מבוא למערכות לומדות

תרגיל בית 3

פרטי המגישים:

אילנית סמול	315820258	ilanitsmul@campus.technion.ac.il
שם	ת.ז	אימייל

אסף אנטר	301019733	asafanter@campus.technion.ac.il
שם	ת.ז	אימייל

מטלת החובה

הערות:

- פתרנו את התרגיל באמצעות חלוקתו ל-4 שלבים. מספור השלבים תואם את מספור השלבים בפונקציית main שב-modeling.py.
- התוצאות שקיבלנו יכולות להשתנות כתלות באופן החלוקה של ה-data לקבוצות train/validation/test וכן באופן פעולת המסווגים שבחרנו (שכן חלקם יכולים להשתמש בפרמטרים רנדומליים המשתנים מריצה לריצה). לכן, כדי שנוכל לנתח את התוצאות, נתייחס לריצה אחת של הקוד שלנו – אשר את הפלט שלה אפשר לראות בקובץ output.txt (לשם נוחות, הפלט מופרד באמצעות כותרות STEP ל-4 חלקים, כותרת אחת לכל שלב בתרגיל). לאורך התרגיל, אנו ננתח את התוצאות המפורטות בקובץ זה בלבד.
- בסוף דו"ח זה צירפנו הסבר על מבנה תיקיית ההגשה והקבצים שבה.

שלב 1: טעינת המידע מהקובץ ה-ElectionsData.csv, חלוקתו ל-3 קבוצות של train/validation/test, ועיבוד המידע בהתאם לתרגיל הקודם ולקבוצת ה-features המצומצמת שנתונה כעת.

- בדומה לתרגיל הקודם, בעזרת הפונקציה "read_csv" (המוגדרת ב-pandas) והקובץ ElectionsData.csv טענו את הנתונים לתוך אובייקט DataFrame, ולאחר מכן חילקנו אותו ל-3 קבוצות: 60% עבור training (6,000 קולות), 20% עבור validation (2,000 קולות) ו-20% עבור test (2,000 קולות).
  - בהתאם לחלוקה זו יצרנו את 3 הקבצים הבאים: "original_data_train.csv", "original_data_validation.csv" ו-"original_data_test.csv", כאשר כל קובץ מכיל את ה-DataFrame הרלוונטי לפי שמו.
- צמצמנו את ה-DataFrame שלנו כך שיכיל רק את 10 ה-features הנתונים בתרגיל, ולאחר מכן ביצענו עליו data preparation בדומה לתרגיל הקודם ואשר מורכב מהשלבים הבאים:
  1. תיקון outliers עבור תכונות נומריות: השלמת ערכים באמצעות החציון.
  2. השלמת ערכים חסרים: עבור תכונות נומריות – הצבנו את החציון, ועבור קטגוריאליות – את הקטגוריה השכיחה בעמודה.
  3. המרת ערכים – נרמול ושינויי טיפוסים: עבור תכונות נומריות – נרמול לפי שיטת ה-min-max, ועבור תכונות קטגוריאליות – המרה למספרים כתלות בסוג הקטגוריה (למשל, תכונות בינאריות הומרו לערכים של 0 ו-1, תכונות בעלות x ערכים הומרו לערכים 1 עד x).
  - בהתאם לחלוקה זו יצרנו את 3 הקבצים הבאים: "prepared_data_train.csv", "prepared_data_validation.csv" ו-"prepared_data_test.csv", כאשר כל קובץ מכיל את ה-DataFrame הרלוונטי לפי שמו.

שלב 2: אימון מגוון של models עם ה-training set בלבד ובעזרת שיטת cross-validation, כדי לקבוע את ה-hyper-parameters שמשיגים את התוצאות הטובות ביותר עבור כל model.

- סקרנו 4 מסווגים מוכרים מהספרייה "scikit-learn": RandomForestClassifier, DecisionTreeClassifier, KNeighborsClassifier ו-SVC.
  - לכל model/מסווג ביצענו מספר הרצות עם hyper-parameters שונים, ועבור כל ריצה כזו בדקנו מהו ה-score הממוצע המתקבל כאשר משתמשים בשיטת cross-validation (עם פרמטר cv=10) על ה-training set בלבד.
    חישוב ה-score הממוצע נעשה באמצעות שורת הקוד: "np.mean(cross_val_score(estimator=clf, X=x, y=y, cv=10))".
  - עבור כל model, בדקנו מהי הריצה שקיבלה את ה-score הממוצע הגבוה ביותר, ובחרנו את ה-hyper-parameters של model זה על פי אותה ריצה.
  - איך בחרנו אילו hyper-parameters לבדוק:
    1. ראשית, בחרנו לבדוק פרמטרים שאנו מבינים את התפקיד שלהם, ושסביר שישפיעו על ה-score באופן ניכר.
    2. לאחר מכן, עבור כל פרמטר p, בדקנו האם כדי להשתמש בערכו הדיפולטי או להשתמש בערך אחר. לצורך כך הגדרנו אובייקט model עם פרמטרים דיפולטיביים פרט לפרמטר p שעבורו ביצענו איטרציה על כל הערכים האפשריים (אם מדובר בערך קטגוריאלי – פשוט עברנו על כולם, ואחרת – עברנו על הערכים שסביר שימקסמו את ה-score). במידה וגילינו שכדאי שלא להשתמש בערך הדיפולטי – המשכנו עם ה-p הזה לשלב הבא.
    3. עבור כל הפרמטרים $p_1,\ldots,p_n$ שקיבלנו מהשלב הקודם – הגדרנו אובייקטי model שונים עם כל השילובים האפשריים של פרמטרים אלו, ובדקנו מהו ה-model שמקבל את ה-score המקסימלי.
    - הערה: בקוד בחרנו לפרט רק את שלב 3, משום שפירוט של שלב 2 היה מגדיל באופן ניכר את הקוד וזמן הרצתו.

- נדגים את שלב זה על המסווג DecisionTreeClassifier:
  - גילינו שכדאי לשנות את הפרמטרים הבאים:
    - criterion – הפונקציה שבאמצעותה מודדים את איכות הפיצול. הפרמטר מקבל את הערכים הבאים: "gini" עבור Gini impurity ו-"entropy" עבור information gain. הערך הדיפולטי הינו "gini". בחרנו לבדוק את שני הערכים האפשריים.
    - min_samples_split – המס' המינימלי של דגימות הדרושות בצומת על מנת שתחשב כעלה. הפרמטר מקבל כל ערך הגדול-ממש מ-1. הערך הדיפולטי הינו 2. בחרנו לבדוק את כל הערכים בין 2 ל-15 (לא כולל).
  - הגדרנו model של DecisionTreeClassifier עבור כל שילוב של הפרמטרים הנ"ל, כלומר 26 שילובים בסך הכל.
  - קיבלנו שה-score המקסימלי מתקבל עבור criterion=entropy ו-min_samples_split=4 וערכו 0.8459 (לשם השוואה, שימוש בערכים הדיפולטיביים criterion=gini ו-min_samples_split=2 הניב score של 0.8336).
- בסיום שלב זה קיבלנו רשימה של 4 אובייקטי model, כאשר לכל אחד מוגדרים ה-hyper-parameters שממקסמים את ה-score על פני ה-training set:

Model	Non-default parameters	Score
RandomForestClassifier	criterion='entropy', min_samples_split=5	0.8905443340004988
DecisionTreeClassifier	criterion='entropy', min_samples_split=4	0.8459971414164851
KNeighborsClassifier	n_neighbors=4, weights='distance'	0.8086625121622291
SVC	kernel='linear'	0.7974844556899039

שלב 3: אימון ה-models מהשלב הקודם (אחרי שקבענו ה-hyper-parameters עבור כל model) עם כל ה-training set (ללא cross-validation), בדיקת הביצועים על ה-validation set, ובחירת ה-model הטוב ביותר לכל משימה.

- את כל אחד מ-4 ה-models מהשלב הקודם אימנו בעזרת ה-training set וביצענו בדיקת ביצועים על ה-validation set.
- <u>חקרנו את המדדים הבאים:</u>
  - **accuracy (ו-error):**
    - נמדד באמצעות הפונקציה "accuracy_score(y_true, y_predict)".
  - **accuracy עם threshold=0.5:**
    - נמדד כמו ה-accuracy הרגיל אלא שאת y_predict לא השגנו ע"י קריאה רגילה ל-"clf.predict(validation_X)" אלא ע"י קריאה ל-"clf.predict_proba(validation_X)", כאשר לדגימות שקיבלו הסתברות הקטנה מ-0.5 להשתייך למפלגה כלשהי הצבנו סיווג מסוג "Unknown", ולשאר הצבנו את הסיווג שקיבל את ההסתברות המקסימלית (כמו מסווג רגיל).
  - **precision, recall, predict votes, true votes:** נמדדו ביחס לכל אחת המפלגות (ללא threshold).
    - precision נמדד בעזרת הפונקציה "precision_score(y_true, y_predict, average=None)".
      באופן דומה, recall נמדד בעזרת הפונקציה "recall_score(y_true, y_predict, average=None)".
    - predict votes מודד את מס' הקולות שקיבלה המפלגה והאחוז ההצבעה אליה ביחס לשאר המפלגות, וזאת בהתאם לתוצאות שנחזו ע"י המסווג שלנו.
      באופן דומה, true votes מודד את אותם המדדים, אבל בהתאם לתוצאות האמת.
  - **distribution of predict votes:** מס' הקולות שקיבלה כל מפלגה, מהגדולה לקטנה (ללא threshold).
  - **גרף ה-confusion matrix:** (ללא threshold)
    - נוצר באמצעות הפונקציה "confusion_matrix(y_true, y_predict, labels=labels)".
  - **גרף ה-histogram:**
    - מכיל 3 תתי-היסטוגרמות, אחת עבור תוצאות האמת, שנייה עבור תוצאות המסווג ללא threshold ושלישית עבור תוצאות המסווג עם threshold=0.5.
    - בנוסף לעמודה עבור כל אחת מהפלגות, הוספנו עמודה לסיווג מסוג "Unknown" הנמצאת בשימוש בהיסטוגרמה עם threshold=0.5. בעזרת עמודה זו נוכל להבין עד כמה המסווג "החלטי" בתהליך הסיווג שלו.
- כעת נציג את התוצאות עבור כל אחד מ-4 המסווגים שלנו, מהמסווג שקיבל את ה-accuracy הגבוה ביותר למסווג שקיבל את ה-accuracy הנמוך ביותר בשלב הקודם.
  - כדי להקל על נוחות הקריאה עיגלנו את הנתונים המספריים וסידרנו אותם בטבלאות וכדומה, אולם כאמור התוצאות המלאות נמצאות בקובץ output.txt.

- RandomForestClassifier: (תזכורת: הדיוק בשלב 2 היה 89.05%)

- accuracy (without threshold): 88.65% (error: 11.35%)
- accuracy (with threshold=0.5): 84.25% (error: 15.75%)
- metrics on each label (without threshold):

	Blues	Browns	Greens	Greys	Khakis	Oranges	Pinks	Purples	Reds	Turquoises	Violets	Whites	Yellows
precision	0.9604	0.8783	0.9434	0.9082	0.8535	0.8966	0.8966	0.955	0.9434	0.7922	0.7451	0.7619	0.6957
recall	0.8017	0.8691	1.0	0.9175	0.9734	0.8966	0.8525	0.9721	0.9615	0.5545	0.5278	0.7619	0.7273
predict votes	101 (5.05%)	189 (9.45%)	106 (5.3%)	98 (4.9%)	471 (23.55%)	58 (2.9%)	58 (2.9%)	511 (25.55%)	106 (5.3%)	77 (3.85%)	51 (2.55%)	105 (5.25%)	69 (3.45%)
true votes	121 (6.05%)	191 (9.55%)	100 (5.0%)	97 (4.85%)	413 (20.65%)	58 (2.9%)	61 (3.05%)	502 (25.1%)	104 (5.2%)	110 (5.5%)	72 (3.6%)	105 (5.25%)	66 (3.3%)

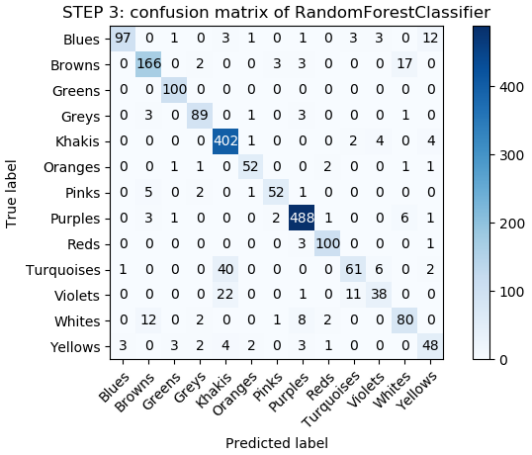
- distribution of predict votes:

1	Purples	511
2	Khakis	471
3	Browns	189
4	Reds	106
5	Greens	106
6	Whites	105
7	Blues	101
8	Greys	98
9	Turquoises	77
10	Yellows	69
11	Pinks	58
12	Oranges	58
13	Violets	51

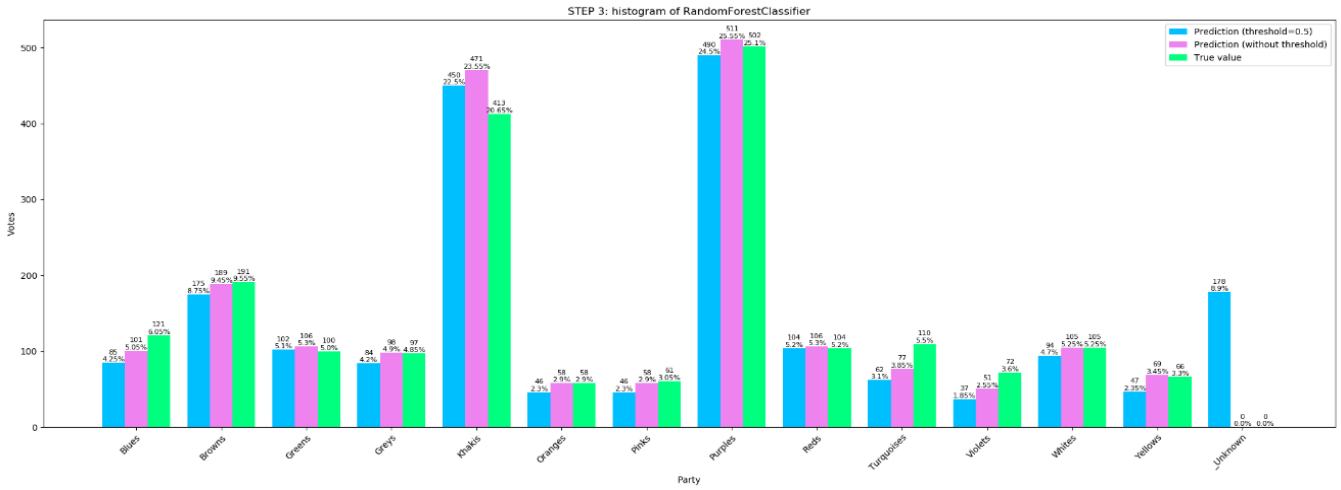
- distribution of true votes:

1	Purples	502
2	Khakis	413
3	Browns	191
4	Blues	121
5	Turquoises	110
6	Whites	105
7	Reds	104
8	Greens	100
9	Greys	97
10	Violets	72
11	Yellows	66
12	Pinks	61
13	Oranges	58

- confusion matrix:



- histogram:



- **מדד ה-accuracy:** 88.65% – הערך הגבוה ביותר ביחס לשאר המסווגים.
- **תוצאות הבחירות:** קיימת התאמה לתוצאות האמת ב-3 המקומות הראשונים: Purples << Khakis << Browns.
- **גרף ה-confusion matrix:** הערכים שמחוץ לאלכסון מקבלים לרוב ערכים נמוכים, פרט לחריגה מרכזית בעמודה Khakis אשר מעלה את אחוז ההצבעות שלה ל-23.55% (בהשוואה ל-20.65% בתוצאות האמת), וגורמת להידרדרות של Turquoises מהמקום ה-5 ל-9 ושל Violets מהמקום ה-10 למקום ה-13. קיימות חריגות נוספות אך משמעותיות פחות בעמודות Browns, Turquoises, Whites ו-Yellows.
- **גרף ה-histogram:** ניתן לראות שלרוב יש התאמה עם תוצאות האמת מבחינת מס' הקולות (גובה העמודות), פרט לחריגה מרכזית ב-Khakis וחריגות נוספות אך קטנות יותר ב-Blues, Turquoises ו-Violets.
- **קביעת ה-threshold=0.5:** מדד ה-accuracy יורד לערך של 84.25% (ירידה של 4.4% – הירידה החדה ביותר מבין המסווגים), בשל 6.9% מההצבעות (178 קולות) שקיבלו סיווג "Unknown".
- **סיכום:** RandomForest בעל דיוק גבוה מאוד (ערכי ה-precision גדולים מ-0.69), וזאת למרות שה-threshold שלו נמוך (כלומר, אף על פי שהוא מחליט את הסיווג בהסתמך על הסתברות קטנה יחסית – הסיווג הזה בכל זאת נכון).

- DecisionTreeClassifier: (תזכורת: הדיוק בשלב 2 היה 84.59%)

- accuracy (without threshold): 85.75% (error: 14.25%)
- accuracy (with threshold=0.5): 85.70% (error: 14.30%)
- metrics on each label (without threshold):

	Blues	Browns	Greens	Greys	Khakis	Oranges	Pinks	Purples	Reds	Turquoises	Violets	Whites	Yellows
precision	0.8349	0.8432	0.9897	0.9059	0.8918	0.7656	0.8095	0.9506	0.9167	0.7282	0.6528	0.7091	0.5366
recall	0.7521	0.8168	0.96	0.7938	0.8983	0.8448	0.8361	0.9582	0.9519	0.6818	0.6528	0.7429	0.6667
predict votes	109 (5.45%)	185 (9.25%)	97 (4.85%)	85 (4.25%)	416 (20.8%)	64 (3.2%)	63 (3.15%)	506 (25.3%)	108 (5.4%)	103 (5.15%)	72 (3.6%)	110 (5.5%)	82 (4.1%)
true votes	121 (6.05%)	191 (9.55%)	100 (5.0%)	97 (4.85%)	413 (20.65%)	58 (2.9%)	61 (3.05%)	502 (25.1%)	104 (5.2%)	110 (5.5%)	72 (3.6%)	105 (5.25%)	66 (3.3%)

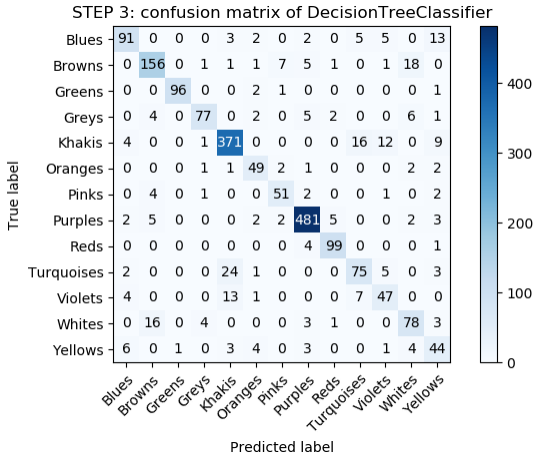
- distribution of predict votes:

1	Purples	506
2	Khakis	416
3	Browns	185
4	Whites	110
5	Blues	109
6	Reds	108
7	Turquoises	103
8	Greens	97
9	Greys	85
10	Yellows	82
11	Violets	72
12	Oranges	64
13	Pinks	63

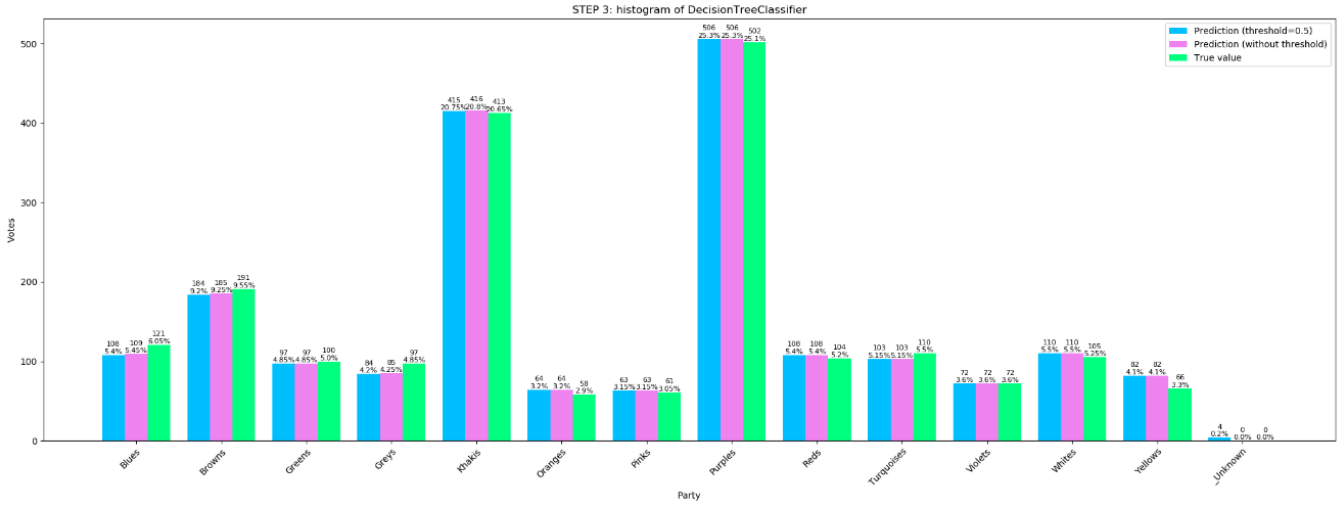
- distribution of true votes:

1	Purples	502
2	Khakis	413
3	Browns	191
4	Blues	121
5	Turquoises	110
6	Whites	105
7	Reds	104
8	Greens	100
9	Greys	97
10	Violets	72
11	Yellows	66
12	Pinks	61
13	Oranges	58

- confusion matrix:



- histogram:



- **מדד ה-accuracy:** 85.75% – הערך השני הגבוה ביותר ביחס לשאר המסווגים (קטן ב-2.9% מ-RandomForest וגדול בכ-5.75% משני המסווגים האחרים).
- **תוצאות הבחירות:** קיימת התאמה לתוצאות האמת ב-3 המקומות הראשונים: Purples >> Khakis >> Browns.
- **גרף ה-confusion matrix:** הערכים שמחוץ לאלכסון מקבלים לרוב ערכים נמוכים, פרט לחריגה מרכזית בעמודה Khakis, ולחריגות נוספות אך משמעותיות פחות בעמודות Browns, Turquoises, Violets, Whites ו-Yellows. בהשוואה למסווג הקודם RandomForest, כאן החריגות מתונות יחסית (כך למשל ה-precision של Khakis הינו 0.8918 כאן לעומת 0.8535 במסווג הקודם).
- **גרף ה-histogram:** ניתן לראות שלרוב יש התאמה עם תוצאות האמת מבחינת מס' הקולות (גובה העמודות). בהשוואה למסווג הקודם RandomForest, כאן התפלגות הקולות בין המפלגות דומה יותר לתוצאות האמת, וזאת למרות שה-accuracy נמוך יותר (כלומר, DecisionTree טועה יותר בסיווג כל הצבעה, אולם בחישוב הכללי התוצאות שלו מתקרבות יותר לתוצאות האמת).
- **קביעת ה-threshold=0.5:** מדד ה-accuracy יורד לערך של 85.70% (ירידה של 0.05% – הירידה המזערית ביותר מבין המסווגים), בשל 0.2% מההצבעות (4 קולות) שקיבלו סיווג "Unknown".
- **סיכום:** הדיוק של DecisionTree נמוך יחסית למסווג הקודם RandomForest, אולם עדיין גבוה יחסית לשני המסווגים הבאים (וכן ערכי ה-precision שלו גדולים מ-0.53). בהקשר ל-threshold, ניתן לראות ש-DecisionTree החלטי הרבה יותר מהמסווגים האחרים.

- KNeighborsClassifier: (תזכורת: הדיוק בשלב 2 היה 80.86%)

- accuracy (without threshold): 80.00% (error: 20.00%)
- accuracy (with threshold=0.5): 78.10% (error: 21.90%)
- metrics on each label (without threshold):

	Blues	Browns	Greens	Greys	Khakis	Oranges	Pinks	Purples	Reds	Turquoises	Violets	Whites	Yellows
precision	0.9077	0.7523	0.9029	0.8763	0.7416	0.9592	0.9545	0.9387	0.8636	0.5217	0.3696	0.6786	0.4535
recall	0.4876	0.8586	0.93	0.8763	0.9104	0.8103	0.6885	0.9761	0.9135	0.3273	0.2361	0.5429	0.5909
predict votes	65 (3.25%)	218 (10.9%)	103 (5.15%)	97 (4.85%)	507 (25.35%)	49 (2.45%)	44 (2.2%)	522 (26.1%)	110 (5.5%)	69 (3.45%)	46 (2.3%)	84 (4.2%)	86 (4.3%)
true votes	121 (6.05%)	191 (9.55%)	100 (5.0%)	97 (4.85%)	413 (20.65%)	58 (2.9%)	61 (3.05%)	502 (25.1%)	104 (5.2%)	110 (5.5%)	72 (3.6%)	105 (5.25%)	66 (3.3%)

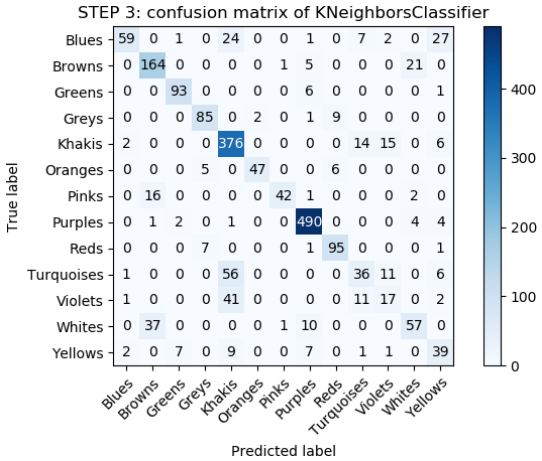
- distribution of predict votes:

1	Purples	522
2	Khakis	507
3	Browns	218
4	Reds	110
5	Greens	103
6	Greys	97
7	Yellows	86
8	Whites	84
9	Turquoises	69
10	Blues	65
11	Oranges	49
12	Violets	46
13	Pinks	44

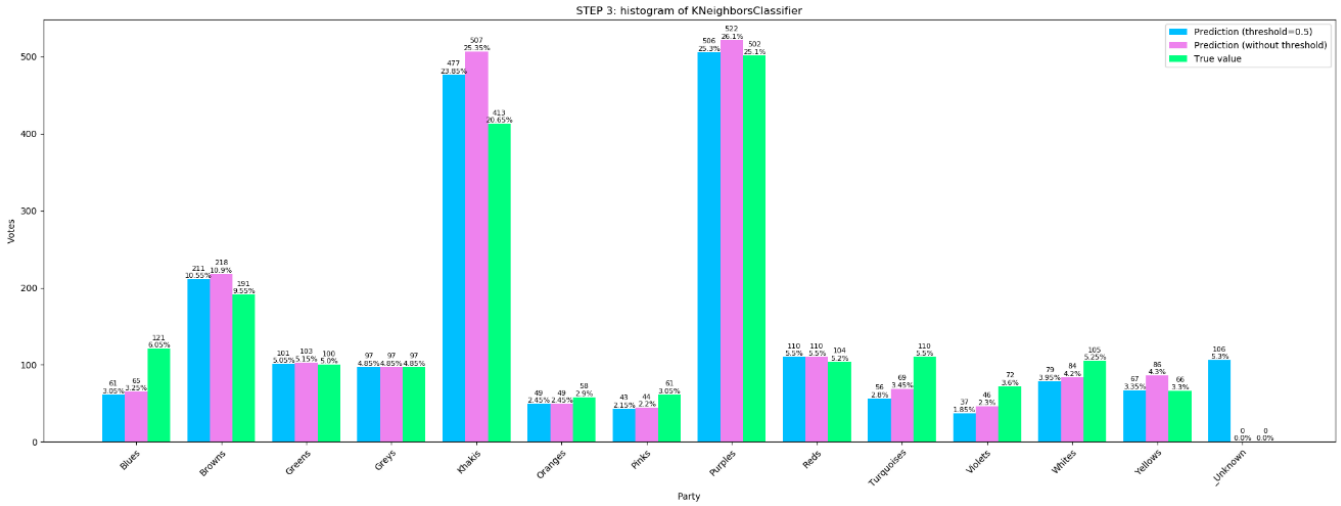
- distribution of true votes:

1	Purples	502
2	Khakis	413
3	Browns	191
4	Blues	121
5	Turquoises	110
6	Whites	105
7	Reds	104
8	Greens	100
9	Greys	97
10	Violets	72
11	Yellows	66
12	Pinks	61
13	Oranges	58

- confusion matrix:



- histogram:



- **מדד ה-accuracy:** 80.00% – הערך הרביעי הגבוה ביותר ביחס לשאר המסווגים (קטן ב-5.75% מ-DecisionTree).
- **תוצאות הבחירות:** קיימת התאמה לתוצאות האמת ב-3 המקומות הראשונים: Purples >> Khakis >> Browns.
- **גרף ה-confusion matrix:** הערכים שמחוץ לאלכסון בעלי חריגות גדולות יותר באופן משמעותי משני המסווגים הראשונים, בעיקר בעמודות Browns, Khakis, Turquoises, Violets, Whites ו-Yellows. חיזוק לכך ניתן לראות במדדי ה-precision של המפלגות שלראשונה מקבלים ערכים מתחת ל-0.5 (בניגוד לשני המסווגים הקודמים שקיבלו ערכים גדולים יותר).
- **גרף ה-histogram:** ניתן לראות שההפרשים בין תוצאות האמת לבין תחזיות המסווגים נעשים גדולים יותר ביחס לשני המסווגים הראשונים, במיוחד בעמודת ה-Khakis (מקבלים 25.35% מהקולות בתחזית לעומת 20.65% בתוצאות האמת).
- **קביעת ה-threshold=0.5:** מדד ה-accuracy יורד לערך של 78.10% (ירידה של 1.9% – ירידה ממוצעת בין המסווגים), בשל 5.3% מההצבעות (106 קולות) שקיבלו סיווג "Unknown".
- **סיכום:** הדיוק של KNeighbors נמוך בהשוואה לשני המסווגים הראשונים, ובעל הבדלים משמעותיים יותר ביחס לסדר היחסי בין המפלגות בתוצאות האמת.

- SVC: (תזכורת: הדיוק בשלב 2 היה 79.74%)

- accuracy (without threshold): 80.05% (error: 19.95%)
- accuracy (with threshold=0.5): 78.15% (error: 21.85%)
- metrics on each label (without threshold):

	Blues	Browns	Greens	Greys	Khakis	Oranges	Pinks	Purples	Reds	Turquoises	Violets	Whites	Yellows
precision	0.9	0.7415	0.9327	0.8725	0.6894	0.875	1.0	0.8835	0.8932	0.0	0.0	0.7317	0.6222
recall	0.8182	0.9162	0.97	0.9175	0.9782	0.8448	0.6557	0.9363	0.8846	0.0	0.0	0.2857	0.8485
predict votes	110 (5.5%)	236 (11.8%)	104 (5.2%)	102 (5.1%)	586 (29.3%)	56 (2.8%)	40 (2.0%)	532 (26.6%)	103 (5.15%)	0 (0.0%)	0 (0.0%)	41 (2.05%)	90 (4.5%)
true votes	121 (6.05%)	191 (9.55%)	100 (5.0%)	97 (4.85%)	413 (20.65%)	58 (2.9%)	61 (3.05%)	502 (25.1%)	104 (5.2%)	110 (5.5%)	72 (3.6%)	105 (5.25%)	66 (3.3%)

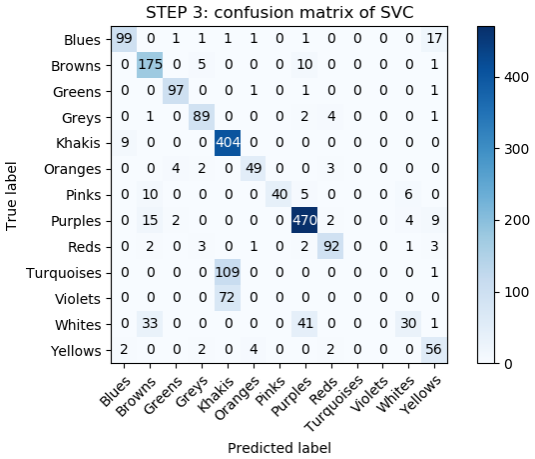
- distribution of predict votes:

1	Khakis	586
2	Purples	532
3	Browns	236
4	Blues	110
5	Greens	104
6	Reds	103
7	Greys	102
8	Yellows	90
9	Oranges	56
10	Whites	41
11	Pinks	40
12	Violets	0
13	Turquoises	0

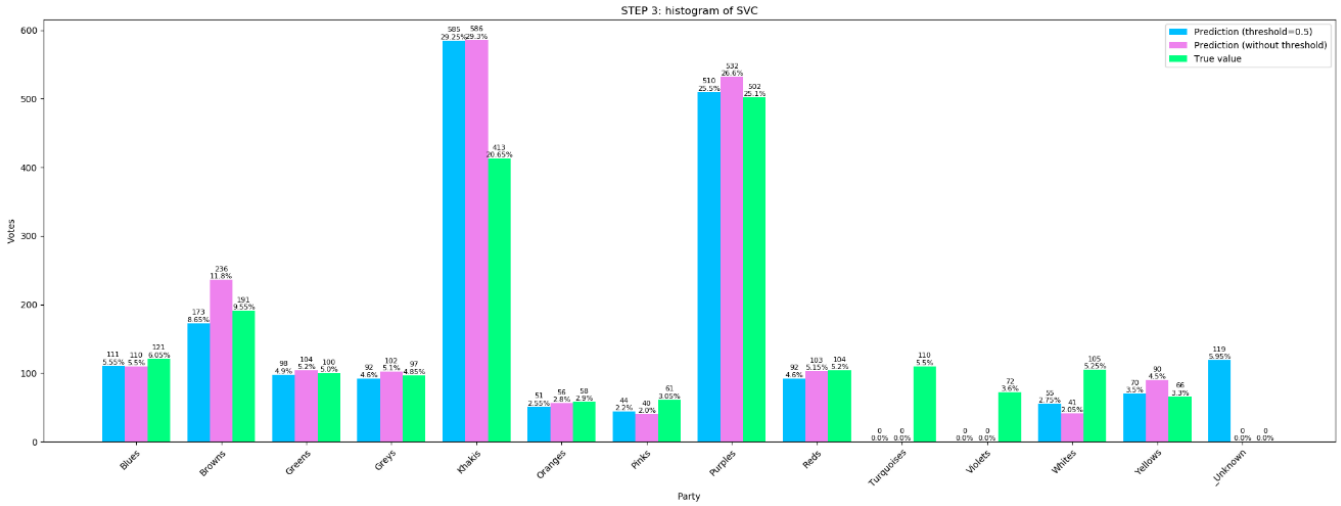
- distribution of true votes:

1	Purples	502
2	Khakis	413
3	Browns	191
4	Blues	121
5	Turquoises	110
6	Whites	105
7	Reds	104
8	Greens	100
9	Greys	97
10	Violets	72
11	Yellows	66
12	Pinks	61
13	Oranges	58

- confusion matrix:



- histogram:



- **מדד ה-accuracy:** 80.05% – הערך השלישי הגבוה ביותר ביחס לשאר המסווגים (קטן ב-5.7% מ-DecisionTree).
- **תוצאות הבחירות:** לא קיימת התאמה לתוצאות האמת ב-3 המקומות הראשונים: Khakis << Purples << Browns (המקומות הראשון והשני החליפו את הסדר ביניהם).
- **גרף ה-confusion matrix:** הערכים שמחוץ לאלכסון בעלי חריגות גדולות יותר באופן משמעותי משלושת המסווגים הראשונים, בעיקר בעמודות Browns, Khakis, Purples, ו-Yellows. החריגה המרכזית אשר נמצאת בעמודה Khakis גרמה להעלאת אחוז ההצבעות שלה ל-29.3% (בהשוואה ל-20.65% בתוצאות האמת), וגורמת להידרדרות של Turquoises ו-Violets מהמקומות ה-5 וה-10 בהתאמה אל המקומות האחרונים, עד כדי קבלת 0 קולות בלבד.
- **גרף ה-histogram:** כמו במסווג הקודם KNeighbors, גם כאן ניתן לראות שההפרשים בין תוצאות האמת לבין תחזיות המסווגים נעשים גדולים יותר ביחס לשני המסווגים הראשונים, במיוחד בעמודת ה-Khakis.
- **קביעת ה-threshold=0.5:** מדד ה-accuracy יורד לערך של 78.15% (ירידה של 1.9% – ירידה ממוצעת בין המסווגים), בשל 5.95% מההצבעות (109 קולות) שקיבלו סיווג "Unknown".
- **סיכום:** הדיוק של SVC נמוך בהשוואה לשני המסווגים הקודמים, ובעל הבדלים משמעותיים יותר ביחס לסדר היחסי בין המפלגות בתוצאות האמת, ואף ב-3 המקומות הראשונים (בניגוד לשלושת המסווגים הקודמים).

- <u>סיכום התוצאות ובחירת model:</u>
  - בחירת ה-model: לאחר סקירת המסווגים הנ"ל הגענו למסקנה שהמדד החשוב ביותר הינו ה-accuracy, ולכן בחרנו להמשיך לשלב הבא עם **RandomForestClassifier** (חסרונו היחיד הינו ה-threshold הנמוך שלו, אך למרות זאת תוצאותיו מדויקות).
  - תוצאות הבחירות: למרות השוני במדד ה-accuracy של כל אחד מהמסווגים, ראינו שכולם נתקלו בקשיים דומים. דוגמאות: כל המסווגים נתנו ל-Khakis עודף קולות על חשבון הקולות של Turquoises ו-Violets; ה-precision של Yellows תמיד קיבל את הערך הנמוך ביותר (או השני-הנמוך ביותר); ועוד.

שלב 4: אימון ה-model הנבחר על ה-training set וה-validation set (ביחד), בדיקת הביצועים על ה-test set, ומענה על המשימות.

- אימנו את RandomForestClassifier בעזרת ה-training set + validation set וביצעו בדיקת ביצועים על ה-test set.
- <u>RandomForestClassifier:</u> (תזכורת: הדיוק בשלב 2 היה 89.05% ובשלב 3 88.65%)

– accuracy (without threshold): 90.30% (error: 9.70%)

– metrics on each label (without threshold):

	Blues	Browns	Greens	Greys	Khakis	Oranges	Pinks	Purples	Reds	Turquoises	Violets	Whites	Yellows
precision	0.9222	0.9022	0.9912	0.9659	0.8517	0.9059	0.8269	0.9607	0.9348	0.8416	0.7925	0.8293	0.8228
recall	0.8469	0.9222	0.9912	0.9659	0.9693	0.9167	0.7963	0.9772	0.9451	0.6855	0.5	0.7727	0.8333
predict votes	90 (4.5%)	184 (9.2%)	114 (5.7%)	88 (4.4%)	445 (22.25%)	85 (4.25%)	52 (2.6%)	535 (26.75%)	92 (4.6%)	101 (5.05%)	53 (2.65%)	82 (4.1%)	79 (3.95%)
true votes	98 (4.9%)	180 (9.0%)	114 (5.7%)	88 (4.4%)	391 (19.55%)	84 (4.2%)	54 (2.7%)	526 (26.3%)	91 (4.55%)	124 (6.2%)	84 (4.2%)	88 (4.4%)	78 (3.9%)

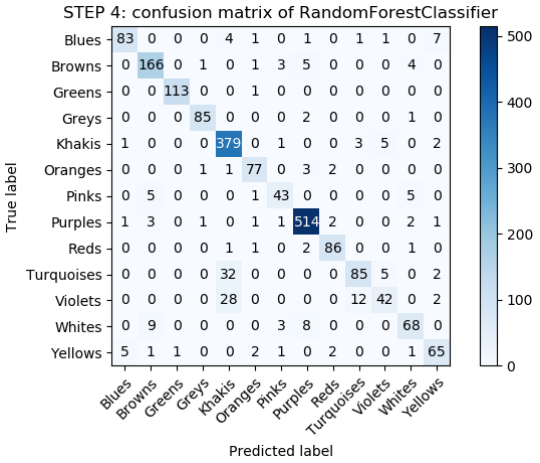
– distribution of predict votes:

1	Purples	535
2	Khakis	445
3	Browns	184
4	Greens	114
5	Turquoises	101
6	Reds	92
7	Blues	90
8	Greys	88
9	Oranges	85
10	Whites	82
11	Yellows	79
12	Violets	53
13	Pinks	52

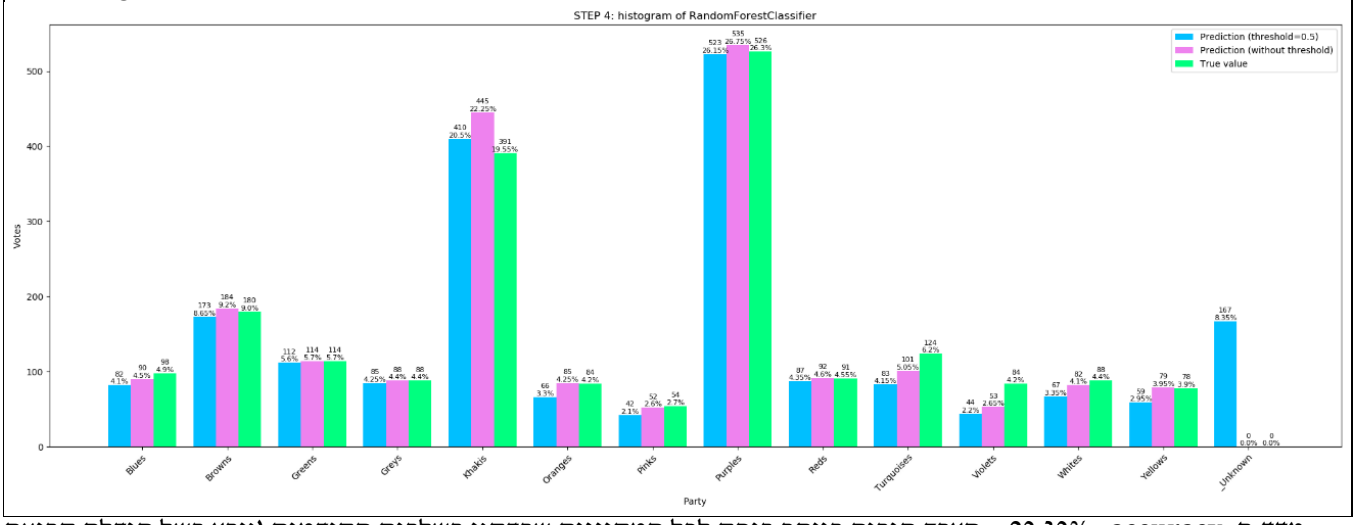
– distribution of true votes:

1	Purples	526
2	Khakis	391
3	Browns	180
4	Turquoises	124
5	Greens	114
6	Blues	98
7	Reds	91
8	Whites	88
9	Greys	88
10	Violets	84
11	Oranges	84
12	Yellows	78
13	Pinks	54

– confusion matrix:



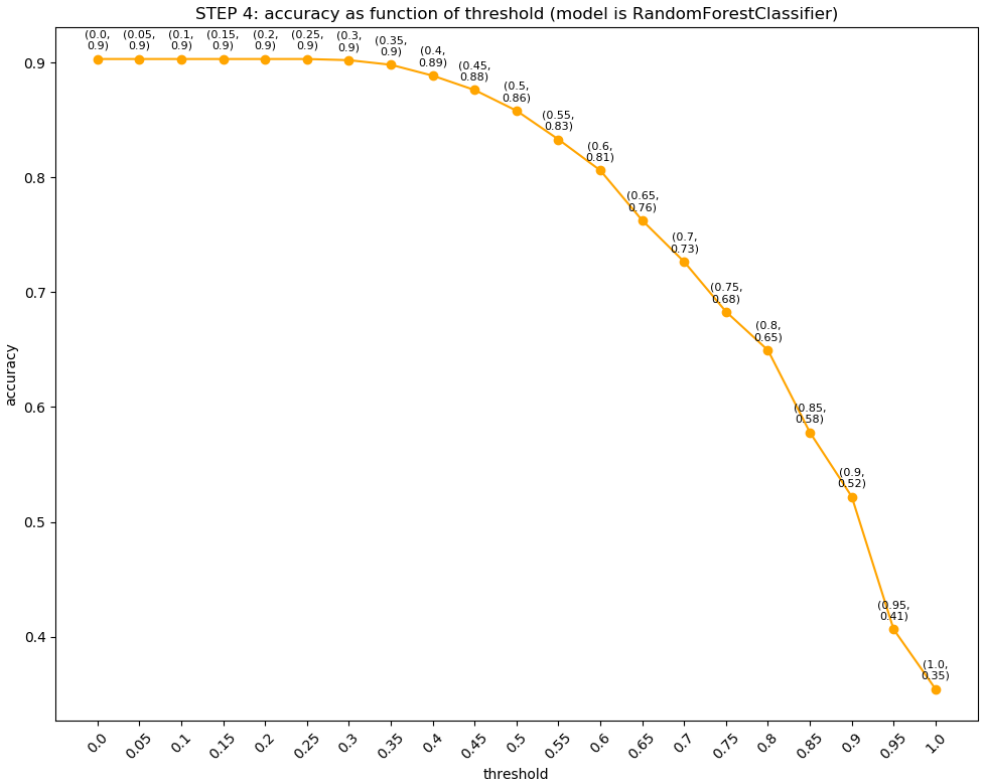
– histogram:



- **מדד ה-accuracy:** 90.30% – הערך הגבוה ביותר ביחס לכל המסווגים שבדקנו בשלבים הקודמים (נובע בשל הגדלת קבוצת האימון הודות לשילוב עם ה-validation set).
- **תוצאות הבחירות:** קיימת התאמה לתוצאות האמת ב-3 המקומות הראשונים: Purples << Khakis << Browns. מקומות 4 (Greens) ו-5 (Turquoises) התחלפו, וכך גם מקומות 6 (Reds) ו-7 (Blues), מקומות 10 (Whites) ו-12 (Violets) מדרדרים בכמה מקומות, אולם שאר המפלגות במקומות 8, 9, 11 ו-13 (Greys, Oranges, Yellows, ו-Pinks) שומרים על המיקום היחסי ביניהם ביחס לתוצאות האמת.

- **גרף ה-confusion matrix:** הערכים שמחוץ לאלכסון מקבלים לרוב ערכים נמוכים, פרט לחריגה מרכזית בעמודה Khakis אשר מעלה את אחוז ההצבעות שלה ל-22.25% (בהשוואה ל-19.55% בתוצאות האמת), וגורמת להידרדרות של Turquoises מהמקום ה-4 למקום ה-5 ושל Violets מהמקום ה-10 למקום ה-12. בנוסף, ניתן לראות שמדדי ה-precision מקבלים את הערכים הגבוהים ביותר ביחס לשלבים הקודמים (מעל 0.79).
- **גרף ה-histogram:** ניתן לראות שלרוב יש התאמה עם תוצאות האמת מבחינת מס' הקולות (גובה העמודות), פרט לחריגה מרכזית ב-Khakis וחריגות נוספות אך קטנות יותר ב-Turquoises ו-Violets.

- <u>בדיקת ה-threshold:</u> בגרף למטה ניתן לראות כיצד ה-threshold משפיע על מדד ה-accuracy.



- כמו שהסקנו בשלב הקודם, המסווג RandomForest משתמש ב-threshold נמוך יחסית, אולם למרות זאת אחוז הדיוק שלו גבוה מאוד. כלומר, למרות שהוא מספק סיווגים עם הסתברויות נמוכות – עדיין הסיווגים האלו נכונים.
- לכן, במענה על המשימות שבתרגיל (מפורט בהמשך) בחרנו להשתמש בתחזיות של RandomForest כאשר לא מוגדר עבורו threshold, וזאת כדי למקסם את הדיוק (כפי שהגרף ממחיש).

- <u>כעת נענה על 3 המשימות שבתרגיל:</u>
- **משימה 1 – המפלגה שתזכה ברוב הקולות:** מפלגת ה-Purples עם 535 קולות שהם 26.75%.
- **משימה 2 – חלוקת הקולות בין המפלגות:**

	party	votes	percentage
1	Purples	535	26.75%
2	Khakis	445	22.25%
3	Browns	184	9.2%
4	Greens	114	5.7%
5	Turquoises	101	5.05%
6	Reds	92	4.6%
7	Blues	90	4.5%
8	Greys	88	4.4%
9	Oranges	85	4.25%
10	Whites	82	4.1%
11	Yellows	79	3.95%
12	Violets	53	2.65%
13	Pinks	52	2.6%

- **משימה 3 – רשימת ה-most probable voters עבור כל מפלגה:**

- כאמור, בחרנו להתבסס על תחזית המסווג RandomForest כאשר לא מוגדר עבורו threshold, משום שראינו שהסיווגים שלו מדויקים למרות שההסתברות שהוא מעניק לסיווגיו נמוכה יחסית: לפי הגרף מהעמוד הקודם, הדיוק המקסימלי מתקבל עבור threshold=0.35.

נסביר זאת באמצעות דוגמא: נניח שדגימה מסויימת ב-test set קיבלה את ההסתברויות הבאות ע"י המסווג: 35% למפלגה A, 34% למפלגה B, 30% למפלגה C, וה-1% הנותרים מתחלקים בין יתר המפלגות. אזי, למרות שההסתברויות מאוד קרובות אחת לשנייה, וכל אחת אינה בעלת רוב משמעותי (שהרי אף 35% זה ערך נמוך יחסית), המסווג שלנו ייבחר במפלגה A. אבל, משום שהראינו שדיוק המסווג גבוה ומשיג תוצאות טובות (accuracy=90.30%), אז ככל הנראה שהסיווג למפלגה A היה נכון, ולכן נמליץ למפלגה A להשקיע בהסעה לבוחר שמייצג דגימה זו, למרות שהסתברות של 35% לא נשמעת בהתחלה כהסתברות מבטיחה.

- בטבלה הבאה פירטנו עבור כל מפלגה את המצביעים שיבחרו בה (ע"פ תחזית המסווג שלנו), כאשר האינדקסים תואמים את האינדקסים שבקובץ processed_data_test.csv (השורה הראשונה בקובץ זה מציינת את שם העמודה, "Vote", ולכן האינדקסים שמפורטים כאן מקבלים את הערכים שבין 2 לבין 2001).

1	Purples 535 votes	[7, 13, 15, 33, 35, 37, 42, 48, 52, 53, 55, 62, 63, 64, 70, 76, 80, 84, 85, 86, 90, 95, 96, 98, 106, 107, 108, 114, 117, 134, 140, 146, 150, 151, 163, 165, 170, 174, 180, 183, 187, 200, 201, 209, 211, 217, 220, 221, 226, 229, 237, 245, 247, 248, 250, 253, 254, 255, 256, 265, 269, 274, 276, 277, 280, 282, 283, 284, 288, 292, 305, 306, 310, 328, 330, 337, 338, 351, 353, 360, 366, 369, 373, 376, 378, 380, 383, 385, 387, 389, 391, 397, 398, 401, 403, 404, 406, 409, 410, 414, 416, 417, 421, 425, 433, 446, 447, 450, 458, 461, 462, 463, 466, 471, 476, 477, 485, 490, 493, 494, 496, 500, 501, 504, 505, 507, 513, 515, 517, 518, 526, 535, 538, 539, 543, 544, 546, 547, 548, 551, 556, 561, 564, 565, 568, 569, 572, 573, 577, 588, 589, 595, 597, 601, 603, 610, 611, 614, 616, 619, 620, 625, 648, 651, 652, 654, 658, 660, 663, 665, 667, 669, 673, 675, 679, 680, 681, 682, 683, 687, 693, 695, 696, 698, 699, 705, 713, 717, 724, 726, 727, 729, 732, 733, 737, 738, 740, 741, 744, 751, 753, 760, 761, 764, 774, 776, 780, 781, 789, 793, 795, 798, 799, 803, 805, 806, 815, 820, 825, 826, 833, 836, 837, 846, 848, 851, 852, 853, 857, 861, 865, 867, 868, 871, 876, 879, 881, 885, 887, 889, 900, 901, 905, 909, 916, 919, 924, 925, 927, 930, 931, 932, 937, 945, 947, 953, 954, 963, 967, 983, 988, 990, 998, 1002, 1004, 1009, 1015, 1024, 1029, 1032, 1034, 1035, 1037, 1038, 1041, 1044, 1045, 1048, 1049, 1054, 1055, 1056, 1057, 1063, 1067, 1070, 1079, 1081, 1091, 1093, 1097, 1098, 1102, 1105, 1111, 1117, 1118, 1123, 1124, 1125, 1129, 1130, 1133, 1136, 1147, 1148, 1149, 1151, 1152, 1154, 1156, 1160, 1167, 1175, 1188, 1194, 1196, 1203, 1205, 1207, 1209, 1214, 1221, 1226, 1230, 1236, 1241, 1242, 1243, 1244, 1249, 1256, 1259, 1261, 1263, 1267, 1269, 1271, 1274, 1275, 1285, 1288, 1290, 1294, 1306, 1310, 1313, 1316, 1321, 1328, 1329, 1333, 1340, 1344, 1348, 1350, 1354, 1357, 1359, 1360, 1364, 1366, 1367, 1372, 1376, 1379, 1381, 1384, 1387, 1388, 1413, 1414, 1420, 1424, 1427, 1428, 1435, 1436, 1437, 1438, 1439, 1444, 1451, 1452, 1454, 1458, 1459, 1462, 1465, 1466, 1467, 1468, 1469, 1470, 1475, 1476, 1479, 1482, 1491, 1496, 1498, 1504, 1505, 1508, 1512, 1515, 1517, 1522, 1530, 1531, 1535, 1538, 1539, 1540, 1542, 1547, 1548, 1551, 1563, 1564, 1570, 1571, 1575, 1578, 1581, 1585, 1586, 1591, 1594, 1595, 1596, 1611, 1612, 1621, 1622, 1626, 1629, 1639, 1640, 1646, 1659, 1661, 1666, 1669, 1672, 1675, 1680, 1681, 1688, 1690, 1694, 1695, 1703, 1706, 1708, 1716, 1721, 1723, 1724, 1725, 1729, 1734, 1738, 1752, 1753, 1757, 1758, 1760, 1762, 1763, 1770, 1772, 1774, 1785, 1786, 1793, 1796, 1800, 1803, 1804, 1806, 1808, 1810, 1812, 1813, 1817, 1829, 1841, 1842, 1846, 1848, 1849, 1852, 1853, 1854, 1859, 1860, 1861, 1865, 1866, 1877, 1888, 1892, 1895, 1900, 1901, 1902, 1906, 1907, 1909, 1917, 1922, 1925, 1930, 1931, 1934, 1939, 1944, 1945, 1947, 1948, 1954, 1957, 1961, 1962, 1963, 1972, 1979, 1983, 1986, 1987, 1990, 1995, 1997, 2001]
2	Khakis 445 votes	[3, 5, 11, 19, 26, 31, 43, 46, 49, 56, 60, 61, 65, 66, 78, 79, 87, 92, 94, 99, 101, 102, 104, 115, 116, 119, 123, 127, 128, 130, 145, 147, 148, 149, 152, 153, 156, 158, 160, 164, 177, 186, 190, 191, 195, 202, 204, 210, 214, 215, 218, 223, 224, 232, 233, 240, 241, 257, 259, 260, 263, 267, 281, 286, 289, 293, 297, 301, 302, 307, 308, 311, 313, 314, 317, 318, 319, 332, 336, 340, 348, 349, 350, 355, 357, 361, 363, 364, 368, 370, 382, 384, 390, 396, 402, 407, 420, 426, 428, 429, 432, 434, 437, 439, 442, 445, 448, 449, 455, 457, 464, 472, 475, 480, 481, 482, 488, 491, 495, 497, 498, 499, 502, 509, 514, 524, 525, 536, 545, 552, 553, 557, 558, 559, 566, 570, 571, 574, 576, 579, 580, 585, 590, 591, 593, 594, 599, 604, 605, 606, 618, 624, 628, 631, 632, 633, 634, 639, 644, 645, 659, 664, 668, 671, 678, 686, 688, 690, 691, 707, 710, 711, 716, 725, 731, 736, 742, 745, 746, 748, 756, 763, 766, 767, 771, 775, 777, 779, 787, 790, 796, 802, 807, 810, 816, 818, 828, 842, 849, 850, 855, 860, 873, 877, 878, 880, 884, 886, 890, 893, 894, 896, 897, 898, 899, 902, 913, 914, 915, 917, 918, 921, 928, 933, 934, 935, 936, 939, 952, 957, 958, 965, 968, 973, 977, 978, 979, 982, 991, 994, 995, 1005, 1010, 1017, 1022, 1023, 1025, 1046, 1058, 1059, 1073, 1074, 1078, 1080, 1090, 1096, 1100, 1107, 1113, 1114, 1121, 1134, 1135, 1138, 1139, 1141, 1145, 1146, 1155, 1161, 1162, 1170, 1176, 1182, 1183, 1187, 1190, 1192, 1193, 1198, 1199, 1200, 1202, 1210, 1213, 1216, 1223, 1224, 1225, 1227, 1229, 1233, 1235, 1250, 1253, 1254, 1255, 1260, 1273, 1278, 1279, 1280, 1282, 1286, 1289, 1291, 1293, 1295, 1299, 1301, 1302, 1303, 1304, 1325, 1330, 1331, 1336, 1345, 1347, 1351, 1358, 1362, 1369, 1390, 1394, 1396, 1398, 1399, 1400, 1401, 1402, 1415, 1416, 1422, 1432, 1440, 1443, 1447, 1449, 1460, 1463, 1492, 1497, 1500, 1507, 1510, 1514, 1519, 1521, 1529, 1533, 1536, 1568, 1579, 1580, 1583, 1588, 1589, 1592, 1599, 1602, 1605, 1608, 1610, 1616, 1617, 1623, 1633, 1634, 1638, 1641, 1642, 1653, 1656, 1657, 1662, 1664, 1667, 1676, 1678, 1679, 1682, 1691, 1692, 1693, 1698, 1705, 1707, 1715, 1718, 1728, 1735, 1736, 1740, 1742, 1743, 1749, 1771, 1779, 1781, 1790, 1792, 1794, 1797, 1801, 1814, 1819, 1827, 1831, 1834, 1837, 1838, 1840, 1844, 1855, 1857, 1864, 1870, 1874, 1875, 1876, 1881, 1885, 1890, 1893, 1899, 1905, 1908, 1910, 1913, 1914, 1926, 1932, 1937, 1940, 1942, 1952, 1965, 1968, 1970, 1988, 1989, 1996, 1999, 2000]
3	Browns 184 votes	[6, 8, 9, 18, 22, 24, 25, 28, 34, 39, 47, 67, 72, 81, 82, 88, 100, 103, 109, 110, 122, 136, 168, 169, 176, 181, 188, 196, 212, 213, 227, 230, 235, 236, 243, 279, 287, 290, 291, 300, 303, 312, 322, 365, 372, 375, 377, 386, 405, 418, 435, 441, 444, 459, 465, 483, 489, 492, 529, 531, 549, 596, 602, 630, 649, 650, 653, 661, 662, 692, 708, 719, 749, 757, 784, 791, 808, 812, 819, 824, 827, 829, 832, 863, 866, 906, 912, 923, 951, 989, 993, 996, 1001, 1011, 1020, 1047, 1072, 1085, 1088, 1089, 1101, 1103, 1104, 1119, 1140, 1159, 1164, 1169, 1171, 1172, 1179, 1184, 1206, 1211, 1219, 1228, 1238, 1239, 1240, 1246, 1308, 1312, 1322, 1335, 1341, 1361, 1363, 1365, 1375, 1378, 1383, 1404, 1411, 1421, 1433, 1446, 1471, 1481, 1484, 1485, 1490, 1495, 1537, 1550, 1558, 1569, 1593, 1619, 1628, 1636, 1644, 1671, 1674, 1685, 1686, 1687, 1700, 1720, 1731, 1747, 1748, 1795, 1820, 1822, 1823, 1851, 1869, 1873, 1879, 1894, 1903, 1912, 1920, 1935, 1938, 1956, 1960, 1969, 1973, 1974, 1981, 1985, 1994, 1998]
4	Greens 114 votes	[2, 14, 20, 54, 58, 71, 97, 120, 121, 137, 157, 167, 185, 234, 238, 262, 296, 315, 342, 346, 356, 419, 423, 431, 436, 452, 468, 503, 519, 523, 533, 534, 542, 578, 638, 642, 666, 670, 672, 674, 697, 712, 715, 750, 778, 788, 830, 831, 859, 892, 903, 948, 997, 1014, 1043, 1051, 1053, 1066, 1077, 1095, 1115, 1120, 1126, 1181, 1222, 1300, 1307, 1318, 1324, 1337, 1377, 1380, 1419, 1441, 1473, 1483, 1493, 1502, 1526, 1554, 1556, 1561, 1573, 1584, 1590, 1597, 1601, 1649, 1652, 1701, 1711, 1714, 1717, 1727, 1745, 1751, 1761, 1767, 1778, 1787, 1789, 1832, 1850, 1858, 1863, 1878, 1882, 1883, 1884, 1886, 1887, 1950, 1978, 1982]
5	Turquoises 101 votes	[16, 38, 51, 68, 69, 93, 126, 154, 171, 172, 192, 193, 206, 208, 222, 242, 258, 268, 331, 341, 343, 344, 345, 354, 358, 395, 422, 440, 478, 486, 522, 530, 532, 617, 626, 646, 689, 722, 730, 752, 759, 792, 801, 814, 835, 840, 858, 869, 875, 944, 971, 976, 1013, 1062, 1106, 1143, 1231, 1258, 1265, 1287, 1292, 1297, 1332, 1339, 1391, 1405, 1445, 1457, 1461, 1474, 1509, 1543, 1545, 1552, 1560, 1567, 1614, 1615, 1620, 1632, 1650, 1651, 1654, 1684, 1702, 1733, 1755, 1764, 1777, 1798, 1809, 1891, 1904, 1943, 1946, 1953, 1964, 1966, 1976, 1980, 1992]
6	Reds 92 votes	[12, 32, 40, 45, 59, 74, 112, 118, 129, 133, 155, 178, 271, 298, 379, 413, 456, 473, 540, 550, 587, 598, 613, 627, 685, 702, 721, 734, 739, 755, 758, 786, 797, 822, 856, 864, 872, 874, 888, 904, 910, 922, 929, 938, 956, 987, 1000, 1006, 1008, 1018, 1019, 1071, 1086, 1087, 1099, 1108, 1127, 1173, 1185, 1212, 1252, 1266, 1272, 1305, 1309, 1334, 1352, 1355, 1406, 1408, 1464, 1478, 1501, 1566, 1582, 1603, 1607, 1631, 1683, 1726, 1754, 1756, 1765, 1776, 1784, 1805, 1847, 1862, 1871, 1897, 1916, 1921]
7	Blues 90 votes	[21, 41, 75, 77, 83, 138, 139, 166, 203, 228, 231, 244, 246, 304, 321, 323, 329, 374, 394, 415, 427, 453, 487, 510, 560, 562, 563, 584, 608, 637, 656, 694, 714, 720, 735, 768, 773, 844, 882, 926, 949, 964, 974, 1007, 1026, 1027, 1050, 1052, 1082, 1092, 1109, 1122, 1128, 1137, 1163, 1220, 1245, 1251, 1314, 1338, 1353, 1373, 1385, 1403, 1423, 1426, 1434, 1448, 1472, 1480, 1565, 1574, 1587, 1613, 1647, 1665, 1696, 1710, 1741, 1744, 1768, 1843, 1845, 1856, 1880, 1896, 1898, 1936, 1941, 1959]
8	Greys 88 votes	[17, 23, 89, 142, 194, 197, 264, 285, 316, 320, 335, 411, 479, 506, 511, 512, 516, 541, 567, 586, 609, 621, 622, 629, 641, 684, 700, 718, 769, 839, 847, 854, 883, 891, 966, 970, 992, 1016, 1033, 1039, 1040, 1083, 1084, 1150, 1174, 1262, 1270, 1277, 1319, 1320, 1327, 1342, 1356, 1371, 1374, 1386, 1392, 1393, 1395, 1397, 1455, 1513, 1524, 1600, 1624, 1625, 1627, 1635, 1658, 1660, 1689, 1697, 1775, 1783, 1799, 1802, 1811, 1818, 1833, 1835, 1867, 1868, 1872, 1915, 1918, 1928, 1933, 1967]
9	Oranges 85 votes	[4, 29, 36, 125, 159, 184, 189, 216, 219, 278, 325, 326, 334, 371, 381, 392, 393, 412, 430, 527, 554, 555, 575, 582, 583, 635, 706, 728, 743, 747, 770, 785, 821, 907, 940, 941, 943, 960, 1003, 1021, 1031, 1060, 1064, 1069, 1110, 1153, 1165, 1166, 1189, 1191, 1201, 1208, 1232, 1234, 1326, 1368, 1389, 1442, 1488, 1520, 1546, 1557, 1572, 1609, 1637, 1670, 1673, 1677, 1699, 1730, 1746, 1766, 1780, 1782, 1807, 1815, 1821, 1824, 1826, 1839, 1927, 1958, 1971, 1984, 1993]
10	Whites 82 votes	[27, 73, 135, 143, 161, 162, 173, 175, 207, 225, 249, 261, 270, 309, 327, 339, 347, 362, 438, 451, 460, 474, 508, 520, 528, 615, 623, 643, 677, 701, 772, 783, 838, 843, 845, 862, 870, 908, 911, 950, 1012, 1030, 1042, 1061, 1076, 1094, 1132, 1144, 1157, 1168, 1215, 1218, 1268, 1276, 1283, 1296, 1311, 1315, 1370, 1382, 1417, 1453, 1486, 1487, 1503, 1506, 1516, 1528, 1559, 1709, 1719, 1737, 1759, 1788, 1828, 1836, 1919, 1924, 1949, 1951, 1955, 1975]
11	Yellows 79 votes	[44, 50, 91, 105, 124, 132, 141, 144, 179, 272, 294, 333, 367, 388, 399, 443, 467, 484, 537, 581, 592, 600, 607, 636, 640, 676, 703, 709, 723, 765, 800, 809, 811, 823, 834, 920, 946, 959, 980, 981, 986, 999, 1112, 1142, 1158, 1204, 1217, 1257, 1264, 1284, 1317, 1323, 1343, 1412, 1418, 1425, 1429, 1430, 1499, 1523, 1525, 1527, 1532, 1562, 1576, 1598, 1604, 1618, 1645, 1668, 1704, 1712, 1722, 1750, 1769, 1825, 1889, 1977, 1991]

