הטכניון, מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות

236756

סמסטר אביב תשע"ט

תרגיל מספר: 5

תאריך הגשה: 30/06/19

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 שקיבלנו בתחילת הקורס 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 .2
Train set	65%
Validating set	10%
Test set	25%

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

כחלק מהתאמת סט המבחן החדש בתרגיל זה ביצענו את אותן מניפולציות שביצענו על סט האימון בדיוק גם על סט המבחן הלא מתויג וזאת על מנת שהמסווג שלנו יתמודד עם סט המבחן כפי שביצענו לאורך כל הסמסטר.

לאחר מכן, ניגשנו למשימת החיזוי כאשר נדרשנו לחזות:

- מה היא המפלגה המנצחת לפי סט המבחן.
- מה הוא פילוח הקולות לפי סט המבחן.
- עבור כל מצביע מסט המבחן לחזות את הצבעתו.
- לחזות קואליציה יציבה והומוגנית הכוללת לפחות 51% מסך כל הקולות.

Mandatory Part – Voting Predictions:

אנו מבינים שעל מנת לבצע את התחזיות הנ"ל על סט המבחן הלא מתויג אנו חייבים לבנות מסווג בעל יכולת הכללה גבוהה שתביא דיוק גבוה וזאת מכיוון שאין אנו יכולים להעריך את ביצועי המסווג על סט המבחן הלא מתויג.

כזכור, בתרגיל בית 3 עסקנו במשימות דומות ושם קיבלנו את התוצאות הבאות:
תרגיל בית 3 Flashback:

Random Forest Classifier accuracy score on validation set is: 90.1%

SGD Classifier accuracy score on validation set is: 74.7%

KNN Classifier accuracy score on validation set is: 77.5%

Decision Tree Classifier accuracy score on validation set is: 85.8%

מכיוון שלפי הדרישות רצינו להגיע למסווג בעל 90-95% דיוק הגענו למסקנה שעלינו לשפר את יכולת הסיווג שלנו.
על מנת להרכיב מסווג עם יכולת הכללה טובה החלטנו להרכיב ועדה הכוללת את שלושת המסווגים הבולטים שראינו במהלך הקורס, ועדה זאת תכלול שלושה מסווגים ותורכב מ - Random Forest, Multi-Level Perception, SVM.

בחרנו דווקא בשלושת מסווגים אלה כי אלו התבלטו ביכולת הכללה טובה ובביצועים טובים כפי שראינו בתרגול.

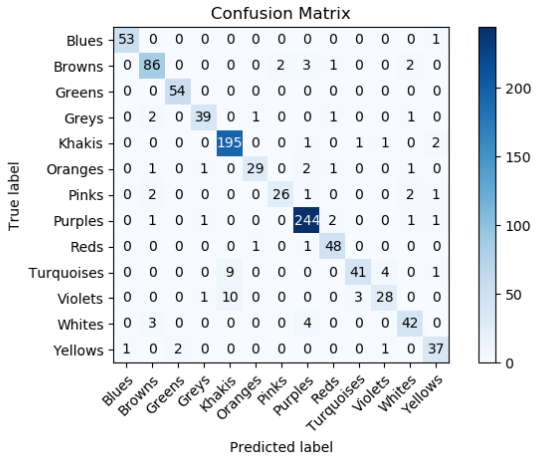
מכיוון שבתרגילי בית קודמים השתמשנו רק ב Random Forest היינו צריכים לכוון את הפרמטרים של השניים האחרים ובנוסף החלטנו לנסות למצוא פרמטרים טובים יותר גם עבור ה Random Forest.

מציאת הפרמטרים הטובים ביותר לכל מסווג התבצעה באופן הבא עבור כל מסווג ביצענו Random Search על מבחר סטים של פרמטרים כאשר כל סט של פרמטרים הוערך על ידי K-fold cross validation.

המוטיבציה לבצע זאת היא למקסם את יכולת ההכללה של כל אחד מהמסווגים בוועדה וכך ליצור ועדה כוללת חזקה יותר מכל אחד מהמסווגים בעצמם.

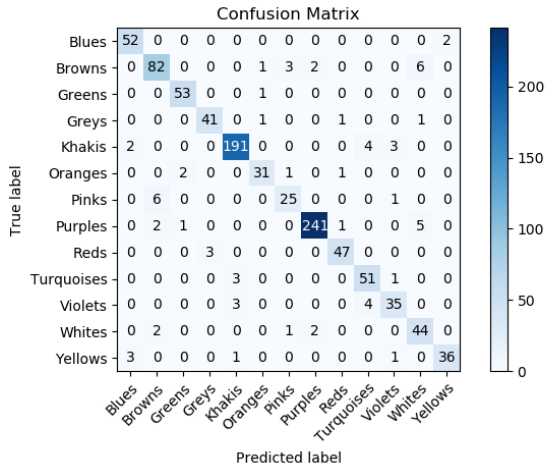
בחינת ביצועי כל אחד מהמסווגים עם ההיפר פרמטרים שנמצאו:

Random Forest- הניב דיוק של 92.2% על סט הוולידציה.



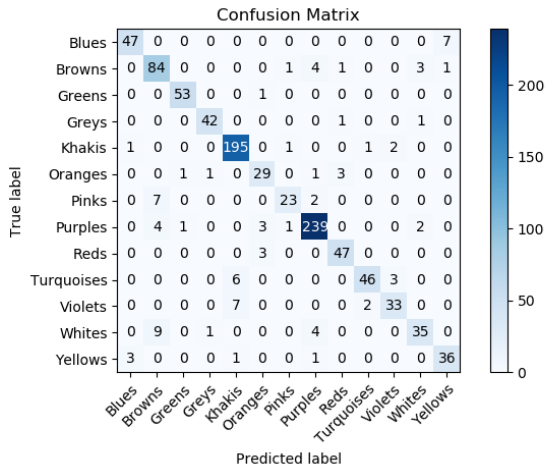
כפי שניתן לראות נוטה לסווג את המפלגות טורקיז וסיגל (violets) לצבע החאקי.

MLP- הניב דיוק של 92.9% על סט הוולידציה.



כפי שניתן לראות נוטה לסווג את המפלגה הוורודה למפלגה החומה.

SVM- הניב דיוק של 90.9% על סט הוולידציה.



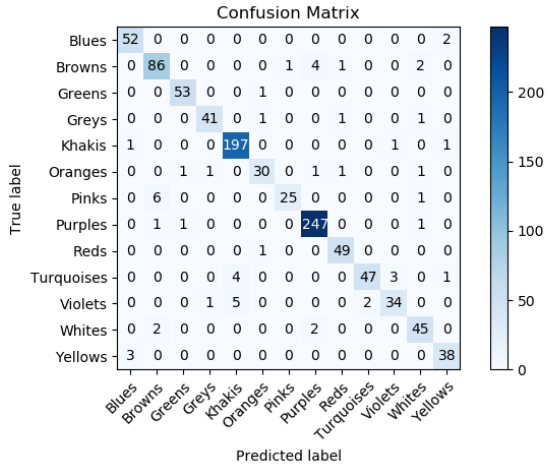
כפי שניתן לראות נוטה לסווג את המפלגות טורקיז וסיגל לבצע החאקי וגם נוטה לסווג את המפלגה הוורודה למפלגה החומה.

בסוף לאחר מציאת ההיפר-פרמטרים עבור כל אחד מהמסווגים, קבענו את הפרמטרים עבור כל מהמסווגים, עטפנו את כולם לכדי מסווג כולל, אשר בעת אימון מאמן את כל המסווגים שבתוכו וכאשר צריך לחזות יבצע את החיזוי על ידי החלת מניפולציה כלשהי על התחזיות שהניב כל אחד מהמסווגים.

