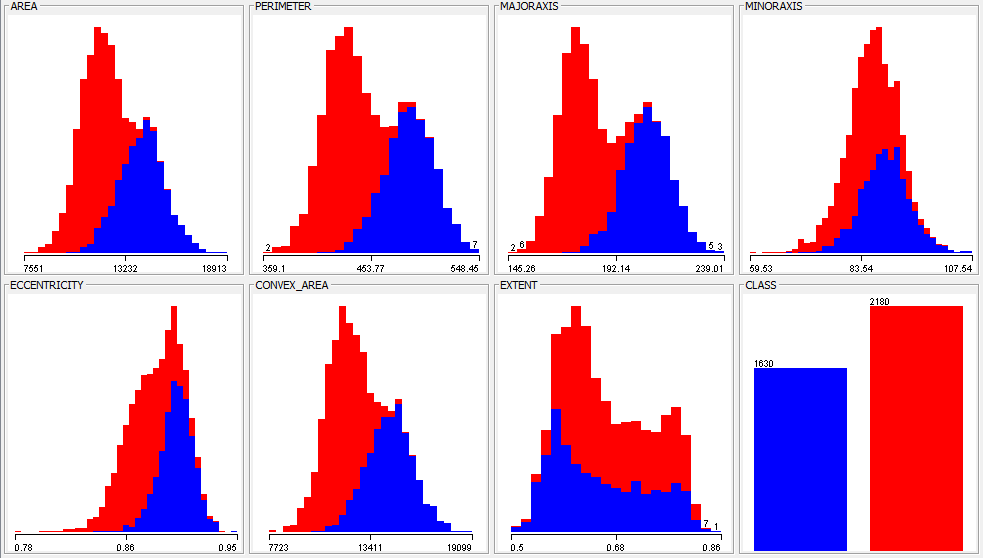
**כותרת**

1. בחרנו במאגר , המתאר שני סוגי דגנים של אורז הגדלים בטורקיה. המאגר מכיל דוגמאות המכילות משתנים - פיצ'רים מורפולוגיים המתארים את צורת הדגן בערכים נומריים ומשתנה תיוג קטגורלי () לסיווג הדגן.   
   פירוט המשתנים בטבלה הבאה:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Returns the number of pixels within the boundaries of the rice grain. | Area | 1 |
| Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain. | Perimeter | 2 |
| The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives. | Major Axis Length | 3 |
| The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives. | Minor Axis Length | 4 |
| It measures how round the ellipse, which has the same moments as the rice grain, is. | Eccentricity | 5 |
| Returns the pixel count of the smallest convex shell of the region formed by the rice grain. | Convex Area | 6 |
| Returns the ratio of the region-formed by the rice grain to the bounding box pixels. | Extent | 7 |
| Cammeo and Osmancik rice | Class | 8 |

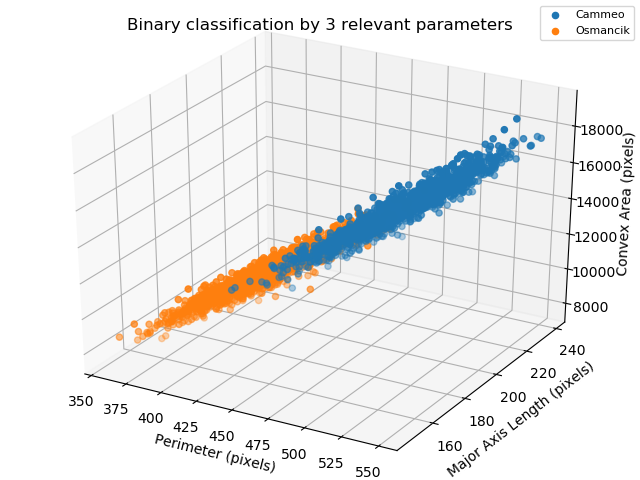
ערכי הפיצ'רים נקבעו בעזרת תמונות של הדגנים, מיוצגים בפיקסלים, כפי שמצוין במאמר . לטובת ניתוח הנתונים בוצעה המרה של התיוג הקטגורלי בתיוג בינארי:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

  
מכיוון שאנו לא מתמצאים בסוגי האורז, על מנת לקבל הבנה כללית של ה ומשתניו השתמשנו בתצוגה ויזואלית להפרדת ה על פי כל פיצ'ר בנפרד בעזרת . בהמחשה הויזואלית בחיפוש אחר המשתנים שמפרידים את ה בצורה טובה, נראה כי המשתנים (מהטבלה מעלה) יוכלו להיות רלוונטיים בהמשך לניבוי סוג האורז.

