הטכניון, מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות

236756

סמסטר אביב תשע"ט

תרגיל מספר: 4

תאריך הגשה: 18/06/19

תא להחזרה: 68

מגישים:

עידו יחזקאל	8 6 3 7 9 3 4 0 2	idoye
שם מלא	מספר ת.ז.	דואר אלקטרוני ב- t2

אמיר אביבי	3 7 8 3 8 1 5 0 3	saavivi
שם מלא	מספר ת.ז.	דואר אלקטרוני ב- t2

Mandatory Part – Loading and Preparing the Data:

ראשית כמתבקש טענו את הקובץ הדוגמאות שוב, והחלנו עליו את מניפולציות העיבוד המקדים להכנת המידע כפי שביצענו בתרגילים הקודמים, מניפולציות אלה כוללות:

1. חלוקת סט הדוגמאות הכולל (על די Stratified Shuffle Split) לשלושה סטים:

Data Set	Percentage from Original Data Set
Train set	65%
Validating set	10%
Test set	25%

2. הוצאות ערכים שהינם outliers, לדוגמא ערכים שליליים.
3. השלמת ערכים חסרים לפי השיטות המקובלות: closest fit, feature correlation, mean and majority.
4. בחירת סט הפיצ'רים הנכון כפי שנבחר בתרגיל מספר 3.
5. ביצוע נורמליזציה לערכים קטיגוריאליים וZ-scale לערכים נומינליים.
6. ייצוא המידע ל 3X2 קבצי CSV לפני ואחרי השנויים.

לאחר מכן, ניגשנו למשימת החיזוי כאשר בחלק החובה נדרשנו לחזות:

- הרכבת הקואליציה יציבה הכוללת 51% מסך כל הקולות.
- זיהוי פיצ'רים דומיננטיים עבור כל מפלגה וביצוע מניפולציות על מנת **לחזק את הקואליציה ובניית קואליציה אלטרנטיבית.**

Mandatory Part – Building Steady Coalition:

לצורך בניית קואליציה יציבה לפי הגדרתה בתרגיל אנו נדרשים למצוא קבוצת מפלגות אשר קיים דמיון בין המאפיינים של המצביעים שלהם. מכיוון שבבעיה זו רוב מאפייני המצביעים (לאחר בחירת הפיצ'רים) הם מאפיינים בעלי ערך מספרי רציף הרי שנוכל למדוד בין שני דגימות על ידי מרחק בין הערכים המספריים של מאפייני הדגימות.

על מנת לקבץ את המפלגות לכדי קואליציה הומוגנית פעלנו בשתי שיטות הבאות:

1. Clustering Model- כאשר אנו משתמשים במודל מסוג זה אנו למעשה "מקבלים בחינם" את היכולת לזהות את הדמיון והשוני בין המצביעים באמצעות היכולת להתייחס למצביעים אשר **נמצאים** באותו אשכול כבעלי מאפיינים דומים ואילו למצביעים אשר **אינם נמצאים** באותו אשכול כחסרי דמיון.
   בצורה זו אנו ננסה ליצור **קבוצת מצביעים** השייכת לאשכול מסוים אשר יחסית הומוגנית והיא זאת שתהווה **בסיס להרכבת הקואליציה**.
2. Generative Model- כמאפיין את מודל זה לאחר אימון המודל אנו יכולים לקבל את מאפייני פונקציית הסתברות של מפלגה מסוימת, לדוגמא במקרה של Gaussian Naïve Base נוכל לחלץ מהמודל את השונות והתוחלת עבור התפלגות מפלגה מסוימת (נשים לב שעבור הבעיה שלנו אלה הם וקטורים בגודל 9 כמספר הפיצ'רים).
   כתוצאה מכך נוכל למדוד דמיון בין שני מפלגות על ידי השוואה בין מאפייני פונקציות ההסתברות של כל אחת מהן.

אלו הן למעשה דרכי הפעולה העיקריות שפיתחנו ובעזרת כל אחת מהן הצלחנו לבנות קואליציה לפי ההגדרה כפי שמוסבר בהמשך בהרחבה.