12	Violets 53 votes	[57, 111, 113, 182, 205, 273, 299, 324, 352, 359, 400, 424, 469, 655, 657, 704, 782, 794, 813, 817, 841, 895, 955, 961, 984, 1065, 1116, 1131, 1177, 1178, 1186, 1195, 1197, 1237, 1247, 1248, 1281, 1407, 1409, 1410, 1456, 1511, 1541, 1544, 1577, 1630, 1648, 1663, 1732, 1773, 1791, 1911, 1929]
13	Pinks 52 votes	[10, 30, 131, 198, 199, 239, 251, 252, 266, 275, 295, 408, 454, 470, 521, 612, 647, 754, 762, 804, 942, 962, 969, 972, 975, 985, 1028, 1036, 1068, 1075, 1180, 1298, 1346, 1349, 1431, 1450, 1477, 1489, 1494, 1518, 1534, 1549, 1553, 1555, 1606, 1643, 1655, 1713, 1739, 1816, 1830, 1923]

מטלת רשות – A

- פתרון מטלת החובה התבסס על הרצת פונקציית ה-main שבקובץ modeling: הפונקציה מתחילה בחלוקת הנתונים לקבוצות train/validation/set, נמשכת במציאת מסווג אופטימלי, ומסתיימת ביצירת תחזית עבור קבוצת ה-test על סמך מסווג זה.
- ריצת התוכנית נעשית ללא התערבות ידנית מצידנו באמצע הרצת הקוד, שכן בשלב מציאת המסווג האופטימלי אנחנו מתבססים על מדד ה-accuracy (שכאמור הראינו שהוא מצליח להראות את השוני בין המסווגים בצורה הטובה ביותר). ניתן לראות זאת בפונקציה "compare_performance_of_models" אשר מקבלת אובייקטים של מסווגים ומחזירה את המסווג שקיבל את ה-accuracy המקסימלי כאשר הוא אומן ע"י ה-training set ונבדק ע"י ה-validation set.
- לפיכך, לדעתנו הפתרון למטלת הרשות A הינו שימוש בקוד של מטלת החובה, שהרי הוא כולל בתוכו תהליך של בחירת model אופטימלי באופן אוטומטי.

מבנה תיקיית ההגשה

- הסבר על תהליך העבודה שלנו. — HW3.pdf
- קבצי ה-python שלנו: ב-modeling נמצאים כל השלבים המתוארים במסמך זה, וב-prepare_data נמצאים השלבים הדרושים להכנת הנתונים כפי שבוצעו ב-HW2. — modeling.py, prepare_data.py
- פלט של הרצת פונקציית ה-main שבקובץ modeling.py (התוצאות המפורטות בדו"ח הנ"ל מתבססות על הפלט הזה). — output.txt
- הנתונים המקוריים, מחולקים ל-train, validation ו-test, כוללים את כל ה-features. — original_data_train.csv, original_data_validation.csv, original_data_test.csv
- הנתונים לאחר העיבוד שלנו, מחולקים ל-train, validation ו-test, כוללים רק את 9 ה-features המצוינים בתרגיל וכן את "Vote". — prepared_data_train.csv, processed_data_validation.csv, processed_data_test.csv
- תחזית הבחירות של מצביעי קבוצת ה-test כפי שנחזו ע"י ה-model שבחרנו. — vote_result.csv