



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 נדרשנו לחזות:

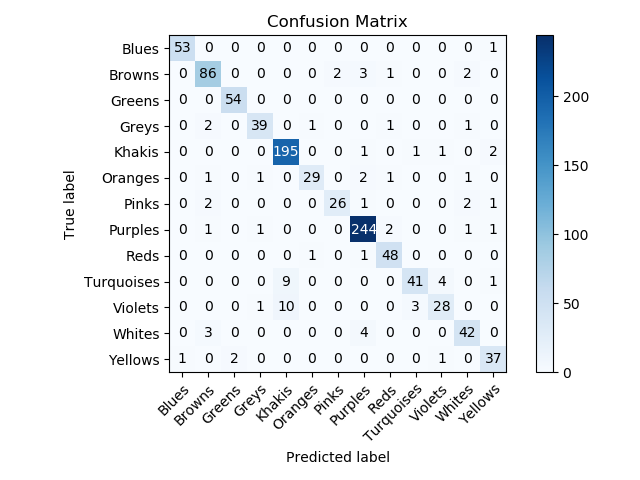
* מה היא המפלגה המנצחת לפי סט המבחן.
* מה הוא פילוח הקולות לפי סט המבחן.
* עבור כל מצביע מסט המבחן לחזות את הצבעתו.
* לחזות קואליציה יציבה והומוגנית הכוללת לפחות 51% מסך כל הקולות.

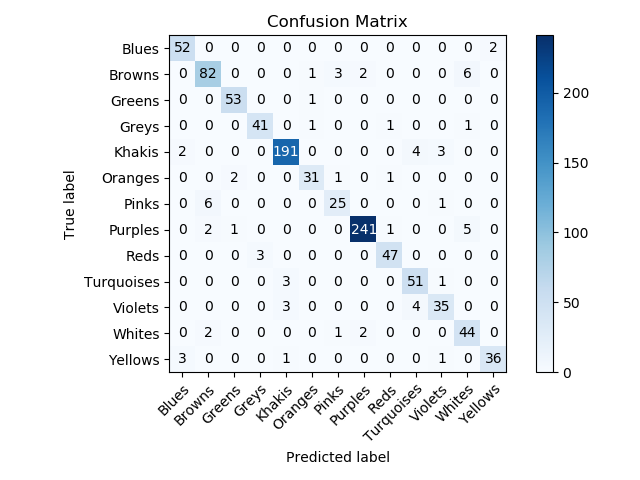
Mandatory Part – Voting Predictions:

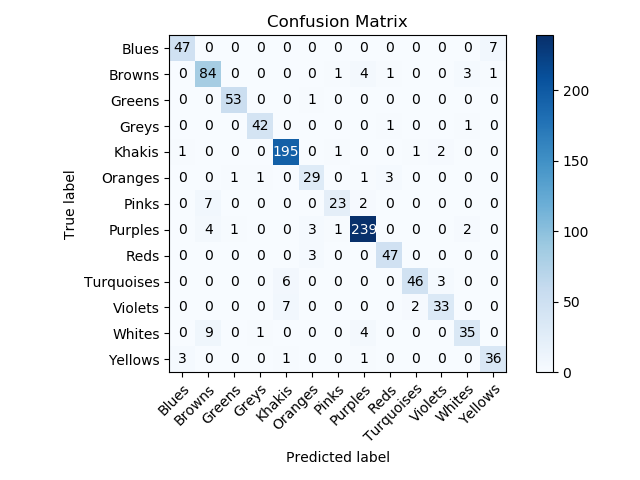
אנו מבינים שעל מנת לבצע את התחזיות הנ"ל על סט המבחן הלא מתויג אנו חייבים לבנות מסווג בעל יכולת הכללה גבוהה שתביא דיוק גבוה וזאת מכיוון שאין אנו יכולים להעריך את ביצועי המסווג על סט המבחן הלא מתויג.  
לכן על מנת להרכיב מסווג עם יכולת הכללה טובה החלטנו להרכיב ועדה הכוללת את שלושת המסווגים הבולטים שראינו 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.

