



5

68

30/06/19

עידו יחזקאל

אמיר אביבי

idoye

saavivi

2 0 4 3 9 7 3 6 8

3 0 5 1 8 3 8 7 3

מבוא למערכות לומדות

236756

סמסטר אביב תשע"ט

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:

|  |  |
| --- | --- |
| 1. Percentage from Original Data Set | Data Set |
| 65% | Train set |
| 10% | Validating set |
| 25% | Test set |

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

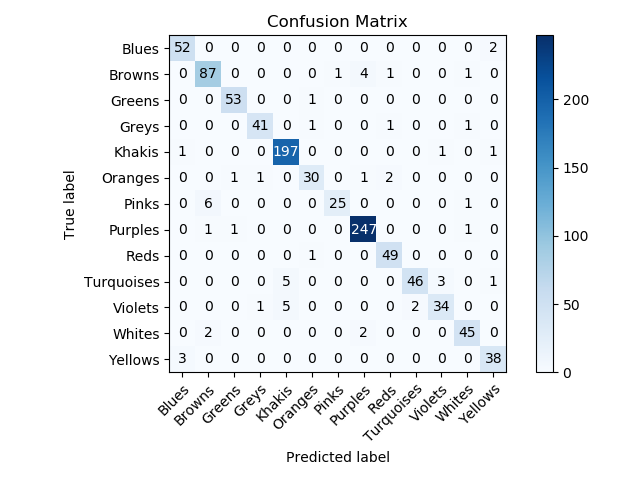
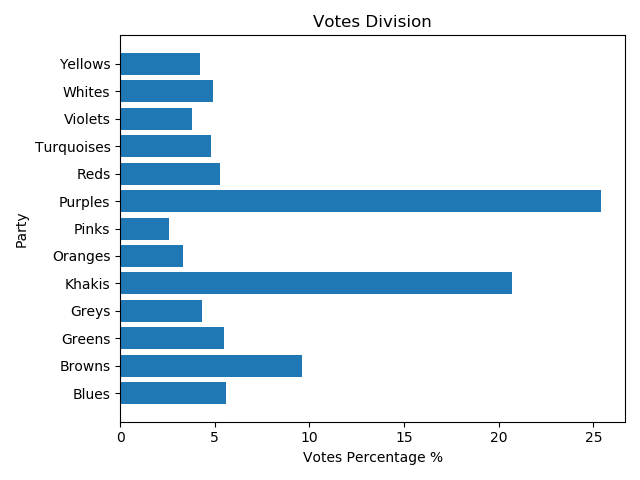
כדי לבצע את שלושת התחזיות הרכבנו עטפנו 3 מסווגים שונים אשר הניבו מדד דיוק הכי טוב מתרגלי בית קודמים, המסווגים הם:

1. RandomForestClassifier**(**random\_state**=**0**,** criterion**=**'entropy'**,** min\_samples\_split**=**4**,** min\_samples\_leaf**=**1**,** n\_estimators**=**450**)**
2. MLPClassifier**(**hidden\_layer\_sizes**=(**150**,** 10**),** activation**=**'relu'**,** solver**=**'lbfgs'**,** alpha**=**0.001**,** batch\_size**=**'auto'**,** learning\_rate**=**'adaptive'**,** learning\_rate\_init**=**0.001**,** power\_t**=**0.5**,** max\_iter**=**1000**,** shuffle**=True,** random\_state**=**0**,** tol**=**0.0001**,** verbose**=False,** warm\_start**=True,** momentum**=**0.9**,** nesterovs\_momentum**=True,** early\_stopping**=True,** validation\_fraction**=**0.1**,** beta\_1**=**0.9**,** beta\_2**=**0.999**,** epsilon**=**1e-08**,** n\_iter\_no\_change**=**10**)**
3. SVC**(**C**=**150**,** kernel**=**'poly'**,** degree**=**3**,** random\_state**=**0**)**

את ההיפר פרמטרים עבור כל אחד המסווגים בחרנו על פי ביצועי המסווגים לפי תרגילי בית קודמים, כאשר כל אחד מהם ממקסם את מדד הדיוק על הסט הוולידציה (מבלי לראות את סט המבחן המתויג)

את החלטת המסווג הכולל מימשנו על פי החלטת הרוב וכאשר כל אחד מהמסווגים סיווג בצורה שונה את הדוגמא אזי נלקח הסיווג של מסווג מספר 1 אשר בעל הדיוק המקסימלי.

**ביצועים על סט וולידציה:**דיוק של 94.4%, מפלגה מנצחת: סגולים.

ללא סטיות בתחזיות יוצאות דופן.  


