Noy Jaffe, Omer Siton, Ori Tabib, Shanni Chared

IML Hackaton 2022

תיאור ה-Dataset תוך התייחסות לאתגרים:

ה-Dataset איתו התמודדנו הכיל 34 פיצ'רים, אשר מרביתם הכילו ערכים שאינם מספריים או לחלופין ערכים המכילים קטגוריות רבות. על מנת להתמודד עם האתגרים השונים בוצעו הפעולות הבאות:

- עבור עמודות אשר הכילו ערכי NA בוצעה החלפה בערכי ממוצע, חציון או בערכים מספריים אחרים בהתאם לסוג ואופי הקטגוריה (פיצ'ר).
- עבור עמודות אשר ערכיהן נחלקו לשתי קטגוריות עיקריות כגון "חיובי" ו-"שלילי" (המפורטות מטה) ביצענו חקירה מעמיקה בניסיון להבין את משמעות הקטגוריה וחשיבות הסיווג שבוצע ע"י הרופא המאבחן. למשל, עבור עמודת האבחון ההיסטולוגי, היה נראה לנכון כי יש לחלק את הקטגוריות לערכי (+,-) באמצעות אפיון וזיהוי מונחים רפואיים (כלומר, "שמות קוד") המעידים על דרגה חמורה יותר/פחות של המחלה.
- עבור עמודות אשר כללו ברובן ערכים חסרים אך חשיבותם לניבוי ניכרת, חשבנו רבות על הדרך הנכונה לקבלת סיווג נכון ומדויק. דבר זה מהווה אתגר משמעותי מאחר ומדובר בסיווג שעלול להטות באופן קריטי את הלומד ואף להוביל לטעויות. דוגמא מובהקת לחשיבות זו הינה עמודת "רוחב הגידול". בעמודה זו, שורות ריקות אינן בהכרח מעידות על קורלציה בין אי ציון ערך "רוחב" לאי קיום גידול ובפרט יעידו על אי ביצוע הבדיקה המאבחנת בלבד. מכאן, משקול הערכים ואופן הצגתם בדאטה הינה משמעותית שהרי הדבר עלול להוביל להטיה.

תהליך ה- Preprocessing:

נתייחס כעת לתהליך ה-Preprocessing שביצענו. למעט עמודת ה-Age, ביצענו פעולות עריכה והתאמה במרבית העמודות. עם זאת, מספר פיצ'רים אשר התגלו כרלוונטיים פחות לניתוח בעקבות התייחסות למספר התצפיות הקיימות וכן לאור משמעותם, הוסרו.

כמו כן, הפיצ'רים בהם בוצעו שינויים הינם:

1. פיצ'ר Basic stage: חילקנו לשלוש דרגות בהתאם למצבים השונים (ערכים 1-3), כאשר null=0.
2. פיצ'ר Her2: ערכים 1-, 0, 1 כאשר אלו מתייחסים לקיום הגן Her2 שמקושר להתפתחות סרטן שד (זהו גן שנבדק ע"י 2 מבחנים: IHC ו-FISH). באותו אופן, עבור הפיצ'ר Lymphatic penetration.
3. פיצ'ר Histological diagnosis: ערכים 1-, 0, 1 כאשר אלו מושפעים מהאם האבחנה מכילה מילים המקושרות לגידול שפיר/ממאיר (ערכים 1-, 1 בהתאמה, 0 עבור מילים ניטרליות יותר).
4. פיצ'ר Histopathological degree: המרה לדירוגים מספריים, בהתאם לרמות G השונות (ערכים 1-3 בהתאמה, ערך 0 מתייחס לחוסר יכולת לסווג, ערך 1- ל-null).
5. פיצ'ר KI67 protein: המרת ביטויי אחוזים לערכים בין 0-10.

6. פיצ׳רי TMN : המרה לדירוגים מספריים.
7. פיצ׳ר Margin Type : מיפוי הערכים באופן הבא : 0 : ׳ללא׳, ׳נקיים׳ : 1, ׳נגועים׳ : -1.
8. פיצ׳רים אבחנה-Nodes exam ואבחנה-Positive nodes : החלפת ערכי NA עם ערך חציון.
9. פיצ׳ר אבחנה-Side : המרה לשתי עמודות המייצגות right left, אינדיקטיביות (ערכי 0,1).
10. פיצ׳ר אבחנה-Stage : המרה לדירוגים בהתאם לחומרת הStage. טווח ערכים 0-12.
11. פיצ׳ר אבחנה-Surgery sum : החלפת NA ב-0.
12. פיצ׳ר אבחנה-Tumor width : הורדה של עמודה זו, והוספת עמודות שמהוות אינדיקטורים לכך שרוחב הגידול בטווחים שונים.
13. פיצ׳רים אבחנה-er ואבחנה-pr : המרה לערכים חיוביים/שליליים.

<u>התייחסות לשיטת הלמידה ותהליך בחירתה :</u>

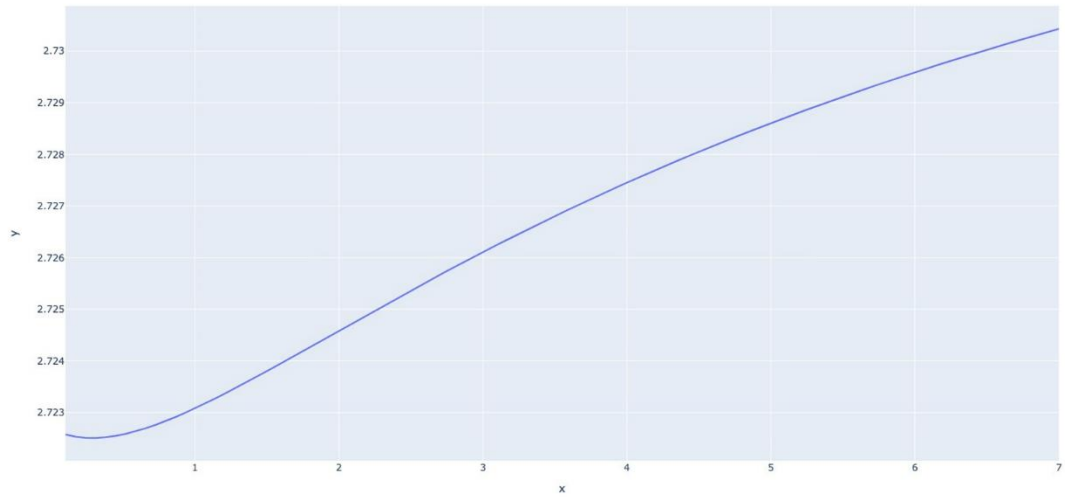
במשימה הראשונה - עקב חשיבות הקשר בין אזורי התפתחות הגרורות, וכן ההסתברות למציאתם יחדיו, חיפשנו מודל שמבצע סיווג לקטגוריות מרובות - במקום ביצוע חיזוי לכל אזור בנפרד. לצורך כך השתמשנו במודלים לביצוע Multioutput Classification, שמאפשר חיזוי משולב לאזורים השונים. בנוסף, בחרנו להשתמש ב-Random Forest שמהווה שיטת Ensembel - זאת לאחר התנסות עם מודלי בסיס שונים : כמו KNN, SVM ,Radius neighbours. מצאנו כי הציונים הגבוהים ביותר התקבלו עבור מודל ה- Random Forest.

במשימה השניה - כיוון שזו הייתה משימת רגרסיה במהותה, ניסינו מספר מודלים : Ridge Resression, Lasso Regression, Linear Regression. בנוסף, ביצענו cross-validation על ערכי הרגולריזציה שונים (כ-100 ערכים במרווחים שווים בין 0.1-7) לצורך כיול המודל – נמצא כי פרמטר הרגולריזציה שנבחר שעבורו התקבלו שגיאות הולידציה הנמוכות ביותר הוא 0.3.

<u>שגיאת ההכללה :</u>

בגרף שלמטה ערכי ה-x מייצגים את ערכי הרגולריזציה, וערכי ה-y מייצגים את ממוצע השגיאות ב- cross-validation. ניתן לראות על פי הגרף כי ערך הרגולריזציה שבחרנו (0.3) הינו בעל שגיאת הולידציה הנמוכה ביותר.

עם זאת, ניתן לראות כי סטיית התקן של ערכי ממוצע אלו הינה : 104.82248957798492, כלומר מדובר בסטייה גבוהה. לכן, קשה לומר מה תהיה שגיאת ההכללה, אך נעריך שתהיה בסביבות שגיאת הוולידציה הממוצעת שהתקבלה עבור ערך הרגולריזציה שנבחר (סביבות 2.7).


2.73
2.729
2.728
2.727
2.726
2.725
2.724
2.723
y
1
2
3
4
5
6
7
x