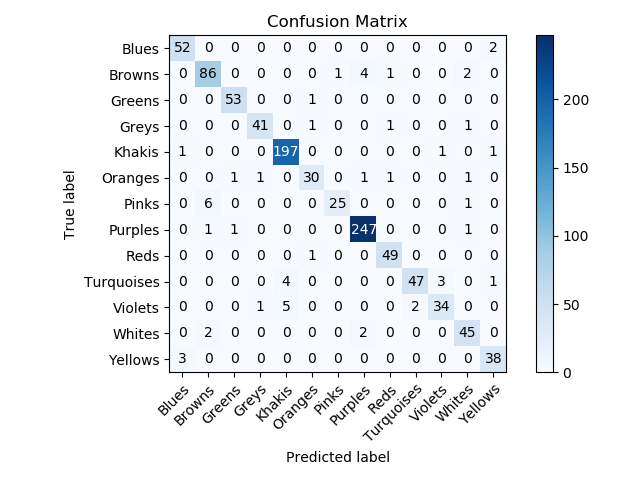
MLP- הניב דיוק של 92.9% על סט הוולידציה.   
  
כפי שניתן לראות נוטה לסווג את המפלגה הוורודה למפלגה החומה.

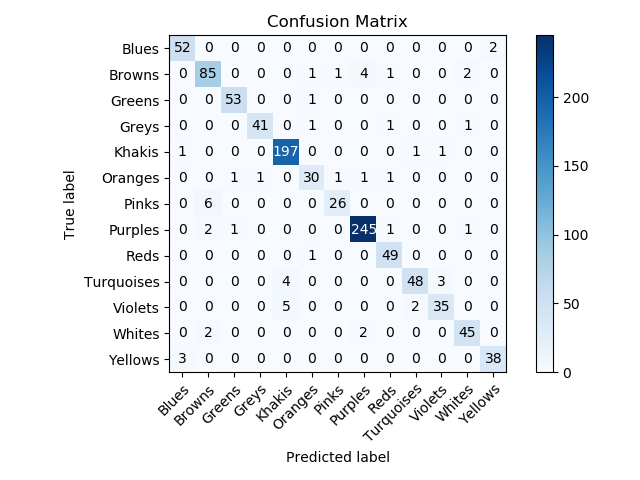
SVM- הניב דיוק של 90.9% על סט הוולידציה.   
  
כפי שניתן לראות נוטה לסווג את המפלגות טורקיז וסיגל לבצע החאקי וגם נוטה לסווג את המפלגה הוורודה למפלגה החומה.

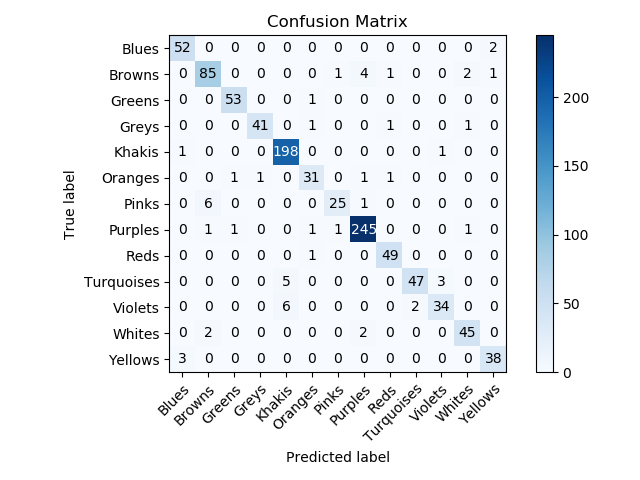
בסוף לאחר מציאת ההיפר-פרמטרים עבור כל אחד מהמסווגים, קבענו את הפרמטרים עבור כל מהמסווגים, עטפנו את כולם לכדי מסווג כולל, אשר בעת אימון מאמן את כל המסווגים שבתוכו וכאשר צריך לחזות יבצע את החיזוי על ידי החלת מניפולציה כלשהי על התחזיות שהניב כל אחד מהמסווגים.

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

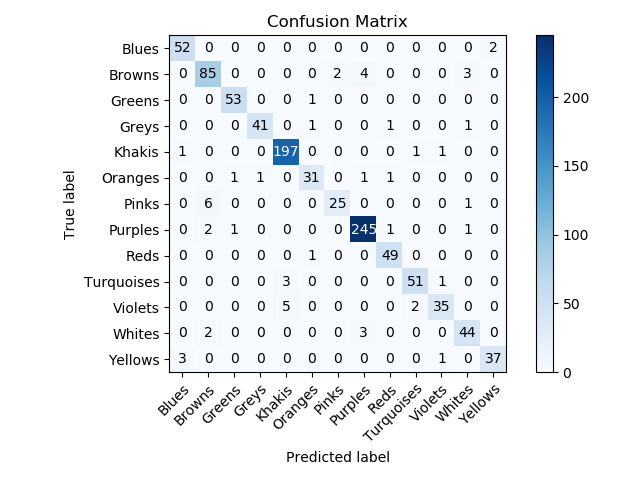
1. כאשר יש רוב התחזית תוכרע על ידי הרוב כנהוג בוועדה, כאשר כל התחזיות של כל המסווגים שונות זו מזו, מסווג מאסטר ייקח את ההחלטה.
2. לחזות את ההסתברות לכל תחזית, למצע את ההסתברות עבור כל תחזית ולבסוף להחליט על פי התחזית עם ההסתברות הגדולה ביותר.