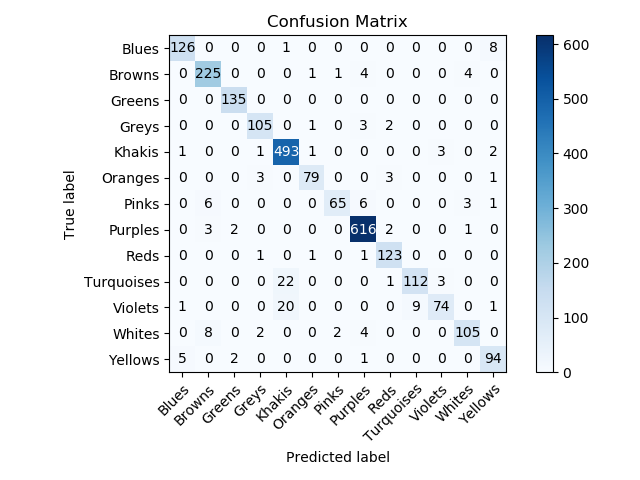
**השוואת תוצאות המסווג על סט מבחן מתויג מתוך סט האימון אל מול סט המבחן לא מתויג:**

**סט מבחן מתויג:**

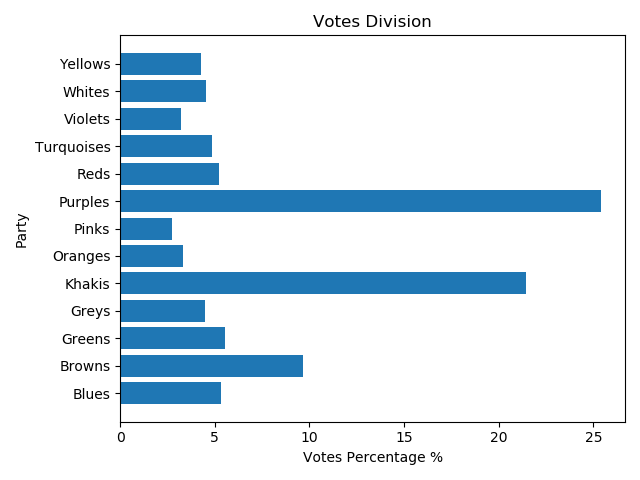
עבור סט מבחן מתויג אשר המסווג הכולל לא האתמן עליו כלל ולא הסתמך עליו כלל הצלחנו להגיע לאחוז דיוק של 94.08% .  
על מנת לחזות את המפלגת המנצחת לפי סט זה חזינו את כל ההצבעות והמפלגה המנצחת היא בעלת רוב הקולות, לפי סט מבחן (מתוך סט האימון הכולל) המפלגה המנצחת הינה הסגולים.

התפלגות הקולות לפי סט מבחן זה:

|  |  |
| --- | --- |
| Vote Percentage | Color |
| 5.32% | Blues |
| 9.68% | Browns |
| 5.56% | Greens |
| 4.48% | Greys |
| **21.44%** | **Khakis** |
| 3.32% | Oranges |
| 2.72% | Pinks |
| **25.4%** | **Purples** |
| 5.24% | Reds |
| 4.84% | Turquoises |
| 3.2% | Violets |
| 4.52% | Whites |
| 4.28% | Yellows |



המסווג שלנו בעיקר מתבלבל בחיזוי צבעי טורקיז וסיגל לצבע החאקי ולכן אנו נרצה לשים דגש על צבעים אלו בסט המבחן הלא מתויג.

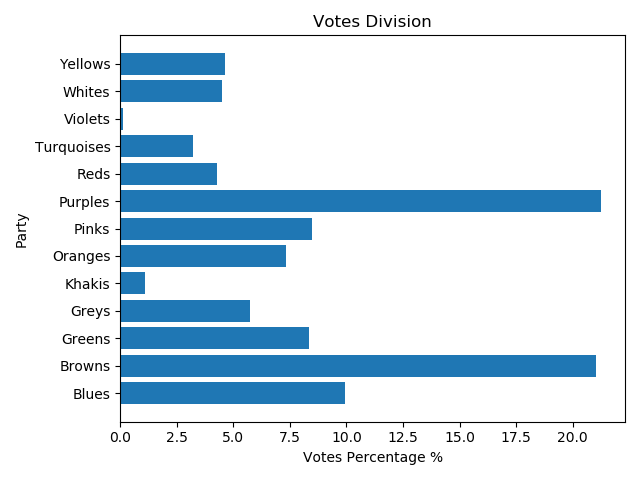
****

**סט מבחן לא מתויג (תחזיות אלה הן התחזיות להגשה):**

כעת כאשר בידנו סט מבחן לא מתויג אימנו את המסווג בשנית כל סט האימון מתחילת הקורס מכיוון שכעת סט המבחן אינו תלוי בסט האימון כלל.

על מנת לחזות את המפלגת המנצחת לפי סט זה חזינו את כל ההצבעות והמפלגה בעלת רוב הקולות היא הסגולים אמנם בהפרש כה קטן של 24 קולות או 0.24% לא ניתן לקבוע באופן חד משמעית כי היא בעלת ניצחון מובהק בבחירות. לכן נוכל להגיד כי קיים תיקו בין הסגולים לחומים עם יתרון קל לסגולים.  
התפלגות הקולות לפי סט מבחן זה:

|  |  |  |
| --- | --- | --- |
| Vote Percentage | Votes | Color |
| 9.92% | 992 | Blues |
| **21.01%** | **2101** | **Browns** |
| 8.36% | 836 | Greens |
| 5.74% | 574 | Greys |
| 1.1% | 110 | Khakis |
| 7.32% | 732 | Oranges |
| 8.5% | 850 | Pinks |
| **21.25%** | **2125** | **Purples** |
| 4.3% | 430 | Reds |
| 3.21% | 321 | Turquoises |
| 0.14% | 14 | Violets |
| 4.5% | 450 | Whites |
| 4.65% | 465 | Yellows |



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: