**מבוא למערכות לומדות – תרגיל בית 2**

מגישים:

**שם:** אסף חאייק ברוך **ת.ז:** 206783441

**שם:** בן בנוז **ת.ז:** 207570573

**הסבר על פתרון הבעיות**

כדי לפתור את הבעיות שהוצגו בתרגיל השתמשנו במסווגים שפותרים את בעיית קביעת ההצבעה של בוחר לפי הנתונים עליו (הבעיה הבסיסית), לכן לכל בעיה הגדרנו מעטפת שתהיה אחראית על מסווג סטנדרטי, הגדרנו לכל בעיה גם פונקציה לשערוך טיב הפלט של מודל שפותר בעיה.  
*הערה: בהמשך נשתמש במונח 'וקטור התוצאות' של בחירות.  
הכוונה היא שניצור היסטוגרמה של הקולות לפי התפלגות ההצבעה למפלגות של הבוחרים (הערך לכל מפלגה הוא סכום ההסתברויות אצל כל הבוחרים שהם יצביעו לה, עבור תוצאות אמיתיות ההסתברות אצל כל בוחר היא 1 למפלגה שהוא הצביע לה ו-0 לאחרות) וננרמל אותה כך שסכום הערכים של כל המפלגות הוא 1 – וזהו וקטור התוצאות.*

1. סיווג רגיל  
   במקרה של בעיית הסיווג הבסיסית של כל בוחר למפלגה לא השתמשנו במעטפת, ופונק' ההערכה הייתה הדיוק, אבל מאוזן לפי הסיווגים האפשריים (balanced\_accuracy\_score של sklearn).
2. המנצח בבחירות

החלטנו שפתרון לבעיה זו יכול להיגזר מהפתרון לבעיה השנייה של התפלגות הקולות בין המפלגות, ולכן החלטנו להשתמש פחות או יותר באותו הפתרון רק עם פונקציית הערכה שונה.

פונקציית ההערכה במקרה הזה משתמש בוקטורי התוצאות של החיזוי והאמת  
כ-logits של softmax כך שטעויות בחיזוי אצל המפלגות המובילות הרבה יותר משפיעות על הערכת טיב החיזוי.

אנחנו מחשבים את המרחק האוקלידי בין וקטור התוצאות של החיזוי לוקטור התוצאות של האמת (לאחר softmax של שניהם) ומעדיפים ערכים יותר קרובים ל-0 (ולכן מחזירים את המרחק כפול ).

1. תוצאות הבחירות

במקרה הזה החלטנו השתמש במרחק האוקלידי בין וקטורי התוצאות של החיזוי ושל האמת בשביל להעריך את איכות הפיתרון.  
אנחנו רוצים לצמצם את המרחק הזה (מרחק 0 זה הכי טוב) ולכן נחזיר את המרחק כפול .

1. מצביעים סבירים לכל מפלגה

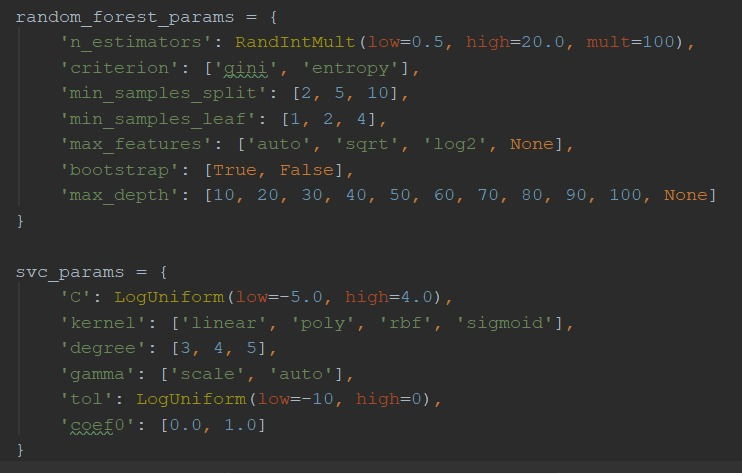
בדומה לתהליך שביצענו בפתרון הבעיות הקודמות חישבנו את ההסתברות של כל מצביע להצביע לכל מפלגה והחזרנו לכל מפלגה רשימה של מצביעים שיש להם הסתברות גבוהה מערך שהמשתמש הגדיר (threshold) להצביע לה, אנחנו בחרנו להשתמש בערך דיפולטי של 60% .