Building Steady Coalition Using Clustering Model:

לשם בניית קואליציה יציבה על ידי שימוש ב Clustering Model אימנו שני מודלים שאנו מכירים מהכיתה: KMeans ו Gaussian Mixture .
ראשית עבור כל אחד מהמודלים נצטרך לבחור היפר-פרמטרים הטובים ביותר לבניית קואליציה ולשם כך השתמשנו ב K Fold Cross Validation , אמנם מכיוון שאלו מודלים השייכים ל – unsupervised learning היינו צריכים בעצמנו לפתח מדד כיצד לבחור היפר-פרמטרים טובים יותר לבעיה איתה אנו מתמודדים.
המדד שפיתחנו הוא העדפת מודל המנסה לרכז כל מפלגה באשכול מסוים ולא מפזר אותה על פני כמה אשכולות בצורה שלא ניתנת להבחנה לאיזה אשכול היא שייכת, כלומר אנו נעדיף מודלים אשר אינם מפזרים כל מפלגה בין כמה אשכולות אלא יוצרים אשכולות אשר ניתן להבחין בצורה מובהקת כי איזו מפלגה שייכת לכל אשכול. לדוגמא: אם מודל מסוים מנסה לחלק את המידע ל -2 אשכולות וקיבלנו כי עבור כל מפלגה 51% מסך הקולות שלה נמצאים באשכול מסוים והשאר באחר אזי המודל אינו מחלק את הקולות בצורה שמאפשרת לנו לבנות קואליציה הומוגנית מספיק. לעומת זאת מודל אשר קיימות בו מפלגות אשר לפחות 70% מקולות המפלגה שייכם לאשכול מסוים והשאר לאחר אזי נעדיף את המודל הזה יותר מכיוון שהוא יוצר אשכולות יותר מקובצים לפי מפלגות.
לכן הדרך שבה ביצענו K Fold Cross Validation היא:

1. אימון המודל על כל חלק אימון.
2. חיזוי על חלק המבחן.
3. עבור כל אחד מהאשכולות שנוצרו חישבנו כמה מפלגות שייכות לאשכול זה (גודל האשכול), כאשר מפלגה תחשב לשייכת לאשכול אם 60% מסך כל הקולות שלה שייך לאשכול זה.
4. חישוב סכום הגדלים של כל האשכולות מכל החלקים.
5. העדפת מודלים עם אשכולות יותר גדולים.

בעזרת שיטה זו בדקנו 3 מספרי אשכולות אפשריים: 2,3,4 עבור KMeans ו Gaussian Mixture .
המודלים שנבחרו הינם:

- KMeans with 2 clusters.
- Gaussian Mixture with 2 clusters.

לאחר בחירת שני המודלים הללו ניגש להרכבת הקואליציה. השלבים להרכבת הקואליציה הינם:

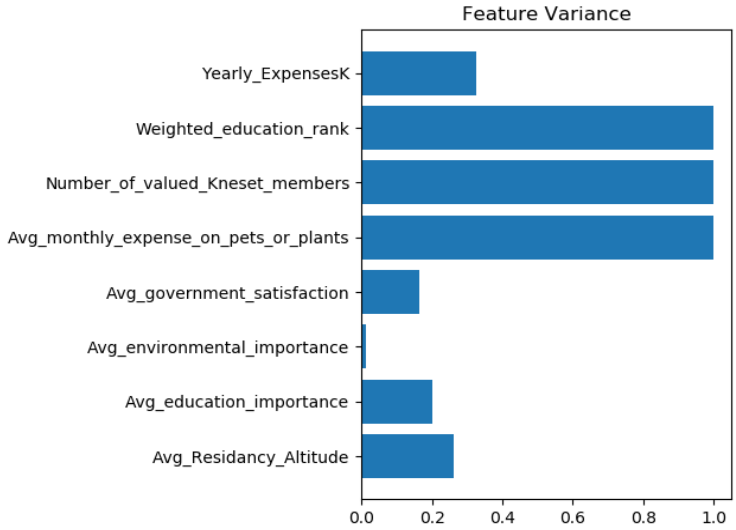
1. אימון כל אחד המודלים על כל סט האימון.
2. הרכבת הקואליציה הצפויה בעזרת סט הווליד‌ציה. כל אשכול שהמודל יצר יכול להוות בסיס להרכבת קואליציה, **נגדיר שמפלגה שייכת לאשכול מסוים אם ורק אם לפחות 80% מסך כל הקולות שהצביעו לה שייכים לאשכול זה** ובנוסף סך כל הקולות של המפלגות ששייכות לאשכול עולה על 51% מסך כל הקולות. בצורה זו בנינו קואליציה עם לפחות רוב קולות, קואליציה יחסית הומוגנית מכיוון שעבור מפלגות ששייכות לאשכול הרי שהמודל בחר לשים את רוב המצביעים שלהם באותו אשכול ולכל קיים דמיון בין המאפיינים שלהם, לעומת מפלגות

שלא שייכות לאשכול הרי שלא קיים רוב מובהק של המצביעים שלהן באשכול ולכן הם יהיו באופוזיציה.

