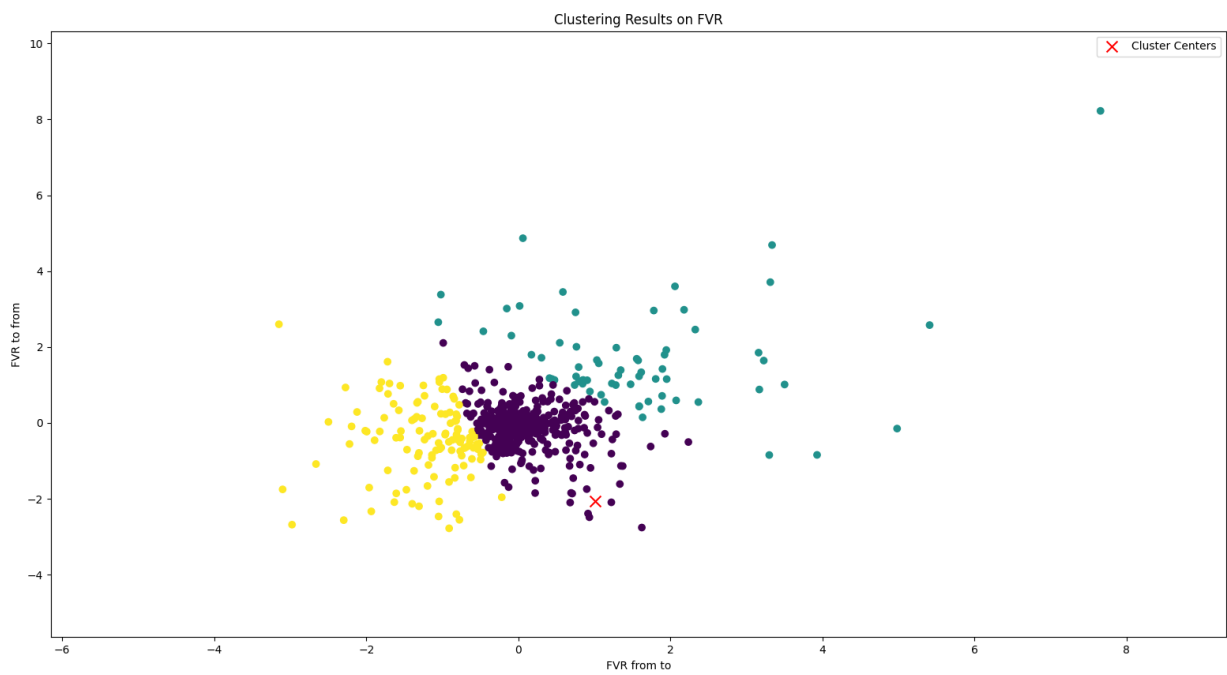
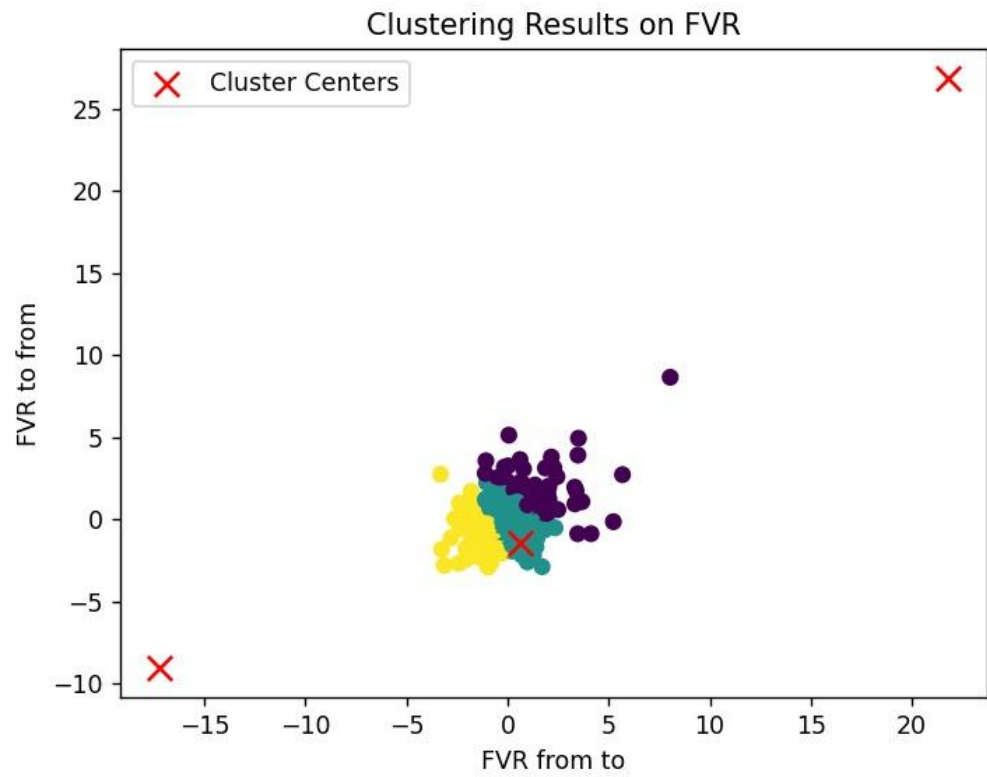
Correlation .5.4

```
# Select the relevant columns for correlation analysis
columns_of_interest = ['FVR from to', 'FVR to from']

# Calculate the correlation matrix
correlation_matrix = fvr_df[columns_of_interest].corr()

# Visualize the correlation matrix as a heatmap
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', square=True)
```