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.

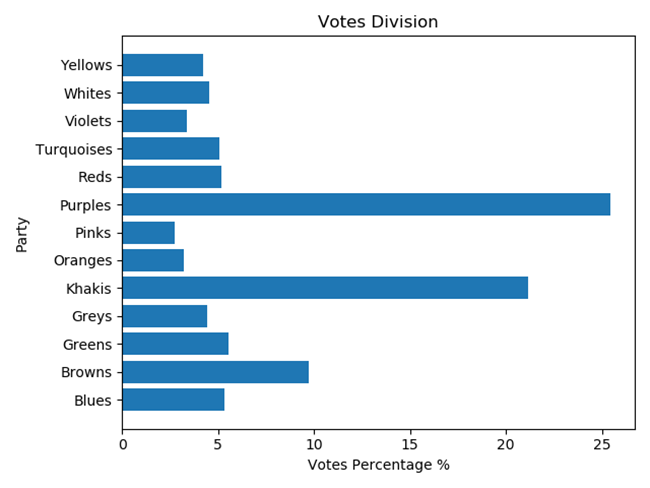
**השוואת תוצאות המסווג על סט מבחן מתויג מתוך סט האימון אל מול סט המבחן לא מתויג:**

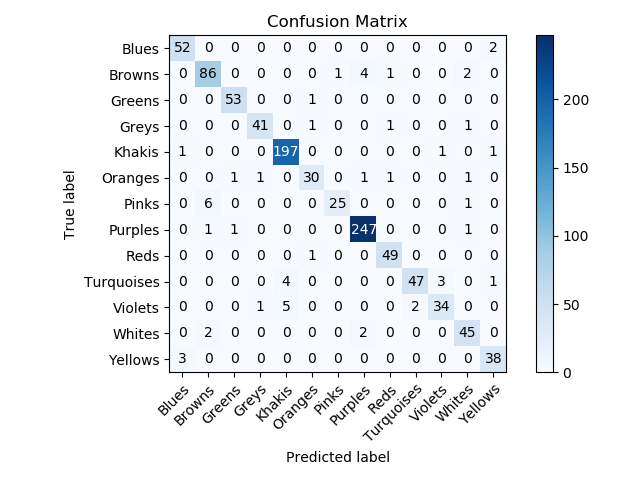
**סט מבחן מתויג:**

עבור סט מבחן מתויג אשר המסווג הכולל לא האתמן עליו כלל ולא הסתמך עליו כלל הצלחנו להגיע לאחוז דיוק של 94.20%, גבוהה 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:

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



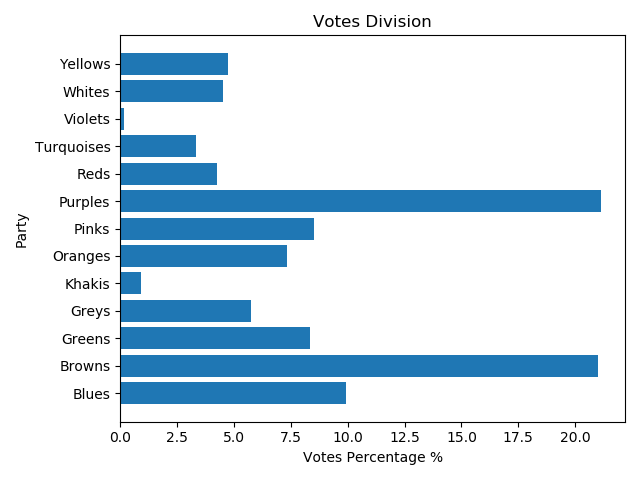


ניתן לראות כפי שראינו על סט הוולידציה כי התופעות שסבלו מהם כל אחד מהמסווגים באופן עצמאי קטנו.

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

כעת 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.



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