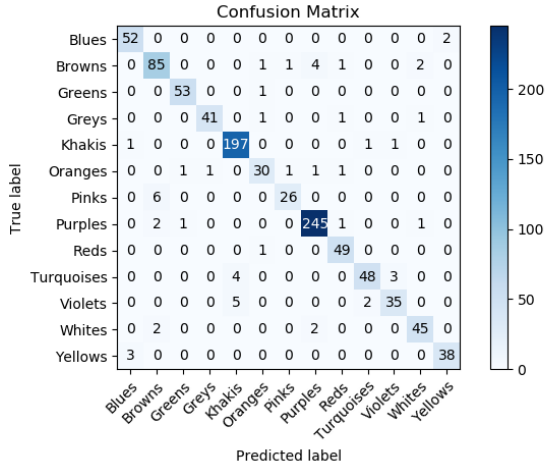
כעת נשאר להבין איך לכוון את המסווג שיבצע תחזיות, חשבנו על שתי דרכים:

1. כאשר יש רוב התחזית תוכרע על ידי הרוב כנהוג בוועדה, כאשר כל התחזיות של כל המסווגים שונות זו מזו, מסווג מאסטר ייקח את ההחלטה.
2. לחזות את ההסתברות לכל תחזית, למצע את ההסתברות עבור כל תחזית ולבסוף להחליט על פי התחזית עם ההסתברות הגדולה ביותר.

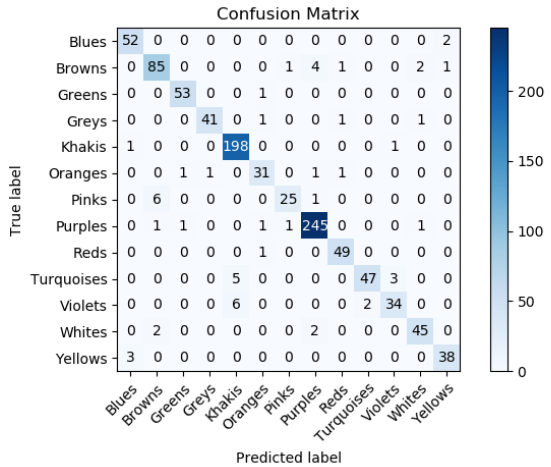
בחירת הדרך להכריע תחזית:
ראשית ניסנו את דרך 1 כאשר המסווג שמכריע כאשר אין רוב הוא ה - Random Forest, דיוק הוועדה על סט הוולידציה עמד על 94.4%.



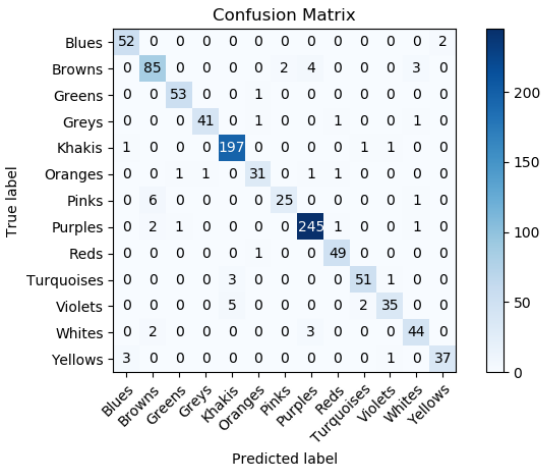
כאשר המסווג שמכריע כאשר אין רוב הוא ה -MLP, דיוק הוועדה על סט הוולידציה עמד על 94.4%.



כאשר המסווג שמכריע כאשר אין רוב הוא ה -SVM, דיוק הוועדה על סט הוולידציה עמד על 94.3%.



דרך 2: ביצוע התחזית על ידי שקלול ההסתברויות הניבה תוצאות דיוק של 94.4% על סט הוולידציה.



מסקנות ובחירת הדרך לביצוע תחזית:

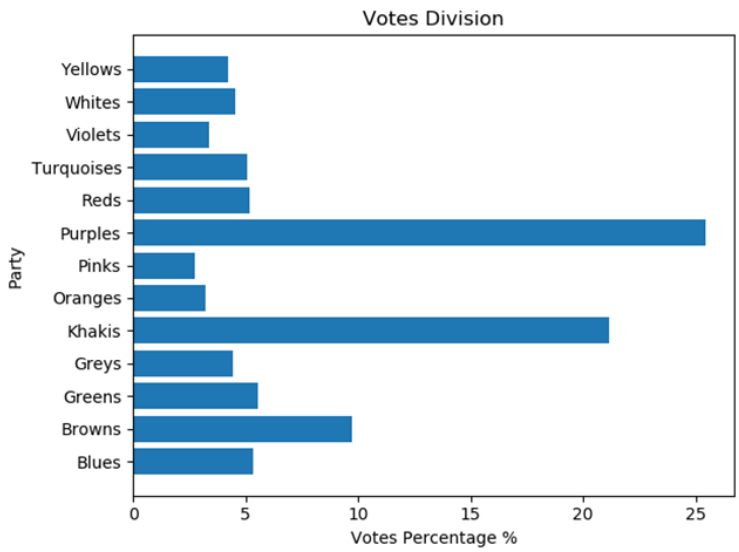
- ניתן לראות כי התופעות שראינו קודם עבור כל אחד מהמסווגים קטנו.
- כל הדרכים הגדילו את הדיוק של ביחס לכל מהמסווגים.
- יכולת הכללה של הוועדה שהרכבנו טובה יותר מכל אחד מהמסווגים באופן עצמאי.

לבסוף החלטנו לבחור בדרך הראשונה כאשר המסווג אשר מכריע את התחזית כשאין רוב הוא דווקא ה-Random Forest וזאת מכיוון שלמרות שהשיטה זו השיגה דיוק גבוהה כמו של האחרות המסווג עצמו הוא בעל אחוז דיוק בין ה MLP לSVM ומכיוון שאנו לא רוצים לסבול מ Overfitting או underfitting נבחר בשיטה זו.

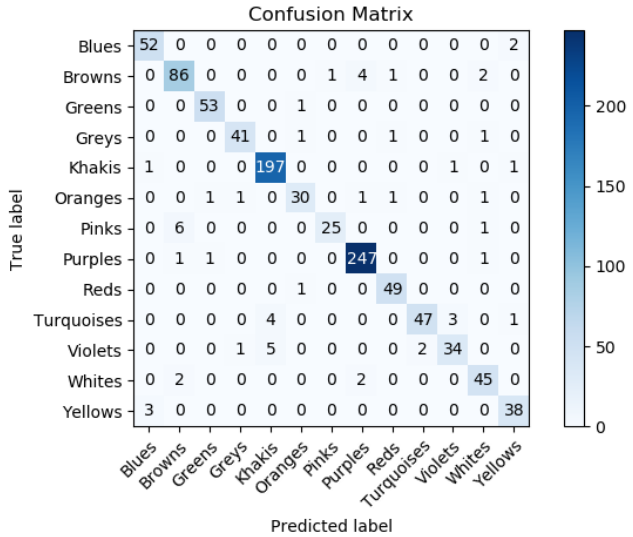
השוואת תוצאות המסווג על סט מבחן מתויג מתוך סט האימון אל מול סט המבחן לא מתויג:

סט מבחן מתויג:

עבור סט מבחן מתויג אשר המסווג הכולל לא התאמן עליו כלל ולא הסתמך עליו כלל הצלחנו להגיע לאחוז דיוק של 94.20%, אשר גבוה מאחוז הדיוק של כל מהמסווגים באופן עצמאי.
על מנת לחזות את המפלגת המנצחת לפי סט זה חזינו את כל ההצבעות והמפלגה המנצחת היא בעלת רוב הקולות, לפי סט מבחן (מתוך סט האימון הכולל) המפלגה המנצחת הינה הסגולים.
התפלגות הקולות לפי סט מבחן זה:



Color	Vote Percentage
Blues	5.32%
Browns	9.72%
Greens	5.56%
Greys	4.44%
Khakis	**21.16%**
Oranges	3.2%
Pinks	2.76%
Purples	**25.44%**
Reds	5.2%
Turquoises	5.08%
Violets	3.36%
Whites	4.52%
Yellows	4.24%



ניתן לראות כפי שראינו על סט הוולידציה כי התופעות שסבלו מהם כל אחד מהמסווגים באופן עצמאי קטנו.

