Putting It All Together

על מנת לענות על השאלות שנשאלו תכננו לבצע את התהליך הבא:

להשתמש בסט המתוייג על מנת למצוא מסווג מתאים לתיוג הסט הלא מתוייג, לאחר בחירת המסווג, לאמן אותו על כל הסט המתוייג ולסווג את הלא מתוייג.

לאחר שיש בידנו את התיוגים שחזינו, נסיק מהן את ההתפלגות והמנצחת, ובעזרתן נמצא קואליציה יציבה.

Preprocessing

ביצענו את תהליך העיבוד המקדים על קבוצת התכונות הרלוונטיות.

Filling missing values

כיוון שלסט המתוייג יש 2 מטרות (מציאת מסווג אופטימלי ואימונו) יצרנו עותק נוסף שלו.

בחרנו להשלים את המידע החסר בעותק שישמש לאימון בשיטת Closest Fit תוך התייחסות לתיוג של הדוגמאות כיוון שהנחנו שאנשים המצביעים לאותן מפלגות ודומים ברוב תכונותיהם יהיו דומים גם בתכונות החסרות.

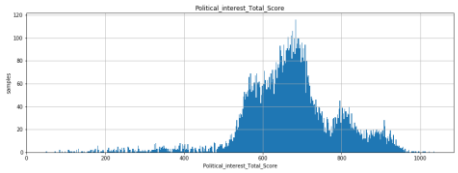
עבור הסט הלא מתוייג בחרנו להשלים את הערכים החסרים בשיטת החציון בתקווה שזה לא ישפיע על שונות התכונות בצורה ניכרת מכיוון שהן קטנות יחסית. את החציון לקחנו מהעותק של הסט המתוייג.

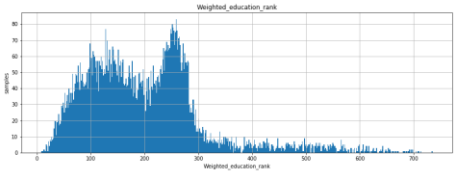
Data Transformation

מתכונות הבדידות שאין בהן סדר יצרנו תכונות חדשות בשיטת One-Hot.

את הנרמול ביצענו בשיטות הבאות (באופן זהה לתרגיל 2):

- Z-score עבור התכונות שראינו בבחינת המידע שלפילוג שלהן יש 'זנבות' ארוכים, כדי שמרכז המסה של ההתפלגות תהיה בקטע $[-1,1]$. למשל עבור התכונות הבאות:





- Min-Max עבור שאר התכונות, גם לקטע $[-1,1]$.

בנוסף, הוספנו תכונה המהווה אינדיקטור למידע שהיה חסר במקור.

בחירת המודל

המודלים אותם בחנו

קודם ביצענו כיוונון פרמטרים בעזרת CV על קבוצת האימון עבור המסווגים: id3, knn, forest, svm.

מהטובים ביותר עבור כל סוג מהמודלים האלו יצרנו מודל מורכב שהוא ועדה (voting) והפרמטרים שכיוונו הם הרכב הועדה וגודלה.

באופן זה אנו מחפשים היפוטזה מתאימה במרחב גדול ומגוון של מסווגים אפשריים.

אופן מדידת ביצועי המסווגים בתהליך התבצע באופן זהה כפי שמותאר בהמשך.

הערכת הביצועים של המודלים

מכיוון שהסט המתוייג אינו בהכרח מייצג את ההתפלגות האמיתית של המפלגות, נרצה לקבל recall גבוה עבור כל מפלגה בפני עצמה. בצורה זו אנו נמצא מסווג אשר ינסה ללמוד כל אחת מהמלפלגות בצורה מספקת, כך אם יש מפלגה שבסט המתוייג היא קטנה ובסט החדש היא גדולה, אז כמות הקולות עבור מפלגה זו תהייה יותר מייצגת, ולכן ההיסטוגרמה תהיה פחות רגישה לסטיות גדולות.

על מנת לבצע זאת באופן שיוויוני, מדדנו את איכות המסווגים ע"פ ה-recall המינימלי שהתקבל מכל אחת מהמפלגות (שורות במטריצת הבלבול).

