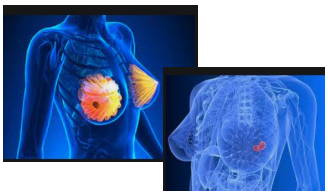


IML Project

האקתון מערכות לומדות - 2022

- משימה 2 -

Detecting Attributes of Breast Cancer

משתתפים : יאיר שטרן, שיר לוי, הילה היימברג

1 ביוני 2022

תיאור הדאטה:

קיבלנו סט אימון שמכיל 65,798 דגימות כאשר כל אחת עם 34 פיצ'רים. מידע זה מסופק ע"י קובץ train שמכיל 49,351 רשומות וע"י קובץ test שמכיל 16,447 רשומות. מדובר על מידע רפואי של מטופלות בסרטן השד כמו מאפיין מזהה של מטופלות, גיל, מספר הניתוחים וכו'.
מאפיינים בעייתיים : התמודדנו עם דאטה שדרש נורמליזציה. מדובר בדאטה שרופאים הזינו ומכילים גם שגיאות כתיב- שילוב של אנגלית, עברית ומספרים, כפילות של אותיות וערכים לא הגיוניים תחבירית.

<u>חלק 1</u>

המשימה:

לזהות מאפיינים של סרטן שד אצל נשים בהינתן מידע רפואי עליהן.
עלינו לחזות התפשטות של גידולים בגוף, כלומר **לחזות לאיזה אזורים הסרטן עלול להתפשט**.

תהליך עיבוד הדאטה:

עברנו על הפיצ'רים תוך בדיקה אילו יהיו רלוונטיים למודל שלנו עבור המטרה לשמה הוא פועל ואילו לא.
את הפיצ'רים הלא רלוונטיים הוצאנו מהדאטה ע"י כך שהרכבנו את הדאטה רק ע"י הפיצ'רים הרלוונטיים.

דוגמאות לפיצ'רים לא רלוונטיים בעיניינו:
-User Name- מספר משתמש של הרופא. איננו חושבים שיש קשר בין שם המשתמש של הרופא לבין השאלה "לאן הסרטן עלול להתפשט בגוף".
-Hospital- מספר המייצג את שם ביה"ח. גם כאן איננו רואים קשר בין ביה"ח לבין השאלה המבוקשת.
-Surgery date1, Surgery date2, Surgery date3- תאריכי ניתוחים לא יתנו לנו מידע על מיקום הגידול המתפשט בגוף ולכן בחרנו לא להכניס אותם.

דוגמאות לפיצ'רים חשובים שהשארנו:
- אבחנה-Side- הצד בו נמצא הגידול. מתבקש שפיצ'ר זה יהיה רלוונטי מאוד למודל כדי לעזור לו לחזות באיזה אזור בגוף יש התפשטות סרטנית.
- אבחנה-Histological diagnosis- אבחנה של הרקמות. הנחנו שבבדיקת רקמות נוכל לזהות אזורים בגוף שנגועים בסרטן.
- אבחנה-Surgery name1- אבחנה-Surgery name2- אבחנה-Surgery name3- שם הניתוח ככה"נ יתן לנו מידע על האזור בגוף שבו הניתוח מתבצע וכך המודל ידע לחזות יותר טוב להיכן הגידול הסרטני התפשט.

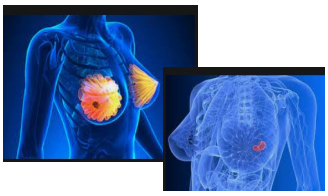
פיצ'רים קטגוריאליים: בחרנו את הפיצ'ר SIDE מאחר והוא קטגוריאלי והשתמשנו בGetDummies.

נקודות מעניינות בעיבוד המידע:
הפיצ'רים שאחראיים על עומק ורוחב הגידול (אבחנה-Tumor width, אבחנה-Tumor depth) כביכול נשמעו לנו רלוונטיים לחיזוי המיקום בגוף אך מכיוון שממעבר על הדאטה הבחנו שאין כמעט מידע עליהם- נאלצנו להוריד אותם.

בחירת המודל, בנייתו והדילמות שהתמודדנו איתן:

הבנו שצריך להתמודד עם בעיית קלסיפיקציה ובחרנו להשתמש ב**RandomForestClassifier**.

האתגר שלנו היה שצריך לחזות תתי קבוצות של לייבלים (אזורים בגוף). לתת למודל את הלייבלים כמו שהם עלול לגרום למודל לטעות ולא להבחין שמדובר בקבוצה של לייבלים. לכן הפיתרון שלנו היה לחלק את עמודת הלייבלים ל11 עמודות (כמספר האזורים בגוף). כל עמודה מייצגת קלאס מסויים. אימנו את



IML Project

המודל כל פעם עם עמודת לייבלים שונה כבעיית קלסיפיקציה בינארית רגילה. ביצענו פרדיקציה לכל קלאס ולבסוף איחדנו לעמודה אחת של תתי קבוצות.

חלק 2

המשימה:

לזהות מאפיינים של סרטן שד אצל נשים בהינתן מידע רפואי עליהן-המטרה היא **לחזות את גודל הגידול**.

תהליך עיבוד הדאטה:

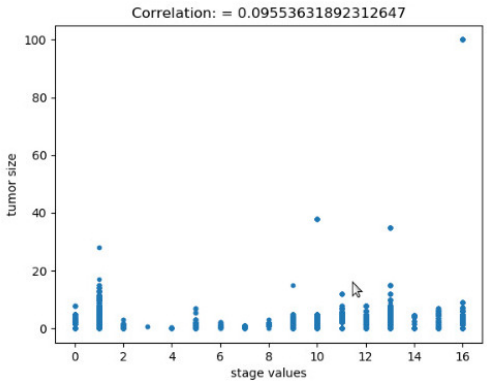
עברנו על הפיצ'רים תוך בדיקה אילו יהיו רלוונטיים למודל שלנו עבור המטרה לשמה הוא פועל ואילו לא.
את הפיצ'רים הלא רלוונטיים הוצאנו מהדאטה ע"י כך שהרכבנו את הדאטה רק ע"י הפיצ'רים הרלוונטים.

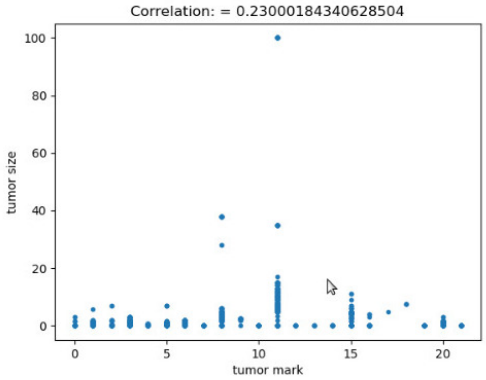
דוגמאות לפיצ'רים לא רלוונטיים בעיניינו:
- אבחנה-Side- בשונה מהמשימה של החלק הראשון, כאן הנחנו שהמידע על הצד בו נמצא הגידול הסרטני אינו יכול לתרום לנו מידע רב לגבי גודל הגידול ולכן לא הכנסנו את הפיצ'ר הזה.
- אבחנה-Lymphovascular invasion -Ivi- פיצ'ר זה אחראי לומר האם הגידול פלש לכלי דם או לבלוטות הלימפה ואיננו חושבים שמידע זה תורם לגודל הגידול הסרטני.

דוגמאות לפיצ'רים חשובים שהשארנו:
- אבחנה-Surgery sum- מספר הניתוחים עשוי להעיד על גודל הגידול הסרטני בגוף שכן דרוש מספר ניתוחים רב יותר כדי להתגבר עליו.
- אבחנה-Stage- השלב בו נמצא הסרטן. ככל שהשלב מתקדם יותר כך סביר שגודל הגידול יהיה גדול יותר ולכן החלטנו שפיצ'ר זה רלוונטי למודל שלנו.
- אבחנה-(TNM) Tumor mark- T – גודל הגידול ההתחלתי. מאחר והמודל אמור לחזות את גודל הגידול, הפיצ'ר הזה רלוונטי שכן הוא נותן לנו את הגודל ההתחלתי שלו.
- אבחנה-(TNM) metastases mark- M- כמות הגרורות הקיימות. ככל שיש יותר גרורות סביר להניח שהגודל יהיה גדול יותר.

נקודות מעניינות בעיבוד המידע:
הפיצ'רים שאחראיים על עומק ורוחב הגידול (אבחנה-Tumor width, אבחנה-Tumor depth) כביכול נשמעו לנו רלוונטיים לחיזוי גודל הגידול אך מכיוון שממעבר על הדאטה הבחנו שאין כמעט מידע עליהם-נאלצנו להוריד אותם.

נקודה מעניינת לגבי הקורלציה:





לאחר שחקרנו על נושא הגידולים הסרטניים בשד, ראינו שיש קשר בין ערך הSTAGE לבין גודל הגידול הסרטני. עם זאת, כאשר בדקנו את ערך הקורלציה ביניהם גילינו שאין קשר חזק כל כך כמו שהיינו מצפים. אותו הדבר לגבי הפיצ'ר של TUMOR MARK האחראי על גודל הגידול הראשוני.

בחירת המודל והשיקולים לבחירתו:

הבנו שמדובר בבעיית רגרסיה. ניסינו להשתמש בKNN ובעצים ולאחר שהשווינו בין המודלים ראינו ש RandomForestRegressor נותן את הביצועים הטובים ביותר ולכן בחרנו בו.