פונקציית ההערכה במקרה זה הינה ממוצע של ניקודי ה-f1 בין המצביעים שקבענו שסביר שיצביעו למפלגה מסויומת לבין האנשים שבאמת הצביעו לאותה מפלגה (ממוצע על כל המפלגות).  
כלומר נחשב f1 לכל מפלגה כאשר יש סיווג בינארי של 'סביר' ו-'לא סביר' ונעשה ממוצע, שנרצה להגיע לערך מקסימלי אצלו.

**המודלים שבחרנו לעבוד איתם**

החלטנו לבחור מודל לינארי (או חישובי דומה) ומודל של עץ החלטה ולפי צרכי הבעיות היינו חייבים שלאותם מסווגים תהיה פונקציה שמחזירה הסתברות לכל סיווג. בגלל מגבלות אלה ואיכות המסווגים החלטנו להשתמש ב-SVC וב-Random Forest.  
יתרון נוסף של המסווגים האלו הוא שהם יכולים לאזן את הדוגמאות שלהם באופן ששקול לאיזון סט האימון, מה שאומר שלא היינו צריכים לאזן אותו עבורם.

כיוונון היפר פרמטרים  
בחרנו לכוונן היפר פרמטרים לכל בעיה לפי חיפוש אקראי על מרחב פרמטרים שהגדרנו.

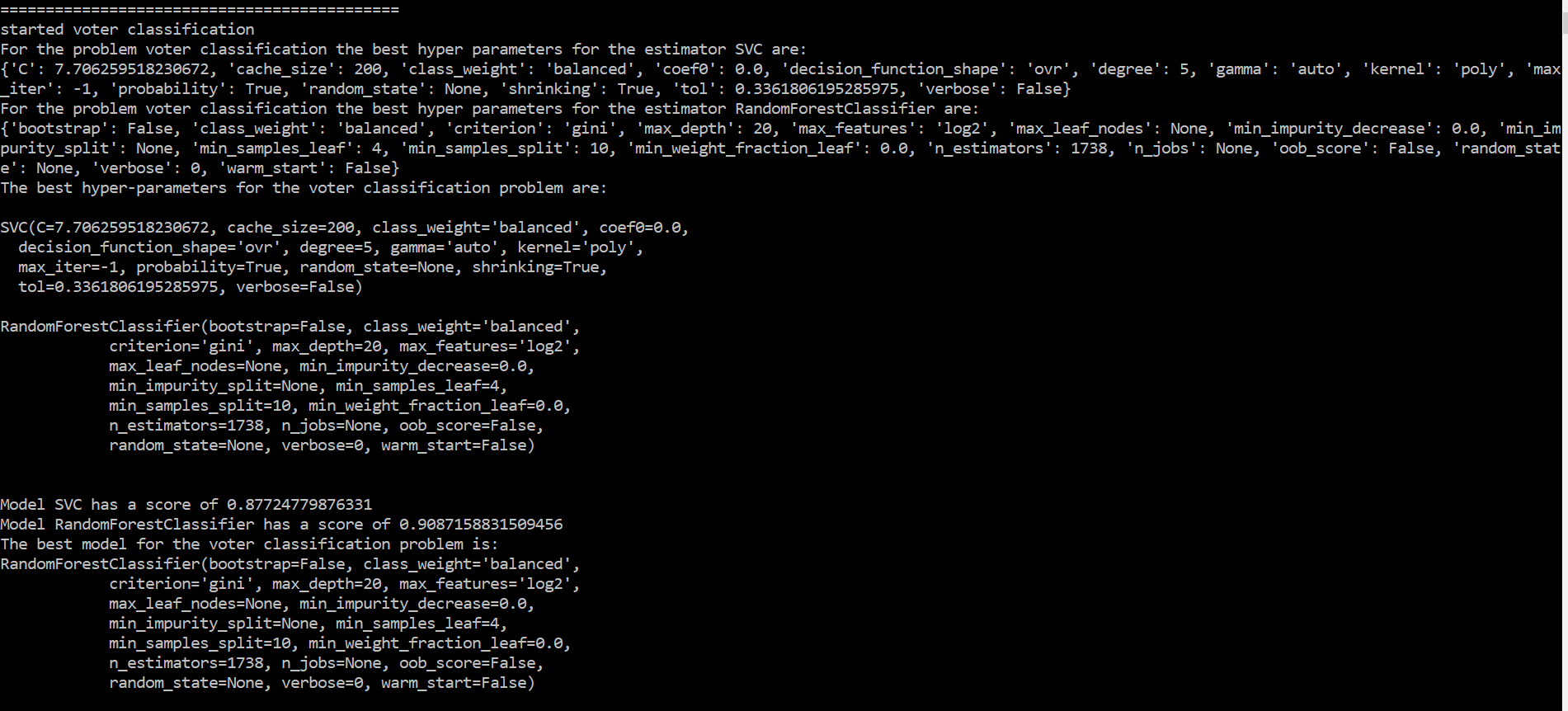
לכל פרמטר שבמרחב הגדרנו התפלגות ערכים אקראית שתעבוד עם RandomizedSearchCV שעושה חיפוש אקראי על היפר פרמטרים עם  
k-Fold cross validation לפי פונקציות ההערכה שהגדרנו.  
החלוקה נעשית ע"י אותו המצב האקראי לכל המודלים שנבדקים.  
גם ווידאנו שהחיפוש נעשה עם אותו מצב אקראי לכל מסווג. ההתפלגויות שבחרנו היו בטווחי ה-log והמכפלות מכיוון שהיה לנו חשוב יותר למצוא את סדר הגודל של ההיפר פרמטרים מאשר הגודל המדויק, שלדעתנו פחות מהותי.  
אנחנו בחרנו להשתמש ב- עבור הכיוונון וב-10 ניחושים של הפרמטרים.  
בנוסף לכך וידאנו שלכל מסווג שלא מאזן את הסט לפי הסיווגים תתקבל גרסא מאוזנת של סט המבחן.  
מימשנו גם ניקוד לפי k-Fold cross validation משלנו אבל בסוף לא השתמשנו בפונקציה הזאת בגלל שזה נעשה אוטומטית בחיפוש, השארנו את הפונק' למקרה שנצטרך אותה בתרגילים הבאים.  
לאחר מציאת ההיפר-פרמטרים אנחנו שומרים אותם בקבצים שבתיקייה, כך שנוכל לדלג על השלב הזה לאחר שמצאנו אותם ופשוט להשתמש בערכים השמורים (נעשה אם לא משתמשים בדגל -s בקריאה לקובץ הראשי, אם משתמשים בו אז החיפוש נעשה).

בחירת המודל האופטימלי לבעיה  
לאחר שכיווננו את הפרמטרים של המסווגים שלנו אימנו אותם על כל סט האימון והשווינו את החיזוי שלהם על סט הוולידציה והערכנו את התוצאות לפי פונקציות ההערכה שלנו.

לאחר מכן לכל בעיה פשוט בחרנו את המסווג עם הניקוד הכי גבוה לפי פונק' ההערכה.

**התוצאות**  
כאן נציג את התוצאות שמצאנו.

היפר פרמטרים ובחירת מודלים:  
קודם נציג לכל בעיה את ההיפר-פרמטרים שנמצאו לה ואת הניקוד של המודלים על סט ה-validation, בנוסף למודל שנבחר בהתאם לניקוד.

1. סיווג רגיל  
   
2. המנצח בבחירות

A screenshot of a cell phone

Description automatically generated

1. תוצאות הבחירות

A screenshot of a cell phone

Description automatically generated

1. מצביעים סבירים לכל מפלגה

A screenshot of a cell phone

Description automatically generated

נשים לב שלפי ההיפר-פרמטרים שנמצאו הבעיות של חיזוי המנצח בבחירות וחיזוי תוצאות הבחירות באמת שקולות – בשתי הבעיות התקבל כי אותם ההיפר-פרמטרים גורמים לתוצאות הכי טובות גם ב-SVC וגם ב-Random Forest.