3. סינון קואליציות זהות שהורכבו ובחירת הקואליציה הכי הומוגנית, קואליציה שבה השונות בין מאפייני המצביעים של הקואליציה כולה היא הקטנה ביותר.
4. בחינת ביצועי המודל על ידי ניסיון הרכבת קואלציה בעזרת סט המבחן.

<u>תוצאות</u>:

נשים לב כי לפני הרכבת קואליציה ניתן להבחין כי זוהי השונות בין מאפייני המצביעים השונים על סט האימון:

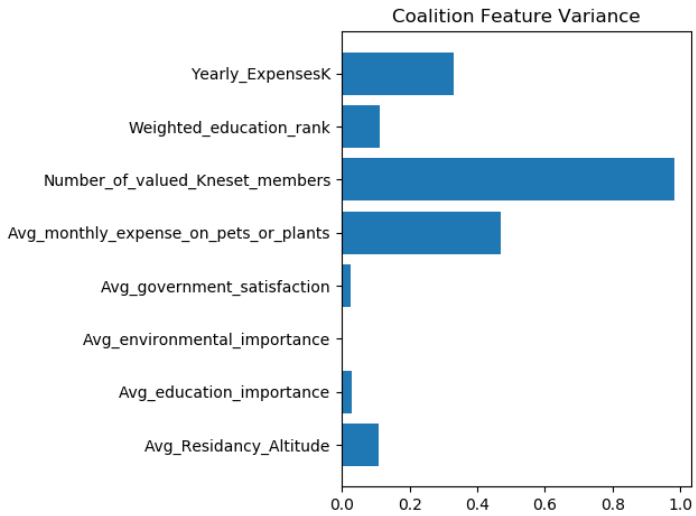


(הפיצ'רים לאחר scaling)

הקואליציה הסופית שנבחרה נבנתה על ידי המודל Gaussian Mixture בעזרת **סט הוולידציה**(הבסיס להרכבתה הוא אשכול 0). הקואליציה כוללת את המפלגות הבאות וכוללת 60.8% מסך כל הקולות.

Browns, Greens, Greys, Oranges, Pinks, Purples, Reds, Whites.

לאחר הרכבת קואליציה ניתן להבחין כי ישנה ירידה בשונות בין רוב מאפייני המצביעים השייכים קואליציה על סט האימון דבר המעיד כי קבוצת מצביעים זו יותר הומוגנית מהקבוצה הכוללת:

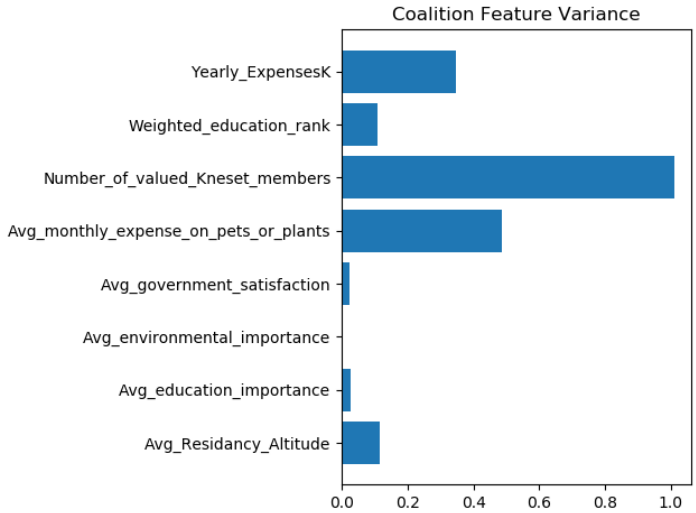


כעת על מנת לבחור את ביצועי המודל, השתמשנו בסט המבחן כאשר על מנת לחזות את הסיווג של כל מצביע השתמשנו במסווג שבחרנו בתרגיל הקודם בעל דיוק של 92.64% והוא:

```
RandomForestClassifier(random_state=0, criterion='gini', n_samples_split=3,
                       min_samples_leaf=1, n_estimators=500)
```

הקואליציה הסופית שנבחרה נבנתה על ידי המודל Gaussian Mixture בעזרת **סט המבחן**(הבסיס להרכבתה הוא אשכול 0). הקואליציה כוללת את המפלגות הבאות וכוללת 60.88% מסך כל הקולות. Browns, Greens, Greys, Oranges, Pinks, Purples, Reds, Whites.

לאחר הרכבת קואליציה ניתן להבחין כי ישנה ירידה בשונות בין רוב מאפייני המצביעים השייכים לקואליציה על סט האימון וכמעט שוויון אל הקואליציה שנבנתה בעזרת סט הוולידציה.



כלומר המודל שבנינו הצליח לבנות קואליציה לפי ההגדרה ואף לחזות את הקואליציה מראש בעזרת סט האימון הוולידציה.

Building Steady Coalition Using Generative Model:

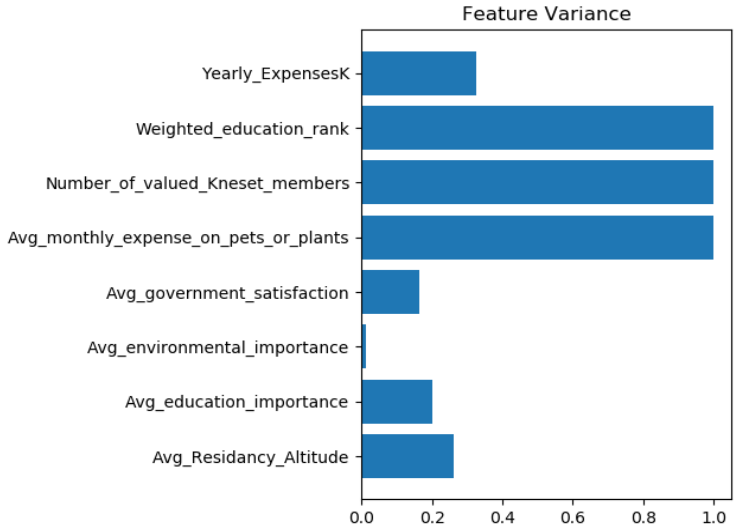
לשם בניית קואליציה יציבה על ידי שימוש ב Generative Model אימנו שני מודלים שאנו מכירים מהכיתה: QDA ו Gaussian Naïve Base .
ראשית עבור כל אחד מהמודלים נצטרך לבחור היפר-פרמטרים הטובים ביותר לבניית קואליציה ולשם כך השתמשנו ב K Fold Cross Validation כאשר המדד אותו רצינו למקסם הינו מדד הדיוק.

לאחר בחירת היפר-פרמטרים עבור כל אחד מהמודלים הללו ניגש להרכבת הקואליציה. השלבים להרכבת הקואליציה הינם:

1. חישוב וקטור התוחלת של כל מפלגה על ידי שימוש ב one Vs. all ואימון כל אחד מהמודלים על כל סט האימון.
2. הרכבת הקואליציה הצפויה בעזרת סט הוולידציה, תוך כדי לקיחת השראה מהמערכת הפוליטית בישראל אשר מוטלת הרכבת הקואליציה על ראש מפלגה מסוימת ניסנו להרכיב קואליציה על ידי הטלת ההרכבה על מפלגה מסוימת אשר מהווה בסיס לקואליציה. מפלגה נוספת לקואליציה כאשר וקטור התוחלת שלה הוא הקרוב ביותר לווקטור התוחלת של המפגלה שמנסה להרכיב את הקואליציה. הרכבת הקואליציה נפסקת כאשר הושג רוב מצביעים של לפחות 51% מסך כל הקולות.
3. סינון קואליציות זהות שהורכבו ובחירת הקואליציה הכי הומוגנית, קואליציה שבה השונות בין מאפייני המצביעים של הקואליציה כולה היא הקטנה ביותר.
4. בחינת ביצועי המודל על ידי ניסיון הרכבת ואלציה בעזרת סט המבחן.