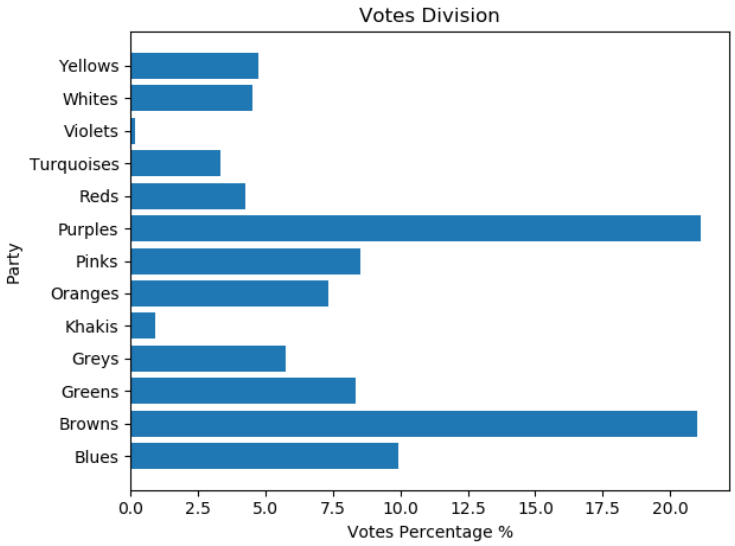
<u>סט מבחן לא מתויג (תחזיות אלה הן התחזיות להגשה):</u>

בעת כאשר בידנו סט מבחן לא מתויג אימנו את המסווג בשנית כל סט הדוגמאות מתחילת הקורס(כולל סט המבחן ששמנו בצד בתחילת הקורס) מכיוון שכעת סט המבחן אינו תלוי בסט האימון.

נשים לב כי כאשר אנו מאמנים את המסווג על כל סט האימון הרי שכעת המסווג שלנו השתנה, יכולת ההכללה שלו השתנתה ובמידה מסוימת איבדנו את הערכת הביצועים שעשינו קודם.

אמנם, המוטיבציה לבצע זאת היא מתוך הנחה כי כאשר המסווג שלנו מתאמן על יותר דוגמאות אזי יכולת ההכללה גדלה וכך נוכל להגיע לאחוז דיוק גבוהה יותר, בנוסף כאשר בחנו את המסווג קודם לכן אימנו אותו על 6,500 דוגמאות וסט המבחן אשר נבחן על פיו מכיל 2,500 דוגמאות לעומת זאת סט המבחן הלא מתויג הוא מכיל 10,000 דוגמאות ולכן נרצה שהמסווג שלנו יתאמן על יותר דוגמאות על מנת לשמור על הביצועים.

על מנת לחזות את המפלגת המנצחת לפי סט זה חזינו את כל ההצבעות והמפלגה בעלת רוב הקולות היא הסגולים **אמנם** בהפרש כה קטן של 11 קולות מהחומים דבר היכול להצביע כי בתוצאות האמתיות של הסט מתקיים ביניהם תיקו או שהחומים מנחים בהפרש קטן מאוד.



Color	Votes	Vote Percentage
Blues	994	9.94%
Browns	**2102**	**21.02%**
Greens	834	8.34%
Greys	576	5.76%
Khakis	93	0.93%
Oranges	734	7.34%
Pinks	853	8.53%
Purples	**2113**	**21.13%**
Reds	427	4.27%
Turquoises	335	3.35%
Violets	17	0.17%
Whites	452	4.52%
Yellows	473	4.73%

כפי שניתן לשים לב מהחיזוי עבור סט המבחן הלא מתויג המפלגה בעלת רוב הקולות היא עדיין הסגולים אמנם המפלגה שכעת במקום השני היא לא החאקי כמו שראינו בסט המבחן הקודם אלא דווקא החומים. מכיוון שלפי סט המבחן הקודם אחוז הדיוק של המסווג שלנו מוגדר כטוב מאוד ולא ראינו נטייה ברורה של המסווג שלנו להתבלבל בין קולות של המפלגה החומה לבין קולות של מפלגות אחרות ולכן למרות הסטייה הגדולה לטובת החומים שלא ראינו בסט האימון נניח כי תוצאות אלה הן אמינות.

לבסוף ייצאנו תוצאות אלו לקובץ CSV בשם "test_predictions.csv" כנדרש בתרגיל.