תוצאות על סט המבחן:  
כעת נציג את תוצאות הסיווג על סט המבחן.  
תוצאות הבעיה הבסיסית שמורות בקובץ test\_predictions.csv.

1. מטריצת הבלבול על הבעיה הבסיסית:  
   A screenshot of a cell phone

   Description automatically generated
2. על טסט המבחן חזינו שהמפלגה הסגולה תנצח, וזה אכן מה שקרה.
3. התפלגות הקולות בין המפלגות:
4. סיווג בעיית המצביעים הסבירים על סט המבחן:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Oranges | Khakis | Greys | Greens | Browns | Blues |
| 2, 13, 19, 40, 96, 124, 131, 145, 158, 248, 249, 268, 305, 329, 424, 497, 533, 538, 557, 575, 582, 603, 625, 656, 703, 774, 776, 778, 788, 797, 804, 826, 897, 929, 990 | 6, 7, 10, 12, 20, 37, 46, 52, 62, 68, 86, 108, 109, 121, 127, 140, 142, 151, 152, 164, 166, 169, 172, 196, 206, 215, 217, 223, 228, 250, 261, 280, 282, 284, 285, 289, 290, 304, 313, 320, 328, 331, 332, 354, 361, 367, 371, 373, 374, 375, 381, 394, 398, 399, 413, 415, 420, 429, 430, 431, 432, 437, 438, 447, 448, 453, 466, 469, 471, 478, 480, 481, 483, 487, 494, 500, 505, 510, 513, 517, 518, 535, 547, 552, 554, 559, 562, 563, 570, 571, 574, 581, 591, 597, 600, 605, 606, 616, 621, 626, 635, 641, 643, 645, 652, 663, 668, 676, 681, 683, 684, 688, 689, 694, 695, 700, 714, 719, 723, 738, 741, 742, 749, 751, 752, 754, 757, 758, 760, 763, 764, 766, 773, 775, 777, 789, 790, 794, 798, 811, 817, 830, 832, 836, 852, 855, 859, 873, 874, 875, 885, 890, 903, 916, 923, 930, 931, 936, 938, 942, 948, 954, 957, 961, 962, 967, 971, 974, 982, 987 | 32, 54, 95, 119, 137, 154, 219, 241, 256, 312, 343, 344, 358, 360, 376, 385, 412, 422, 423, 427, 461, 501, 502, 504, 508, 539, 550, 584, 585, 619, 674, 678, 697, 765, 810, 834, 837, 871, 878, 880, 906, 920, 946, 966 | 1, 18, 31, 43, 63, 67, 70, 93, 103, 104, 128, 130, 133, 161, 182, 184, 187, 195, 232, 233, 257, 270, 288, 296, 307, 347, 404, 410, 451, 455, 477, 489, 531, 549, 579, 583, 593, 598, 612, 613, 615, 623, 636, 637, 661, 672, 682, 712, 718, 768, 772, 847, 918, 921, 955 | 8, 25, 29, 30, 50, 51, 55, 58, 69, 83, 85, 98, 102, 134, 138, 139, 144, 149, 156, 163, 171, 173, 186, 190, 205, 252, 255, 258, 259, 273, 311, 337, 348, 349, 350, 370, 372, 379, 383, 388, 392, 393, 401, 405, 406, 408, 416, 445, 456, 458, 459, 470, 488, 492, 495, 506, 528, 556, 564, 568, 578, 590, 601, 629, 653, 655, 660, 664, 666, 670, 699, 705, 715, 724, 739, 755, 780, 783, 802, 838, 840, 850, 856, 861, 900, 907, 917, 925, 933, 939, 976, 983 | 4, 49, 53, 57, 59, 61, 73, 76, 84, 200, 234, 242, 264, 272, 283, 292, 302, 359, 369, 387, 395, 414, 439, 515, 522, 527, 530, 534, 542, 548, 561, 587, 589, 628, 657, 707, 736, 740, 748, 782, 819, 833, 842, 843, 849, 913, 940, 944, 952, 968, 977 |