אופן בחירת המודל המנצח

בחרנו את המודל בעל הביצועים הטובים ביותר מבין אלו שבחנו כמו שתואר מעלה, כאשר הבדיקה נעשתה על קבוצת הוולידציה. לאחר בחירה זו ביצענו בדיקת שפיות על קבוצת המבחן כדי לראות שהמסווג אכן נותן תוצאות סבירות והצגנו את הדיוק והשגיאה של מסווג זה.

```
Min Recall: 0.795918367347
Accuracy: 0.934666666667
Error: 0.0653333333333
```

חיזוי הצבעות ומנצחת

את המסווג הנבחר אימנו על עותק הסט המתוייג (כפי שתיארנו בשלב העיבוד המקדים) כדי למצות את כל המידע המתויג שיש ברשותנו ללמידה. לאחר מכן ביצענו חיזוי של הדוגמאות הלא מתוייגות, שמרנו אותן בקובץ csv כנדרש, ניתחנו את התפלגות ההצבעות ועל פיה הסקנו את המפלגה המצחת:

```
The predicted distribution: {
'Blues': 0.0965,
'Pinks': 0.0942,
'Purples': 0.1928,
'Reds': 0.0422,
'Greens': 0.0843,
'Browns': 0.2119,
'Yellows': 0.0513,
'Turquoises': 0.0445,
'Oranges': 0.0798,
'Greys': 0.0613,
'Whites': 0.0412
}
The predicted winner: Browns
```

נשים לב, כי מכיוון שיש לנו שגיאה של כ-6% ייתכן כי המנצח תהיה 'Purples' מכיוון שההפרש בינה לבין המנצחת הוא כ-2%.

בחירת קואליציה יציבה

Clustering

על מנת למצוא קואליציה עבור מפלגות בעלות מצביעים דומים, השתמשנו ב-GaussianMixture כדי לחלק את הדאטה לקלאסטרים. כיוונו את פרמטר מספר הקלאסטרים, החל מחצי מספר המפלגות ועד לפי שניים ממספר המפלגות, על מנת לאפשר גמישות מירבית, את טיב התוצאות מדדנו באמצעות כרוס-ולידישן עבור:

- Calinski Harabaz score
- Silhouette score

השתמשנו בציונים אלו על מנת לקבל קלאסטרים שהם יחסית צפופים ורחוקים אחד מהשני.

```
--------------------
Number of components: 9
Calinski Harabaz score: 769.237676253
Silhouette score: 0.257786885483
--------------------
```

על פי התוצאות בחרנו לבחון את הקואליציות עבור 9 קלאסטרים (בפועל בדקנו עבור מספרים שונים כדי לחזק את ההערכות שלנו).

יצרנו היסטוגרמת הצבעות עבור כל קלאסטר:

```
----------------------------------
{'Browns': 434, 'Blues': 4, 'Purples': 393, 'Greens': 177, 'Whites': 79, 'Pinks': 61, 'Yellows': 7}
Cluster  0
Size =  1155
Browns  =  38 %
Blues  =  0 %
Purples  =  34 %
Greens  =  15 %
Whites  =  7 %
Pinks  =  5 %
Yellows  =  1 %
----------------------------------
{'Purples': 400, 'Turquoises': 5, 'Greens': 181, 'Browns': 434, 'Pinks': 64, 'Whites': 84, 'Blues': 5, 'Yellows': 1}
Cluster  1
Size =  1174
Purples  =  34 %
Turquoises  =  0 %
Greens  =  15 %
Browns  =  37 %
Pinks  =  5 %
Whites  =  7 %
Blues  =  0 %
Yellows  =  0 %
----------------------------------
{'Greens': 167, 'Purples': 395, 'Pinks': 74, 'Whites': 93, 'Browns': 380, 'Blues': 5, 'Turquoises': 2, 'Yellows': 8}
Cluster  2
Size =  1124
Greens  =  15 %
Purples  =  35 %
Pinks  =  7 %
Whites  =  8 %
Browns  =  34 %
Blues  =  0 %
Turquoises  =  0 %
Yellows  =  1 %
----------------------------------
{'Reds': 422, 'Oranges': 403, 'Greys': 305, 'Greens': 2, 'Purples': 8, 'Blues': 1, 'Yellows': 3, 'Whites': 2, 'Browns': 7, 'Turquoises': 3, 'Pinks': 3}
Cluster  3
Size =  1159
Reds  =  36 %
Oranges  =  35 %
Greys  =  26 %
Greens  =  0 %
Purples  =  1 %
Blues  =  0 %
Yellows  =  0 %
Whites  =  0 %
Browns  =  1 %
Turquoises  =  0 %
Pinks  =  0 %
----------------------------------
```

```
-----------------------------------
{'Yellows': 5, 'Browns': 430, 'Pinks': 76, 'Purples': 348, 'Whites': 66, 'Greens': 153, 'Blues': 9, 'Turquoises': 2}
Cluster  4
Size =  1089
Yellows  =  0 %
Browns  =  39 %
Pinks  =  7 %
Purples  =  32 %
Whites  =  6 %
Greens  =  14 %
Blues  =  1 %
Turquoises  =  0 %
-----------------------------------
{'Pinks': 586, 'Purples': 4, 'Whites': 1}
Cluster  5
Size =  591
Pinks  =  99 %
Purples  =  1 %
Whites  =  0 %
-----------------------------------
{'Blues': 937, 'Yellows': 483, 'Turquoises': 430}
Cluster  6
Size =  1850
Blues  =  51 %
Yellows  =  26 %
Turquoises  =  23 %
-----------------------------------
{'Oranges': 395, 'Greys': 308, 'Purples': 1}
Cluster  7
Size =  704
Oranges  =  56 %
Greys  =  44 %
Purples  =  0 %
-----------------------------------
{'Purples': 379, 'Browns': 434, 'Pinks': 78, 'Whites': 87, 'Greens': 163, 'Turquoises': 3, 'Yellows': 6, 'Blues': 4}
Cluster  8
Size =  1154
Purples  =  33 %
Browns  =  38 %
Pinks  =  7 %
Whites  =  8 %
Greens  =  14 %
Turquoises  =  0 %
Yellows  =  1 %
Blues  =  0 %
```