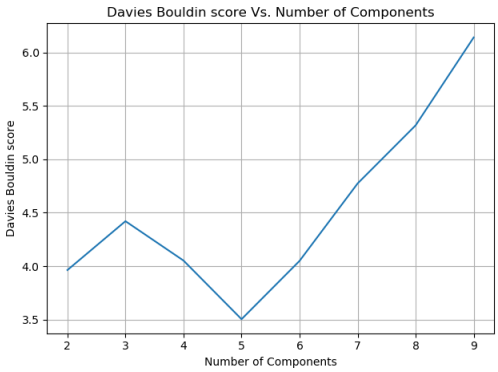
Building Steady Coalition Using Clustering Model:

לצורך בניית קואליציה בעזרת סט המבחן הלא מתויג החלטנו לשנות גישה מתרגיל בית 4 ולבחון צעדים חדשים להרכבת הקואליציה על מנת להיות יותר תואמים להגדרת קואליציה יציבה.

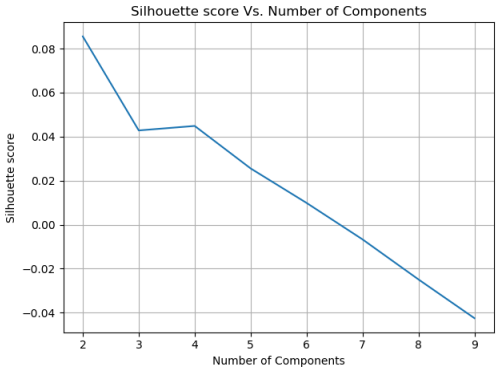
למרות השינוי בגישה עדיין בחרנו להשתמש באלגוריתם Gaussian Mixture אשר הינו מודל Clustering וזאת מכיוון שהניב תוצאות טובות בתרגיל בית קודם המוטיבציה להשתמש במודל שכזה היא היכולת לזהות דמיון בין המצביעים של המפלגות על ידי קיבוצם לתוך Clusters.

ראשית עלינו היה להבין עם כמה Clusters עלינו לאמן את המודל שלנו, זהו למעשה היפר-פרמטר של המודל אשר עלינו היה למצוא. לצורך מציאת הפרמטר אימנו את האלגוריתם עם מספר Clusters שונה ובחנו את ביצועיו. כל מודל אימנו בעזרת כל סט האימון מתחילת הקורס וזאת מכיוון שכעת אין משמעות לסיווג הדוגמאות (unsupervised learning) אלא יש יותר משמעות לכמות הדוגמאות שהמודל צריך להתמודד איתם ולמידת טיב הClustering.
את המודל הערכנו בעזרת המדדים הבאים:

<u>מדדים פנימיים:</u>
<u>Davies Bouldin score</u>- בוחן את הדמיון הממוצע בין שני Clusters הכי דומים, כאשר הדמיון הוא יחס בין המרחק בתוך הCluster למרחק ביניהם. נעדיף ערכים נמוכים.

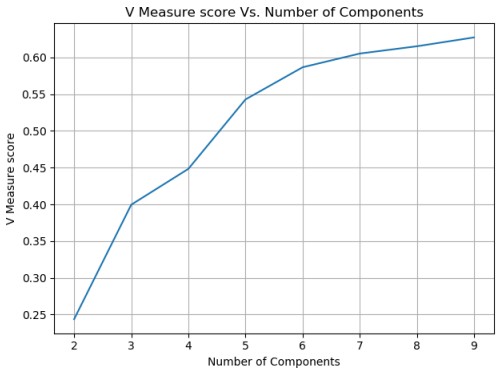


<u>Silhouette score-</u> בוחן עבור כל דוגמא כמה היא מתאימה לCluster שהיא שייכת. ערכיו הם בין 1 ל -1 ונעדיף ערכים קרובים ל1.