ב אין נתונים חסרים, נראה שהמכשול העיקרי עשוי להיות חוסר איזון בכמות הדוגמאות, הן מתפלגות ביחס של .

1. בוצעה חלוקה של ה לסט אימון וסט מבחן ביחס של . בוצע נרמול לכל הדוגמאות.
2. על מנת להתרשם מה בחרנו להציג את דוגמאות האימון ב תלת מימדי בעזרת משתנים בהפרדה למשתנה מנובה. ציר מייצג את משתנה , ציר מייצג את משתנה () וציר מייצג את משתנה (). ניתן לראות שכבר במימד שלישי בעזרת המשתנים הנ"ל קיימת הפרדה מסוימת הנראית לעין בין שני המשתנים המנובים.



1. על מנת לסווג את ה בחרנו ב שיטות קלסיפיקציה -

כדי לבחון את מידת ההצלחה של כל שיטה על ה , ראשית בחרנו היפר-פרמטרים סבירים, פונקציית פשוטה ומדד להערכת המודל. בעזרת כל שיטת קלסיפיקציה חילקנו את ה ל קבוצות שוות בגודלן, וביצענו למידה בשתי דרכים – פעם אחת בעזרת למידה על 4 מתוך 5 הקבוצות ואבלואציה על הקבוצה החמישית, ופעם נוספת בעזרת ומיצוע.  
לאחר מכן, חזרנו על התהליך עם מקדם רגולריזציה, ולבסוף השתמשנו בתוצאות כדי לבצע נוסף להיפר פרמטרים הראשונים שבחרנו לטובת מודל סופי.

כל המודלים שנציג משתמשים במדד דיוק להערכת המודל:

עבור מסווג בחרנו בשימוש ב פולינומי, מהצורה הבאה:

כאשר עבור סיפוק ההיפר-פרמטרים בתנאי הראשוני השתמשנו בערכים לקבלת פונקציית ה: . עבור המסווג בחרנו בפונקציית :

בעיית האופטימיזציה שמנסים לפתור :

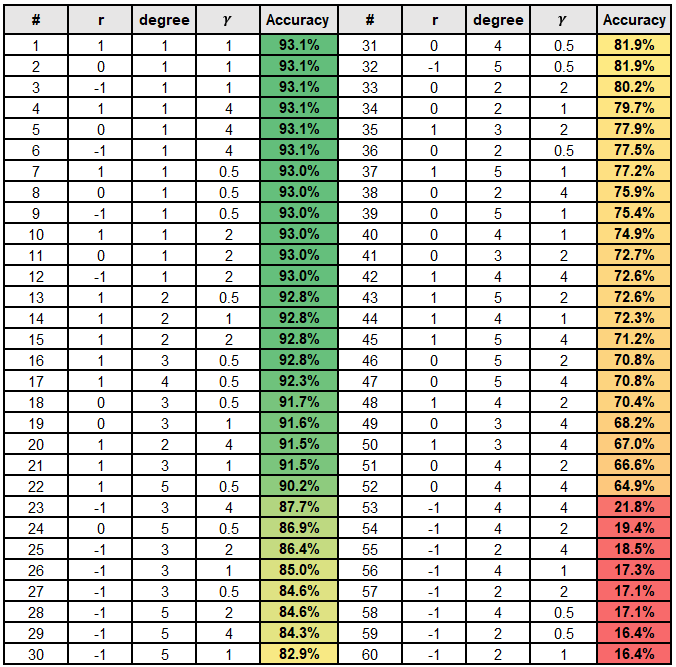
שלב הראשון בחנו את המודל ללא רגולריזציה () והגבלנו את מס' האיטרציות ל במידה ולא בוצעה התכנסות עד לנק' זו (אכן, ללא רגולריזציה המודל לא התכנס). בשלב השני הוספנו מקדם רגולריזציה () למודל , בחרנו את המקדמים .  
נציג את מדד הערכת המודל כפונקציה של מס' הדוגמאות שסיפקנו למודל.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