<u>תוצאות</u>:

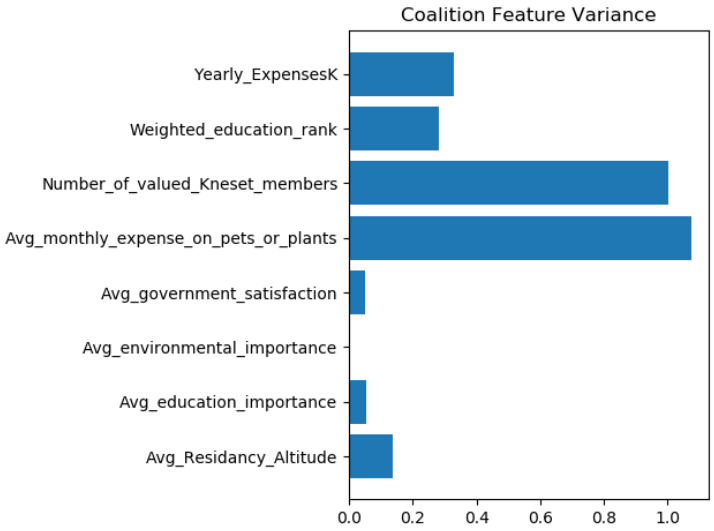
כפי שהוצג קודם לכן זוהי השונות בין מאפייני המצביעים השונים על סט האימון:



(הפיצ'רים לאחר scaling)

הקואליציה הסופית שנבחרה נבנתה על ידי **שני המודלים** (כלומר שני המודלים בנו את אותה קואליציה) בעזרת **סט הוולידציה** כוללת את המפלגות הבאות וכוללת 60.8% מסך כל הקולות.
Blues, Browns, Greys, Oranges, Pinks, Purples, Reds, Whites.

לאחר הרכבת קואליציה ניתן להבחין כי ישנה ירידה בשונות בין רוב מאפייני המצביעים השייכים לקואליציה על סט האימון דבר המעיד כי קבוצת מצביעים זו יותר הומוגנית מהקבוצה הכוללת:



כעת על מנת לבחור את ביצועי המודל, השתמשנו בסט המבחן כאשר על מנת לחזות את הסיווג של כל מצביע השתמשנו במסווג שהוזכר לעיל.
מכיוון ששני המודלים יצרו את אותה קואליציה השתמשנו ב Gaussian Naïve Base לבניית הקואליציה הסופית בעזרת **סט המבחן** הקואליציה שהורכבה זהה לקודמת.

לסיכום, מודל זה גם אפשר לנו לבנות קואליציה שונה מהקואליציה שנבנתה על ידי המודל הקודם ובעלת אחוז קולות כמעט זהה אמנם כפי שניתן להשוות את ההומוגניות של 2 הקואליציות שבנינו הרי זאת שנבנתה על ידי ה Clustering model יותר הומוגנית ולכן נעדיף אותה.
ולכן הקואליציה שאנו מציעים הינה:

Browns, Greens, Greys, Oranges, Pinks, Purples, Reds, Whites.

Identifying Each Party Leading Features:

בתרגילים הקודמים ניגשנו לבעיה זו על ידי הצגת כל פיצ'ר אל מול ההצבעות האפשריות, בתרגיל זה ניגש לבעיה בצורה שונה תוך התחשבות בנעשה בתרגילים קודמים.
על מנת לזהות לכל מפלגה את הפיצ'רים המאפיינים אותה נעזר ב Embedded feature selection שהינה חלק אינטגרלי מעץ החלטה.
עבור כל אחת מהמפלגות אימנו את המסווג הנבחר שלנו:

```
RandomForestClassifier(random_state=0, criterion='gini', n_samples_split=3,
                       min_samples_leaf=1, n_estimators=500)
```

בשיטת one Vs. all והוצאנו את הפיצ'רים המובלים על ידי התכונה feature importance's שקיימת למסווג זה.
להלן עבור כל מפלגה הפיצ'רים החשובים לפי סדר חשיבות:

Party	Leading Features
Blues	1. Avg_monthly_expense_on_pets_or_plants 2. Weighted_education_rank 3. Avg_Residancy_Altitude 4. Avg_government_satisfaction 5. Avg_environmental_importance 6. Avg_education_importance 7. Yearly_ExpensesK 8. Number_of_valued_Kneset_members 9. Most_Important_Issue
Browns	1. Number_of_valued_Kneset_members 2. Avg_education_importance 3. Avg_environmental_importance 4. Avg_government_satisfaction 5. Weighted_education_rank 6. Avg_Residancy_Altitude 7. Most_Important_Issue 8. Avg_monthly_expense_on_pets_or_plants 9. Yearly_ExpensesK
Greens	1. Avg_education_importance 2. Avg_environmental_importance 3. Avg_government_satisfaction 4. Weighted_education_rank 5. Avg_Residancy_Altitude 6. Avg_monthly_expense_on_pets_or_plants 7. Number_of_valued_Kneset_members 8. Yearly_ExpensesK 9. Most_Important_Issue
Greys	1. Most_Important_Issue 2. Weighted_education_rank 3. Number_of_valued_Kneset_members 4. Avg_government_satisfaction 5. Avg_education_importance 6. Avg_environmental_importance 7. Avg_Residancy_Altitude 8. Yearly_ExpensesK 9. Avg_monthly_expense_on_pets_or_plants
Khakis	1. Weighted_education_rank 2. Avg_Residancy_Altitude 3. Avg_education_importance

	4. Avg_government_satisfaction 5. Avg_environmental_importance 6. Avg_monthly_expense_on_pets_or_plants 7. Yearly_ExpensesK 8. Number_of_valued_Kneset_members 9. Most_Important_Issue
Oranges	1. Most_Important_Issue 2. Weighted_education_rank 3. Avg_environmental_importance 4. Number_of_valued_Kneset_members 5. Avg_government_satisfaction 6. Avg_education_importance 7. Avg_Residancy_Altitude 8. Avg_monthly_expense_on_pets_or_plants 9. Yearly_ExpensesK
Pinks	1. Avg_environmental_importance 2. Weighted_education_rank 3. Avg_government_satisfaction 4. Avg_education_importance 5. Number_of_valued_Kneset_members 6. Avg_Residancy_Altitude 7. Avg_monthly_expense_on_pets_or_plants 8. Yearly_ExpensesK 9. Most_Important_Issue
Purples	1. Weighted_education_rank 1. Avg_environmental_importance 2. Number_of_valued_Kneset_members 3. Avg_Residancy_Altitude 4. Avg_government_satisfaction 5. Avg_education_importance 6. Most_Important_Issue 7. Yearly_ExpensesK 8. Avg_monthly_expense_on_pets_or_plants
Reds	1. Avg_education_importance 2. Most_Important_Issue 3. Weighted_education_rank 4. Number_of_valued_Kneset_members 5. Avg_environmental_importance 6. Avg_government_satisfaction 7. Avg_Residancy_Altitude 8. Yearly_ExpensesK 9. Avg_monthly_expense_on_pets_or_plants
Turquoises	1. Avg_government_satisfaction 2. Avg_environmental_importance 3. Weighted_education_rank 4. Avg_Residancy_Altitude 5. Avg_education_importance 6. Avg_monthly_expense_on_pets_or_plants 7. Yearly_ExpensesK 8. Number_of_valued_Kneset_members 9. Most_Important_Issue
Violets	1. Avg_government_satisfaction 2. Avg_education_importance 3. Weighted_education_rank 4. Avg_environmental_importance 5. Avg_Residancy_Altitude 6. Avg_monthly_expense_on_pets_or_plants 7. Yearly_ExpensesK 8. Number_of_valued_Kneset_members 9. Most_Important_Issue

Whites	1. Avg_education_importance 2. Avg_environmental_importance 3. Avg_government_satisfaction 4. Number_of_valued_Kneset_members 5. Weighted_education_rank 6. Avg_Residancy_Altitude 7. Most_Important_Issue 8. Yearly_ExpensesK 9. Avg_monthly_expense_on_pets_or_plants
Yellows	1. Avg_monthly_expense_on_pets_or_plants 2. Avg_government_satisfaction 3. Avg_Residancy_Altitude 4. Weighted_education_rank 5. Avg_environmental_importance 6. Avg_education_importance 7. Yearly_ExpensesK 8. Number_of_valued_Kneset_members 9. Most_Important_Issue

לפני ביצוע מניפולציות על סט הפיצ'רים ולנסות לשנות את הקואליציה הקיימת ננסה להבין איזה פיצ'רים הפרידו בינה לבין האופוזיציה על ידי מרכזי האשכולות:

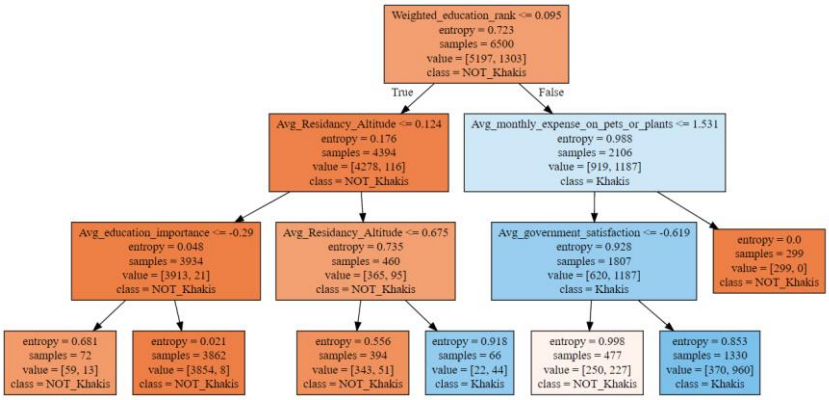
Feature	Coalition Cluster	Opposition Cluster
Avg_Residancy_Altitude	-0.3998517	0.37721118
Avg_education_importance	0.01104139	0.00684677
Avg_environmental_importance	-0.79099666	-0.75936145
Avg_government_satisfaction	-0.12007848	-0.02143256
Avg_monthly_expense_on_pets_or_plants	-0.18421413	0.27515572
Most_Important_Issue	3.16998264	2.78523297
Number_of_valued_Kneset_members	-0.29232246	0.43663424
Weighted_education_rank	-0.65683802	0.98110137
Yearly_ExpensesK	0.00516143	-0.01156653

אבחנות:

- אלה הם הפיצ'רים שיוצרים הבדל מובהק בין הקואליציה לאופוזיציה.
- שמנו לב כי הפיצ'ר אשר **אינו** מוביל ברוב מפלגות הקואליציה הינו Avg_monthly_expense_on_pets_or_plants ולכן על מנת לבנות קואליציה חלופית בעזרת שיטה זו נרצה ליצור הומוגניות בעבור פיצ'ר זה אצל המפלגות באופוזיציה.
- פיצ'ר מוביל במפלגות הקואליציה וגם במפלגות האופוזיציה הינו Weighted_education_rank.

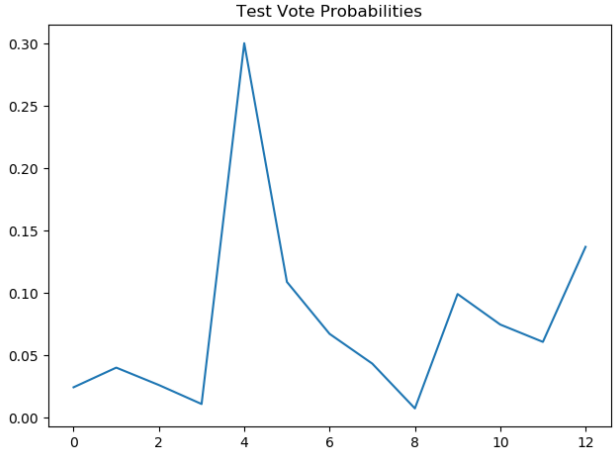
שינוי המפלגה המנצחת ובניית קואליציה חליפית:

על מנת לשנות את המפלגה המנצחת נרצה לבצע מניפולציות על הפיצ'רים: Weighted_education_rank Avg_monthly_expense_on_pets_or_plants אשר דומיננטיים אצל המפלגה השנייה בגודלה החאקי.
בנוסף נרצה ששינויים ירכיבו אשכולות חדשים שיהוו בסיס לקואליציה חלופית. מהתרגיל הקודם זיהנו כי שני המפלגות המובילות הינן הסגולים וחאקי ולכן מכיוון שהחאקי אינם בקואליציה שלנו נרצה לחזק אותם, את המספרים המדויקים למניפולציה לקחנו מעת ההחלטה הבא:

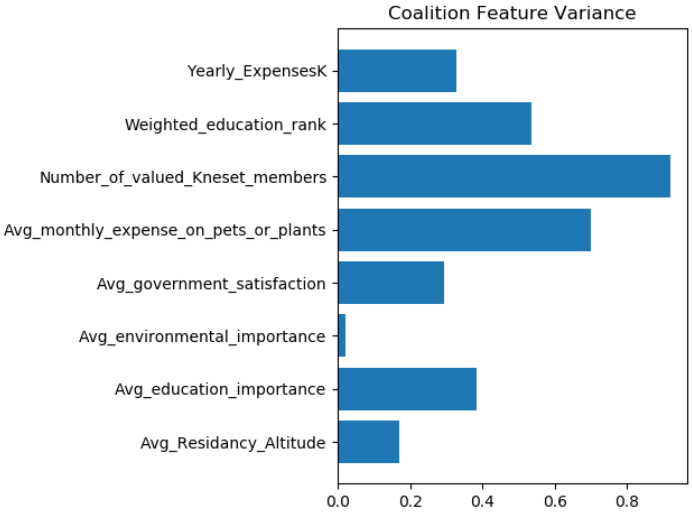


המניפולציה כוללת הגדלת ערכי Weighted_education_rank והקטנת ערכי Avg_monthly_expense_on_pets_or_plants.

הצלחנו לשנות את המפלגה המנצחת לחאקי(4):



ולאחר המניפולציה שלנו התקבלה הקואליציה החלופית הבאה עם 52.1% מהקולות:
Blues, Greys, Khakis, Oranges, Reds, Turquoises, Violets, Yellows
וגרף השונות הבא:



אשר ניתן לשים כי ההומוגניות של קואליציה זו קטנה (שונות גדלה ברוב הפיצ'רים) יותר מאשר הקודמת.

חיזוק הקואליציה הנוכחית:

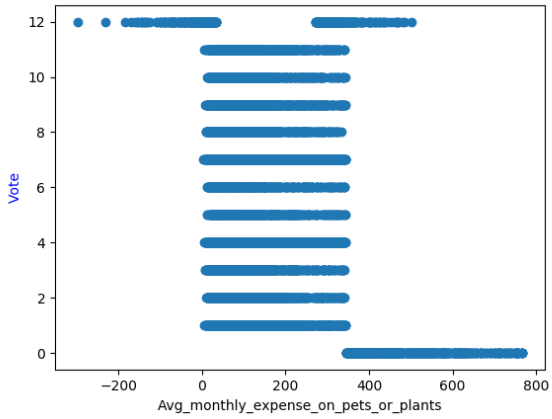
על מנת לחזק את הקואליציה חשבנו על הוספת מפלגה שמחוץ לקואליציה אליה. על מנת לעשות זאת בחרנו על מניפולציות שיוספו את מפלגת החאקי לקואליציה, מכיוון שהתכונות של הדומיננטיות שלה יחסית קרובות למפלגה הכי גדולה בקואליציה, הסגולים. תכונות אלו הן Weighted_education_rank Avg_Residancy_Altitude אשר גם יוצרות הפרדה מובהקת בין האשכולות.

המניפולציה שביצענו הייתה לקרב את תכונות אלו אל מרכז האשכול שהרכיב את הקואליציה ובצורה זו למנוע מתכונות אלו ליצור הפרדה בין האשכולות וכך לתת לשאר התכונות המפרידות בין הקואליציה לאופוזיציה ליצור את ההפרדה בין האשכולות.

לאחר המניפולציה שלנו התקבלה הקואליציה המחוזקת הבאה עם 81.5% מהקולות:

Browns, Greens, Greys, Khakis, Oranges, Pinks, Purples, Reds, Violets , Whites.

נשים לב כי המפלגות שנשארו בחוץ הן: Blues(0), Turquoises (9), Yellows(12).
נסביר זאת על ידי הפיצ'ר שחשוב עבורן שהוא Avg_monthly_expense_on_pets_or_plants אשר כפי שראינו בתרגיל השני יוצר הפרדה בין Blues(0), Yellows(12) לבין שאר המפלגות והוא זה פיצ'ר דומיננטי להפרדה בין האשכולות:



(Turquoises לא נכנסו לקואליציה בגלל שהיו מפוזרים מידי בין האשכולות).