תיעוד:

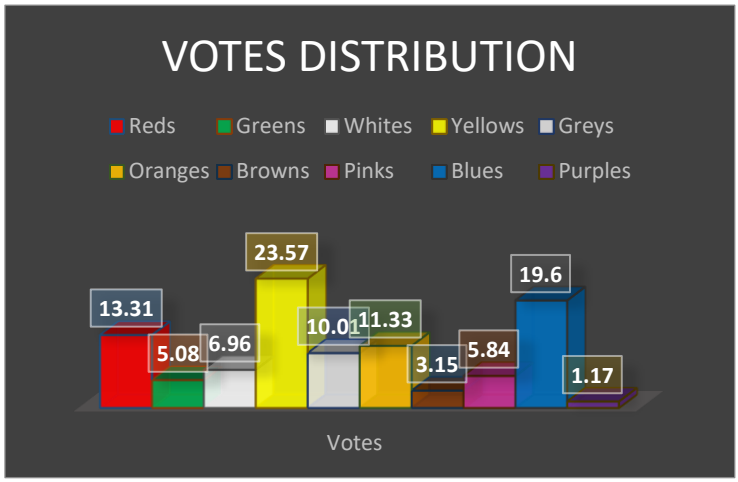
1) הסבר כללי על התהליך:

1. אחרי טעינת הנתונים השארנו רק את עמודות הלוונטיות ויצרנו 2 עותקים ממנו: data_raw ו-data_prepared. את הראשון השארנו כמו שהוא ועל השני ביצענו ה-preparation data אשר כולל:
   a. set_correct_types – השמה של טיפוסים נכונים.
   b. impute_data – השלמה של שדות חסרים.
   c. cleanse_data – ניקיון הנתונים מסטיות ותצפיות לא הגיוניות.
   d. normalize_data – נרמול הנתונים.
2. בשלב זה פיצלנו את שני העותקים ל-train, test ו-validate ושמרנו לקבציי csv.
3. יצרנו 4 מודלים:
   a. Neural Network – תוך שימוש ב-optimize gradient decent של sklearn.
   b. Decision Tree – מסוג CART המבצע סיווג ורגרסיה.
   c. Naive Bayes – את likelihood חישבנו בעזרת הנוסחה: $\frac{e^{\frac{\bar{x}}{2var}}}{\sqrt{2\pi var^2}}$
   d. KNN – עם k=5.

   את כל המודלים יצרנו בעזרת המודול sklearn.
4. אימנו את המודלים ובחרנו את הכי טוב בעזרת cross-validation כאשר המדד שלנו היה הדיוק תחזית. יש מתודה במודול sklearn שמבצעת את ה-cross-validation .
5. לאחר שבחרנו את המודל הכי טוב, במקרה שלנו קיבלנו שרשת נוירונים בעלת ביצועים הכי טובים עם אחוז דיוק של 88, אימנו שוב את המודל הזה היות והאימון שעשינו קודם בוצע בתוך המתודה של sklearn כחלק מ-cross-validation ולכן היה עלינו בשלב זה לאמן שוב את המודל הנבחר.
6. בשלב זה דרשנו מהמודל שלנו לספק תחזית לסט ה-test.
7. חישבנו את מטריצת הבלבול, אחוז דיוק ושגיאה.

2) מהתוצאות שלנו נובע שהמפלגה הזוכה היא Yellows עם בערך 24 אחוז מהקולות.

3) חלוקת הקולות פר מפלגה:



4) מטריצת בלבול (confusion matrix):

	Reds	Greens	Whites	Yellows	Greys	Oranges	Browns	Pinks	Blues	Purples
Reds	284	0	0	3	0	0	0	21	0	21
Greens	0	62	0	0	0	1	0	0	0	1
Whites	0	0	100	0	0	0	0	0	4	1
Yellows	2	0	0	174	15	0	0	0	0	8
Greys	0	0	0	11	195	0	0	0	0	9
Oranges	0	0	0	0	0	112	0	0	3	0
Browns	0	0	0	0	0	0	23	0	0	0
Pinks	39	0	0	0	0	0	0	240	0	0
Blues	0	0	0	0	0	0	2	0	130	0
Purples	61	0	0	9	13	0	0	1	0	424

5) אחוז שגיאה: 0.114271203657.