תוצאות מדד הקורלציה: 0.37
(קורלציה חיובית אך חלשה)



על אף התפלגות אחידה יחסית של יחסי הערכים של FVR בין כל זוג מדינות, רצינו להבדיל בין מס' רמות הדדיות בין כל זוג מדינות, אותם סיווגנו לקטגוריות הבאות:

- הדדיות חיובית: ערכי FVR מאוזנים וגבוהים מההצבעה הממוצעת (לפחות לאחת מהמדינות)
- הדדיות ניטרלית: ערכי FVR מאוזנים ומייצגים את ההצבעה הממוצעת (לשתי המדינות)
- הדדיות שלילית: ערכי FVR מאוזנים ונמוכים מההצבעה הממוצעת (לפחות לאחת מהמדינות)

Random Forest 5.6



average accuracy score over 20 iterations: 97.31%

אימנו מודל מסוג Random Forest, בעזרת דוגמאות מתויגות שנוצרו על ידי תיוג ה-K-means.

חילקנו 80 אחוזים לאימון ו-20 אחוזים לבדיקה.

ניתן לראות שלאחר 20 הרצות, ממוצע הדיוק הוא קצת יותר גבוה מ-97 אחוזים.

6. מסקנות ותובנות:

ראשית, ניתן להסיק מאופן ההתנהגות של כל אחד מהמודלים שניתחנו כי תחרות האירוויזיון לאורך השנים מייצגת תחרות הוגנת שלרוב לא קיימים בה דפוסים הצבעות הדדיות חיוביות בין מדינות (מרבית מכלל ההצבעות מתארות התנהגות צפויה). למרות זאת, קיימים צמדים של מדינות ביניהן דפוס ההצבעות היה שונה בצורה מהותית ועליהן נוכל להסיק מסקנות מעניינות יותר בהמשך. נתאר כעת את המסקנות העולות מכל מודל באופן ספציפי:

- הצגת התפלגות השוני בין ערכי FVR של כל זוג מדינות
מאחר ומדד FVR מחושב על סמך היחס בין דפוס ההצבעה ממדינה מסוימת למדינה מסוימת ביחס לדפוס ההצבעה הממוצע - נוכל להסיק כי קיים יחס חזק בדפוס ההצבעות בין מדינות או לחלופין כי רוב ההצבעות בכלל תחרויות האירוויזיון תאמו את ההתנהגות הנורמלית.
טענות אלו אינן מבוססות מפסיק ועל כן נשתמש במודלים הבאים כדי לנתח זוויות נוספות של דפוס ההצבעות.
- רגרסיה לינארית
ניתן להבחין בצורה ויזואלית וכמו כן עפ"י ערך MSE של המודל כי לא קיימת רגרסיה משמעותית ביחסי הערכים של FVR בין צמדי מדינות באירוויזיון לאורך השנים.
- HeatMap
מאחר ומרבית מבין ערכי FVR בין כל זוג מדינות נמצאים בראשית הצירים, ניתן להסיק כי רוב ההצבעות (ללא תלות בדפוס ההצבעות ההדדיות בין כל זוג מדינות) מתארות התנהגות נורמלית (הצבעה קרובה לממוצע).
- קורלציה
ערך הקורלציה מתאר קשר חלש יחסית בדפוס ההצבעות בין כל זוג מדינות לאורך השנים.
- Clustering - K-means

Cluster	count
0	419.0
1	67.0
2	117.0

- 0 - קבוצה סגולה (הדדיות ניטרלית)
- 1 - קבוצה ירוקה (הדדיות חיובית)
- 2 - קבוצה צהובה (הדדיות שלילית)

מאחר וגודל הקלאסטר המזוהה עם קטגוריית ההדדיות הניטרלית גדול באופן מהותי מאלו של ההדדיות החיובית והשלילית - ניתן להסיק כי מרבית היחסים בין דפוס ההצבעות בין מדינות באירוויזיון הם "הוגנים", כלומר מאוזנים ותואמים את הדפוס הממוצע.

• מדינות יוצאות דופן

Top 12 outlier countries pairs:					
	From	To	FVR from to	FVR to from	Distance
0	Greece	Cyprus	106.748612	129.836425	168.085584
1	Sweden	Denmark	46.412189	74.100416	87.435479
2	Norway	Sweden	75.359416	40.837576	85.713180
3	Italy	Portugal	0.645213	76.932971	76.935676
4	Spain	Greece	46.092688	58.672827	74.612576
5	France	Portugal	69.383450	-2.163997	69.417188
6	Italy	Spain	28.580929	56.932971	63.704259
7	Spain	Switzerland	-44.178320	41.177487	60.392958
8	Spain	Sweden	-41.771019	-42.084545	59.295252
9	Finland	Italy	54.682971	-13.100578	56.230352
10	Denmark	Iceland	30.249106	47.194865	56.056790
11	Portugal	Spain	-14.427256	53.501097	55.412211

על סמך הממצאים הנ"ל ניתן לפרש מספר מסקנות:

1. ע"ס נתונים מויקיפדיה לפיהם 77% מאוכלוסיית קפריסין הם ממוצא יווני, ניתן להסיק כי דפוס ההצבעות המופיע בטבלה מתאר יחסי קרבה חזקים בצורה הגיונית.
2. נראה כי קיים דפוס הצבעות הדדיות חיוביות בין צמדי מדינות מצפון אירופה (שוודיה ודנמרק, שוודיה ונורווגיה, דנמרק ואיסלנד) שעלול להעיד על יחסים טובים בין המדינות באזור.
3. ניתן להבחין בקשר שלילי בדפוס ההצבעות בין שוודיה לספרד - אך מחיפוש כללי על היסטוריית היחסים בין שוודיה לספרד לא קיים מניע בין המדינות.

מקרא:

- From - FVR from <country>
- To - FVR to <country>
- FVR from to - value of FVR from ... to ...
- FVR to from - value of FVR to ... from ...
- Distance - Significance rate of both FVR values
(Calculated using Euclidean distance)

7. לקחים:

מאחר וחלק מהמדינות (ביניהן דפוס ההצבעה היו שונים באופן מהותי מהדפוס הממוצע) נמצאות בקרבה גאוגרפית, חלוקה מקדימה לאיזורים גיאוגרפיים היה יכול לתת עוד פרשנות מעניינת להתנהגות ההצבעות השונות.