*התוצאות שקיבלנו מראות ששימוש ב מביא לאחוזי דיוק יציבים יחסית לאורך הלמידה (למול גודל משתנה) בהשוואה ללמידה שבוצעה על מהדטה והוערכה למול סט ולידציה בודד, כלומר, טווח אחוזי הדיוק משתנה בצורה ניכרת יותר לפי גודל סט האימון כשלא מבצעים מיצוע.  
בנוסף, שימוש ברגולריזציה מסייע ללמידה: בשורה הראשונה בטבלה ניתן לראות למידה ללא רגולריזציה, מכיוון שמדובר בבעיית מינימום, שקול להשפעה גדולה של המשקולות. בלמידה הזאת המודל לא התכנס, ואחוזי הדיוק של המודל נעים סביב על סט הולידציה, ויורדים ככל שיש יותר דוגמאות. בשורה השניה בטבלה ניתן לראות למידה עם מקדם רגולריזציה , ובשורה השלישית עם מקדם רגולריזציה של . עבור קטן (אולי מדי) ניתן לראות בשורה השלישית ירידה באחוזי הדיוק לעומת השורה השניה, כיוון שאנחנו מאפשרים יותר טעויות על דוגמאות סוררות. מהתוצאות עושה רושם ששימוש במקדם מביא לאחוזי הדיוק טובים עבור מסווג ל שלנו.*

*על מנת לבחור היפר-פרמטרים טובים ולמקסם את יכולות המודל נשתמש במקדם שמצאנו לטובת הצגת עבור 3 הפרמטרים של פולינומי : דרגת הפולינום, משתנה חופשי, מקדם למכפלה הפנימית. נציג בטבלה את אחוזי הדיוק שקיבלנו עבור הפרמוטציות השונות של משתנים אלה בערכים:*

*מתוך כלל האפשרויות המודל יבחר את הפרמטרים הטובים ביותר.*



# הערה על כמות איטרציות max\_iter בכל המודלים

ביצוע למציאת קו הרגרסיה. המסווג משתמש בפונקציית :

בעיית האופטימיזציה שמנסים לפתור :

גם כאן בחנו את המודל ללא רגולריזציה () והגבלנו את מס' האיטרציות ל, זה הספיק כדי שהמודל התכנס בהרצות שביצענו. כשהוספנו מקדם רגולריזציה בחרנו את המקדמים .  
נציג את מדד הערכת המודל כפונקציה של מס' הדוגמאות שסיפקנו למודל.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

השימוש ב במודל הנוכחי לא השפיע בצורה ניכרת על התוצאות, אך ראינו שללא שימוש ברגולריזציה, בסט ולידציה בודד אחוזי הדיוק על סט הולידציה מתחילים לרדת ככל שמשתמשים ביותר דוגמאות, בעוד שבסט האימון האחוזים יותר יציבים, דבר שיכול להצביע על . באשר לשימוש במקדם רגולריזציה, עבור ה שלנו ניתן להעריך מהתוצאות ששימוש ברגולריזטור או חוסר שימוש מביאים לביצועים די דומים, כל עוד קבוע הרגולריזציה בגודל סביר (למשל, בגרפים מעלה הוצג שימוש ב ). עם זאת, עבור קבוע קטן () אחוזי הדיוק על הדוגמאות מתחילים להיפגע, וניתן אף לראות שללא מיצוע אחוזי הדיוק על סט הולידציה נמוכים מאשר על סט האימון.

1. ד
2. ד
3. ד
4. ד
5. ד
6. ד
7. ד

Mean Accuracy as function of training set size, cv (5), C=0.005

Mean Accuracy as function of training set size, cv (5), C=10

Accuracy as function of training set size, single validation set, C=0.005

Accuracy as function of training set size, single validation set, C=10

רפרנסים:

CINAR, I. and KOKLU, M., (2019). “Classification of Rice Varieties Using Artificial Intelligence Methods.” International Journal of Intelligent Systems and Applications in Engineering, 7(3), 188-194.

ציטוט וקה