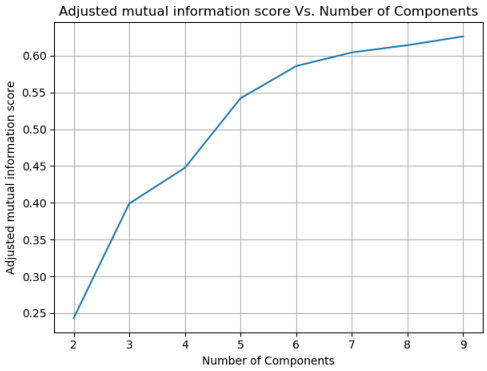
מדדים חיצוניים:

V Measure score- ממוצע הרמוני בין מדדי homogeneity ו completeness. כאשר מדד homogeneity מודד כמה כל Cluster נאמן לתיוג יחיד. מדד ב completeness מודד כמה כל תיוג נאמן לCluster יחיד. המדד מניב ערכים בין 0 ל1 כאשר נעדיף ערכים הקרובים ל1.



מדדי יציבות ויחס:

Adjusted Rand score- מודד את מידת ההסכמה בין שתי ריצות של האלגוריתם. בין הערכים 1- ל1 כאשר נעדיף ערכים הקרובים ל1.



לבסוף החלטנו לבנות את המודל שלנו עם 5 Clusters וזאת מכיוון שעבור מספר זה של Clusters השגנו תוצאות יפות בעבור כל המדדים.

הרכבת הקואליציה התבצעה באופן הבא:

לאחר בחירת מספר הClusters (5) אימנו מחדש את ה Gaussian Mixture עם סט המבחן הלא מתויג.

על מנת להבין את קווי הדמיון בין המפלגות עבור כל מפלגה ניתחנו את הפילוג שלה לפי הClusters כאשר התיוג של כל דוגמא מהסט הוא למעשה התחזית של הוועדה שבנינו.

Cluster	Blues	Browns	Greens	Greys	Khakis	Oranges	Pinks	Purples	Reds	Turquoises	Violets	Whites	Yellows
0	994	36	27	12	93	14	30	48	0	335	17	11	467
1	0	4	2	557	0	717	1	13	425	0	0	3	4
2	0	2062	2	2	0	2	227	2046	2	0	0	438	2
3	0	0	0	2	0	0	595	0	0	0	0	0	0
4	0	0	803	0	0	1	0	6	0	0	0	0	0

Party/Cluster	0	1	2	3	4
Blues	100%	0	0	0	0
Browns	1.7%	0.2%	98.1%	0	0
Greens	3.2%	0.24%	0.24%	0	96.3%
Greys	2.1%	97.2%	0.35%	0.35%	0
Khakis	100%	0	0	0	0
Oranges	1.9%	97.7%	0.27%	0	0.13%
Pinks	3.5%	0.12%	26.6%	69.7%	0
Purples	2.27%	0.61%	96.8%	0	0.28%
Reds	0	99.54%	0.46%	0	0
Turquoises	100%	0	0	0	0
Violets	100%	0	0	0	0
Whites	2.4%	0.66%	96.9%	0	0
Yellows	98.7%	0.85%	0.42%	0	0

cluster 0 distribution: ([('Blues', 994), ('Browns', 36), ('Greens', 27), ('Greys', 12), ('Khakis', 93), ('Oranges', 14), ('Pinks', 30), ('Purples', 48), ('Turquoises', 335), ('Violets', 17), ('Whites', 11), ('Yellows', 467)])

cluster 1 distribution: ([('Browns', 4), ('Greens', 2), ('Greys', 557), ('Oranges', 717), ('Pinks', 1), ('Purples', 13), ('Reds', 425), ('Whites', 3), ('Yellows', 4)])

cluster 2 distribution: ([('Browns', 2062), ('Greens', 2), ('Greys', 2), ('Oranges', 2), ('Pinks', 227), ('Purples', 2046), ('Reds', 2), ('Whites', 438), ('Yellows', 2)])

cluster 3 distribution: ([('Greys', 2), ('Pinks', 595)])

cluster 4 distribution: ([('Greens', 803), ('Oranges', 1), ('Purples', 6)])

Browns dist: [(0, 0.017126546146527116), (1, 0.0019029495718363464), (2, 0.9809705042816366)]

Greens dist: [(0, 0.03237410071942446), (1, 0.002398081534772182), (2, 0.002398081534772182), (4, 0.9628297362110312)]

Pinks dist: [(0, 0.035169988276670575), (1, 0.0011723329425556857), (2, 0.2661195779601407), (3, 0.6975381008206331)]

Purples dist: [(0, 0.022716516800757217), (1, 0.006152389966871746), (2, 0.9682915286322764), (4, 0.002839564600094652)]

Oranges dist: [(0, 0.01907356948228883), (1, 0.9768392370572208), (2, 0.0027247956403269754), (4, 0.0013623978201634877)]

Blues dist: [(0, 1.0)]

Reds dist: [(1, 0.9953161592505855), (2, 0.00468384074941452)]

Yellows dist: [(0, 0.9873150105708245), (1, 0.008456659619450317), (2, 0.004228329809725159)]

Whites dist: [(0, 0.024336283185840708), (1, 0.00663716814159292), (2, 0.9690265486725663)]

Turquoises dist: [(0, 1.0)]

Greys dist: [(0, 0.020942408376963352), (1, 0.9720767888307156), (2, 0.0034904013961605585), (3, 0.0034904013961605585)]

Khakis dist: [(0, 1.0)]

Violets dist: [(0, 1.0)]

ניתן לשים לב לתוצאה מדהימה – כל מפלגה חוץ מהוורודים, שויכה לCluster בודד עד כדי 5% מקולות המפלגה, מכיוון שאין מדובר בחלוקה ל2 Clusters (וגם שם תוצאה שכזו הייתה מפתיעה) הגענו למסקנה שהמסווג הנבחר מבצע עבודה טובה מאוד.
לאחר הסתכלות על ההתפלגות של אחת מהמפלגות קיבצנו אותם לפי בלוקים בעלי התפלגות דומה בין ה-Clusters, המוטיבציה לכך היא לזהות דמיון בין המפלגות ולהרכיב קואליציה בעלת קווי דיון בן המצביעים של המפלגות החברות בה. (הבלוקים הנ"ל אינם Clusters שהאלגוריתם יצר אלא נגזרים מהם).אחוז הקולות המשוערים של כל אחד מהבלוקים:

Block 0: Blues, Khakis, Violets, Turquoises, Yellows -> 19.12% of the votes

Block 1: Reds, Oranges, Greys-> 17.34% of the votes

Block 2: Browns, Whites, Purples-> 46.67% of the votes

Block 3: Pink-> 8.53% of the votes

Block 4: Greens-> 8.34% of the votes

נשים לב כי אף אחד מהבלוקים אינו יכול לבנות קואליציה לבדו ולכן נצטרך למצוא חיבור של בלוקים זאת על מנת להרכיב קואליציה.
נבחין כי קואליציה קטנה יותר תקיים יותר הומוגניות ויותר יציבות (כמו בפוליטיקה במציאות) ולכן על מנת לשמור על הומוגניות הקואליציה ננסה להרכיב כזאת בעזרת בלוק 2 אשר מכיל את מספר הקולות הגדול ביותר ביחס לכמות המפלגות שבו.
נרצה למצוא את הבלוק שהכי קרוב אליו, לשם כך חשבנו עבור כל אחד מהבלוקים את "מרכז המסה" שלו ואת המרחק מכל אחד מהבלוקים האחרים:

	Block 0	Block 1	Block 2	Block 3	Block 4
Block 0	-	2.54	2.37	2.52	1.99
Block 1	2.54	-	0.73	0.54	1.61
Block 2	2.37	0.74	-	0.9	1.07
Block 3	2.52	0.54	0.9	-	1.63
Block 4	1.99	1.61	1.07	1.63	-

לפי טבלה זו בלוק 2 הכי קרוב לבלוק 1 ולאחר מכן לבלוק 3.
אמנם מכיוון שבבלוק 1 יש יותר מפלגות מבלוק 3 וההפרש בקרבה אינו משמעותי נבחר בבלוק 3 כי מכיל פחות מפלגות ומקיים יותר את הגדרת קואליציה יציבה.

<u>לבסוף הקואליציה שהרכבנו מורכבת מהמפלגות Purples, Pinks, Browns, Whites וכוללת 55.2% מהקולות.</u>

ולסיום, תודה אישית על סמסטר מעניין ומלמד :