למידה חישובית – תרגיל 5

מיכל אברמוב, 301834297
בוריס בורשבסקי, 311898746

מה מצורף

1. מחברת המפרטת בשלבים את מה שעשינו (all_togather.ipynb)
2. קובץ all_togather.py המכיל את כל הקוד של התרגיל (בעצם המחברת בצורה ניתנת להרצה מהקונסול)
3. קובץ all_togather.html שמאפשר צפיה נוחה יותר במחברת
4. קבצי הdata מהתרגילים הקודמים והדאטה החדש.
5. קובץ predicted_new.csv שמכיל את מהתוצאות החזויות.

התרגיל

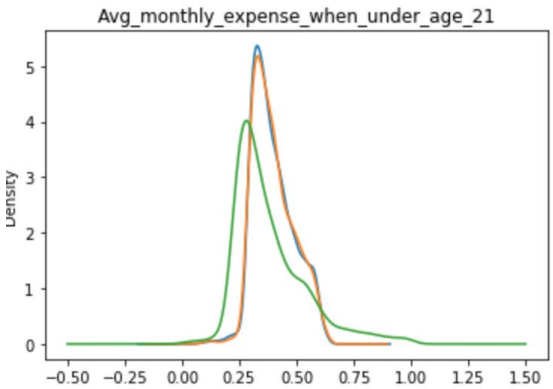
התרגיל מחולק 3 שלבים עיקריים:

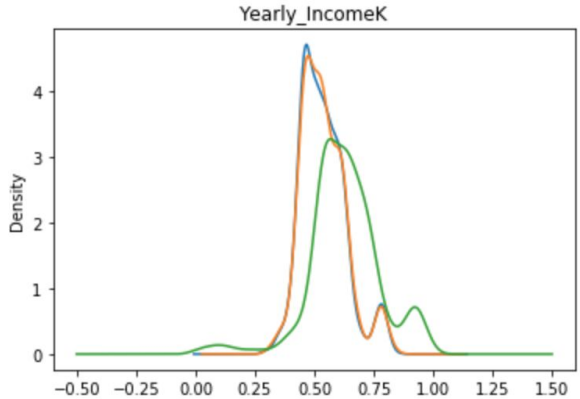
1. השלמת מידע וביצוע scale וטרנספורמציות לפי מה שעשינו בתרגיל 2
2. תחזית לפי מודל כפי שעשינו בתרגיל 3
3. הרכבת קואליציה לפי מה שעשינו בתרגיל 4

השלמת מידע, scale וטרנספורמציות

בתחילת התרגיל ניסינו להשלים את המידע בסט הואלידציה ובסט החדש באותה צורה, לאחר שביצענו זו ניסינו לבדוק את הביצועים שלנו את על סט הואלידציה והשליך מזה שהתוצאות על הסט החדש יהיו דומות אבל גילינו שהתוצאות שלנו על סט הואלידציה ממש לא טובות. למעשה שגיאה של כ30%.

ניסינו להסתכל מה השתבש בפיצ׳רים ולמעשה בהתפלגויות והבנו שפשוט לא השלמנו את המידע טוב למשל:





הקו הירוק מייצג את סט הואלידציה, הגענו למסקנה שפשוט השלמנו את המידע לא טוב!
בשלב זה החלטנו להשלים את המידע על כל הדאטא מחדש.

השתמשנו האותן שיטות של השלמת מידע שעשינו בתרגיל השני פשוט על כלל המידע, כאשר הוספנו גם את חוקים שהבנו רק אחרי אותו תרגיל.

תחזית

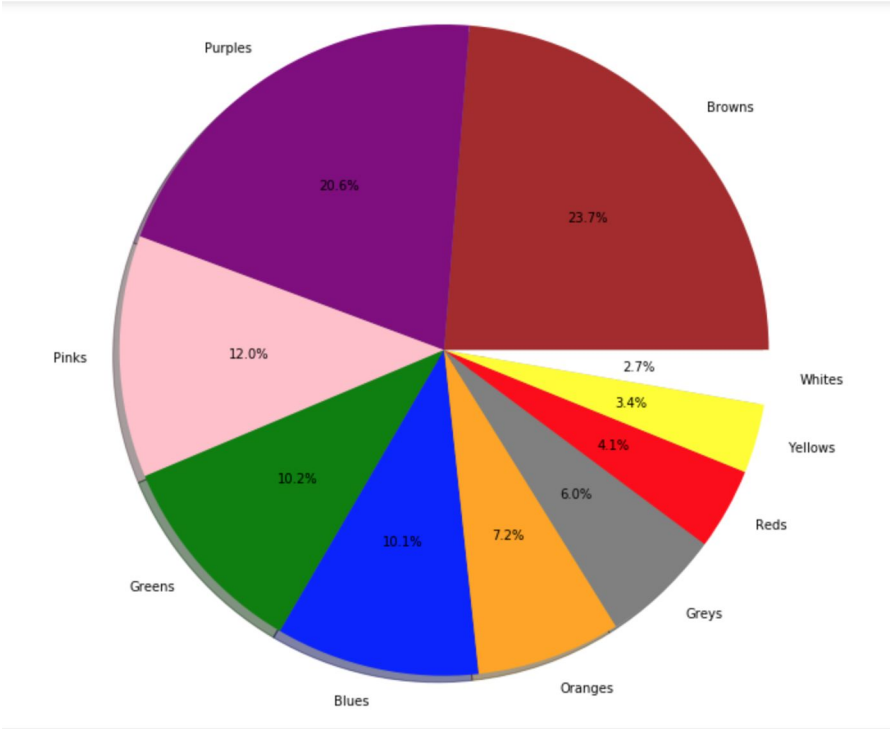
בגלל שהשלמנו את מידע מחדש בחננו בשנית את האלגורים בחרנו בתרגיל 3 RandomForestClassifier בעזרת cross validation ונראה שהוא נתן תוצאות יפות

```
***** RandomForestClassifier *****
             precision    recall  f1-score   support

      Blues    0.90807   0.98232   0.94374       905
     Browns    0.92690   0.98278   0.95402      1858
     Greens    0.99816   0.99754   0.99785      1629
      Greys    0.97037   0.95100   0.96059       551
    Oranges    0.93002   0.94569   0.93779       534
      Pinks    0.97536   0.90909   0.94106      1045
    Purples    0.98466   0.99371   0.98916      2067
       Reds    0.97212   0.96494   0.96852       542
     Whites    0.93380   0.79762   0.86035       336
    Yellows    0.96484   0.82364   0.88866       533

avg / total    0.96107   0.96030   0.95979     10000
```

כשהפעם השתמשנו בכל הדאטה הישן.

אנחנו חזינו שה **Browns** ינצחו בבחירות בדאטה החדש, להלן הפלגות הקולות:

```
Vote distribution
Browns - Votes: 2375 - Percents: 23.75%
Purples - Votes: 2057 - Percents: 20.57%
Pinks - Votes: 1201 - Percents: 12.01%
Greens - Votes: 1023 - Percents: 10.23%
Blues - Votes: 1014 - Percents: 10.14%
Oranges - Votes: 718 - Percents: 7.18%
Greys - Votes: 595 - Percents: 5.95%
Reds - Votes: 406 - Percents: 4.06%
Yellows - Votes: 343 - Percents: 3.43%
Whites - Votes: 268 - Percents: 2.68%
```


Purples
Browns
20.6%
23.7%
Pinks
12.0%
2.7%
Whites
3.4%
Yellows
4.1%
10.2%
6.0%
Reds
10.1%
7.2%
Greens
Greys
Oranges
Blues

קואליציה

בתרגיל הקודם ראינו שניתן לייצר קואליציה ע״י הקבוצה: Brown, Green, Pink, Purple,Whites
בתרגיל הזה בחרנו באותו אלגוריתם וקודם כל בדקנו אם המצב עדיין מתקיים, עם השלמת המידע החדשה ראינו טיפה זליגות של קולות לקלאסטרים שונים אך בגדול התוצאות נשארו זהות (אולי כי ביצענו outlier detection אחרת.)

```
____________________dist per cluster____________________
0 Counter({'Purples': 410, 'Browns': 352, 'Greens': 334, 'Pinks': 193, 'Whites': 58, 'Reds': 2, 'Yellows': 2, 'Grey
s': 1})
1 Counter({'Blues': 700, 'Yellows': 422})
2 Counter({'Purples': 415, 'Browns': 365, 'Greens': 296, 'Pinks': 244, 'Whites': 77, 'Yellows': 1})
3 Counter({'Purples': 414, 'Browns': 390, 'Greens': 313, 'Pinks': 204, 'Whites': 63})
4 Counter({'Purples': 418, 'Browns': 381, 'Greens': 350, 'Pinks': 189, 'Whites': 73, 'Yellows': 3})
5 Counter({'Purples': 410, 'Browns': 370, 'Greens': 336, 'Pinks': 215, 'Whites': 65, 'Blues': 1})
6 Counter({'Reds': 280, 'Oranges': 252})
7 Counter({'Greys': 300, 'Oranges': 282})
8 Counter({'Blues': 204, 'Yellows': 105})
9 Counter({'Reds': 260, 'Greys': 250})
```

בגלל שהזליגות הן בחלקיקי האחוזים החלטנו להשאיר את המצב כך.
בשלב זה אימננו את kmeans עם כל המידע ובדקנו:

- איזה קלאסטר צפוי להכנס כל בוחר במידע החדש
- למי אנו חוזים שכל בוחר במידע החדש הצביע
- האם הנתונים דומים?
- האם זה מעל 50% מהקולות ואפשר להרכיב מזה קואליציה?

מהמידע שקיבלנו הגדרנו את ההגדרות הבאות:

```
coalition = ['Purples', 'Browns', 'Greens', 'Pinks', 'Whites']
non_coalition = ['Greys', 'Oranges', 'Reds', 'Yellows', 'Blues']
coalition_clusters = [0,4,5,2,3]
```

המפלגות Brown, Green, Pink, Purple,Whites מתפזרות בין הקלאסטרים המוצגים ולכן סיממנו את הקלאסטרים.

עבור המידע החדש קיבלנו את התוצאות הבאות:

- מתוך הקולות החדשים **6926** הצביעו למפלגות שהם בקואליציה שלנו (69%) מה שנותן לנו רוב.
- מתוך הקולות החדשים **6924** נמצאו שייכים לקלאסטרים של הקואליציה.
- מתוך הקולות החדשים **6922** גם הצביעו למפלגה בקואליציה וגם שייכים להם ברמת הקלאסטרים.

מסקנות:

- בגלל התפלגות הקולות בין הקלאסטרים מאוד דומה להתפלגות הקולות, וכבר ראינו בתרגיל הקודם שניתן לחבר כמה קלאסטרים ״קרובים״ אז ניתן להסיק שמרבין הקולות שהצביעו לקבוצה שלנו די דומים ביניהם ודי שונים מכל השאר. ולכן הם יהיה הקואליציה שלנו, שהיא קואליציה יציבה.
- ראינו שקיים 4 של קולות שהם התפזרו לקלאסטר אחר, המצב סביר כיוון הקלאסטרים נבחרים לפי מרחק אוקלידי בין כל הפיצ׳רים, גם הפחות משמעותיים לבחירת הקול