ומטריצת מרחקים בין תוחלות הקלאסטרים:

```
[[ 0.          0.67408951  0.65918242  0.6228881   0.66446979  0.87641738  0.80651234  0.71049464  0.65477411]
 [ 0.67408951  0.          0.65846879  0.63580626  0.6553892   0.88436925  0.79521323  0.73833633  0.66591832]
 [ 0.65918242  0.65846879  0.          0.62355725  0.6546736   0.87885399  0.79128177  0.72006394  0.65541706]
 [ 0.6228881   0.63580626  0.62355725  0.          0.62471231  0.8617359   0.69269429  0.57021828  0.62040604]
 [ 0.66446979  0.6553892   0.6546736   0.62471231  0.          0.88097098  0.78600186  0.72346219  0.65863447]
 [ 0.87641738  0.88436925  0.87885399  0.8617359   0.88097098  0.          1.          0.92822106  0.87609487]
 [ 0.80651234  0.79521323  0.79128177  0.69269429  0.78600186  1.          0.          0.76983817  0.79383077]
 [ 0.71049464  0.73833633  0.72006394  0.57021828  0.72346219  0.92822106  0.76983817  0.          0.71174685]
 [ 0.65477411  0.66591832  0.65541706  0.62040604  0.65863447  0.87609487  0.79383077  0.71174685  0.        ]]
```

ביצענו חלוקה של המפלגות לקואליציה ואופוזיציה, עבור הרכבים נבחרים, ובדקנו את ציון ההמוגוניות של הקלאסטרים עבורם:

```
0.928184513677 ['Purples', 'Greens', 'Browns', 'Whites', 'Pinks']
0.764708226911 ['Purples', 'Greens', 'Browns', 'Whites']
0.742564554802 ['Purples', 'Greens', 'Browns', 'Pinks']
0.642290793234 ['Purples', 'Greens', 'Browns']
```

(ציון, הרכב הקואליציה)

בחנו את התוצאות והגענו למסקנה כי הקואלציה: Purples, Greens, Browns, Whites היא העדיפה ביותר.

לקואלציה זו צפוי 53.02 אחוזים מהקולות והיא מתפרסת על פני מספר קלאסטרים יחסית קרובים מה שמצביע על כך שהמצביעים למפלגות אלו יחסית דומים. כמו כן מפלגות האופוזיציה אינן חולקות בעיקרן את הקלסטרים עם הקואליציה ומרחק הקלסטרים בהן הן נמצאות רחוק יחסית מאלו של הקואליציה.

למרות שקיבלנו על סט המבחן כ-6% שגיאה, אנו מצפים שלקואלציה זו עדיין יהיה רוב, זאת כיוון שאנו מניחים שהשגיאה נובעת מסיווג לא נכון של מצביעים, שגיאה כזו כנראה תהיה בתוך קלאסטר מכיוון שמצבעים של מפלגות אלו דומים. לכן, מכיוון שבחרנו קואלציה הומוגנית יש סיכוי טוב שהשגיאות בתיוג יהיו למפלגה אחרת בקואלציה. ובאופן דומה באופוזיציה.

התלבטנו אם להוסיף את Pink לחיזוק הקואלציה אך לבסוף החלטנו שלא מכיוון שלמרות שכ-30% מהמצביעים למפלגה זו הינם בעלי תכונות דומות מאוד לקואלציה לעיל, החלק האחר נמצא בקלסטר נפרד המרוחק משאר הקלאסטרים, מה שמעיד על השונות בתכונות של מצביעים אלו, ובנוסף אינם מהווים אחוז ניכר בקלסטרים של הקואלציה לכן אי צירופם לקוליציה לא יפגע יותר מידי בשונות התכונת המצבעים ממצביעי האופוזיציה ולכן יציבות הקואליציה לא תפגע.

מגישים

אביאל שמחי, 305376063

עמית סולומון, 305025785