|  |  |  |  |
| --- | --- | --- | --- |
| Turquoises | Reds | Purples | Pinks |
| 15, 117, 141, 204, 231, 236, 245, 253, 260, 262, 274, 339, 340, 362, 384, 390, 441, 468, 553, 577, 709, 733, 753, 759, 762, 800, 823, 845, 904, 934, 972, 989 | 17, 38, 42, 47, 78, 132, 136, 143, 146, 153, 155, 177, 207, 298, 300, 316, 325, 333, 334, 365, 396, 397, 407, 433, 442, 444, 454, 512, 576, 610, 634, 665, 680, 713, 771, 792, 809, 846, 851, 858, 860, 864, 896, 919, 953, 960, 975, 979, 986 | 3, 5, 9, 21, 22, 33, 35, 36, 39, 45, 56, 60, 65, 72, 77, 79, 80, 81, 89, 97, 99, 100, 105, 110, 112, 116, 118, 120, 126, 129, 147, 150, 159, 160, 162, 165, 167, 168, 170, 174, 175, 176, 181, 183, 185, 189, 191, 203, 213, 214, 216, 218, 220, 221, 225, 226, 230, 235, 237, 246, 263, 265, 267, 269, 271, 275, 276, 279, 281, 291, 294, 295, 297, 299, 308, 309, 310, 314, 315, 317, 318, 319, 321, 322, 327, 336, 342, 345, 346, 355, 356, 366, 377, 378, 382, 389, 400, 403, 409, 417, 419, 421, 426, 434, 436, 440, 443, 449, 462, 463, 465, 473, 476, 479, 486, 490, 491, 493, 496, 499, 503, 514, 516, 519, 520, 524, 525, 532, 537, 541, 544, 555, 558, 565, 566, 567, 569, 572, 573, 588, 592, 594, 595, 596, 602, 607, 609, 611, 614, 620, 622, 624, 627, 638, 644, 648, 654, 658, 659, 662, 669, 679, 685, 687, 690, 691, 692, 696, 701, 708, 710, 711, 722, 725, 726, 728, 729, 730, 731, 732, 734, 737, 743, 745, 746, 747, 750, 767, 770, 779, 781, 784, 795, 799, 803, 805, 806, 807, 808, 815, 822, 824, 825, 828, 829, 848, 854, 862, 867, 868, 869, 870, 877, 886, 887, 889, 891, 892, 894, 899, 901, 902, 905, 908, 909, 915, 922, 927, 932, 937, 941, 943, 945, 947, 950, 951, 964, 965, 969, 973, 980, 981, 984, 985, 988 | 11, 16, 74, 91, 101, 106, 114, 122, 338, 353, 363, 435, 450, 457, 484, 507, 509, 630, 721, 812, 884, 895 |

|  |  |  |  |
| --- | --- | --- | --- |
| None (not above 60% for any party) | Yellows | Whites | Violets |
| 14, 23, 24, 26, 34, 41, 44, 48, 64, 66, 71, 87, 92, 107, 111, 115, 125, 148, 157, 178, 179, 180, 188, 192, 193, 194, 197, 198, 199, 201, 208, 209, 210, 212, 227, 229, 239, 240, 244, 266, 286, 287, 293, 301, 303, 326, 330, 352, 357, 364, 368, 391, 418, 425, 460, 464, 467, 472, 482, 498, 523, 536, 540, 545, 551, 560, 586, 604, 608, 618, 631, 633, 639, 646, 650, 651, 667, 671, 675, 693, 698, 706, 717, 720, 727, 735, 744, 785, 786, 787, 793, 796, 801, 816, 820, 827, 839, 841, 844, 853, 863, 872, 876, 879, 882, 888, 893, 898, 910, 912, 924, 926, 928, 949, 956, 958, 959, 963, 970 | 75, 90, 123, 238, 251, 324, 341, 351, 411, 474, 617, 647, 677, 716, 756, 883, 935 | 0, 27, 28, 82, 88, 94, 113, 135, 222, 224, 243, 254, 306, 323, 380, 386, 402, 475, 485, 511, 521, 546, 580, 640, 649, 761, 769, 814, 818, 821, 831, 835, 865, 866, 881, 911, 914, 978 | 202, 211, 247, 277, 278, 335, 428, 446, 452, 526, 529, 543, 599, 632, 642, 673, 686, 702, 704, 791, 813, 857 |