שיטות במדעי הנתונים 2022

ד"ר יהודית סומך - פרויקט סופי

מורן תותרי 209518018 חאתם ח'אתר 209002799

שלב ראשון - עיבוד נתונים

האיבר שבחרנו – ריאות

קצת על הריאות ! אצל האדם תופסות הריאות חלק גדול מבית החזה, מעצם הבריח עד הסרעפת. לאדם בריא יש שתי ריאות, ימנית ושמאלית. הריאה השמאלית מחולקת לשני חלקים הנקראים אונות, והיא מעט קטנה מהריאה הימנית כדי להותיר מרחב להכלת הלב. הריאה הימנית מורכבת משלוש אונות. האונות מתחלקות לאוניות, שכל אחת מהן מהווה יחידה עצמאית מבחינת הספקת הדם. לפיכך אפשר, במקרה של מחלות ריאות מסוימות, לכרות אונה או אונית ושאר חלקי הריאה ימשיכו לתפקד ללא בעיה.

תחילת הניתוח

אחרי שקראנו את הנתונים והתאמנו בין הטבלאות , חילצנו חלק מהטבלה שקשור לאיבר שבחרנו . ככה זה נראה

```{r}
tmp.tissue = ts.list[36]
print(paste0("loading ", tmp.tissue, " edata"))
```

```
[1] "loading Lung edata"
```

לאחר מכן , מחקנו את הגנים עם ערכים קטנים מדי דרך החישוב הזה

x > log(2 ,0.1+1)

ואז מסירים את כל הגנים עם השונות = 0 כדי להקטין את המטריצה .

✓ **זיהוי ומחיקת חריגים - outliers**

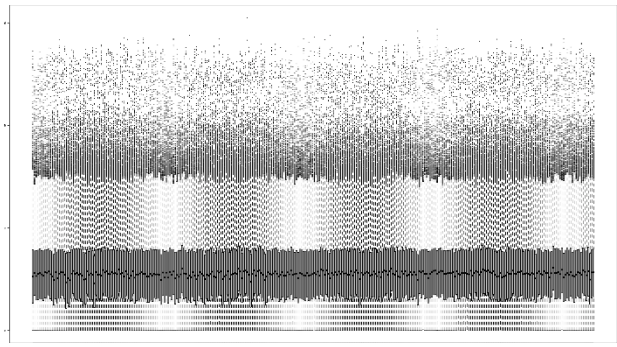


Figure 1
הצגת הנתונים לזיהוי outliers

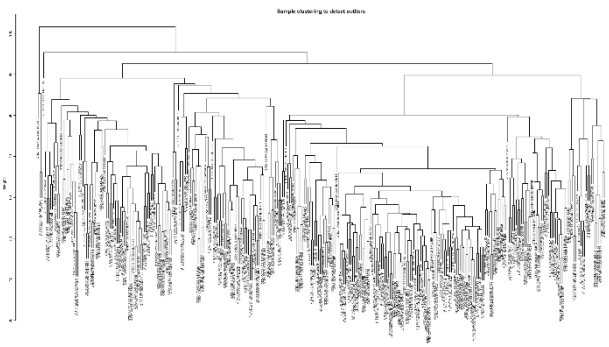


Figure 2
זיהוי אשכולות של outliers

למה חשוב להסיר חריגים (outliers) ?

חריגים מגדילים את השונות בנתונים שלך, מה שמקטין את הכוח הסטטיסטי. כתוצאה מכך, אי הכללת חריגים עלולה לגרום לתוצאות שלך להיות מדויקות יותר סטטיסטית.

✓ נרמול נתונים

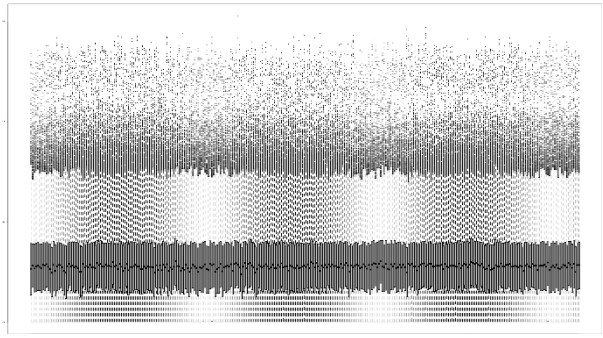


Figure 3
נתונים לפני נרמול

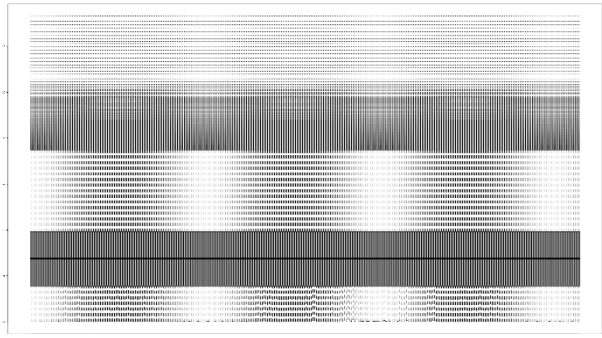


Figure 4
נתונים אחרי נרמול

✓ ויזואליזציות להבנת הדאטה וקשר בין תכונות

בחרנו להציג כמה plots שיעזרו לנו להסתכל על הנתונים מלמעלה ולהבין את התמונה הכוללת לפני שאנחנו מתעמקים בניתוח (זה עזר לנו להבין מתי אנחנו מקבלים תוצאות שהן באמת רלוונטיות לנתונים והתפלגותם ומתי אנחנו מקבלים תוצאה שהיא לא תואמת לנתונים ובכך ידענו שיש בעיה שאנחנו צריכים לתקן) .

בחרנו להסתכל על

- התפלגות הגילאים

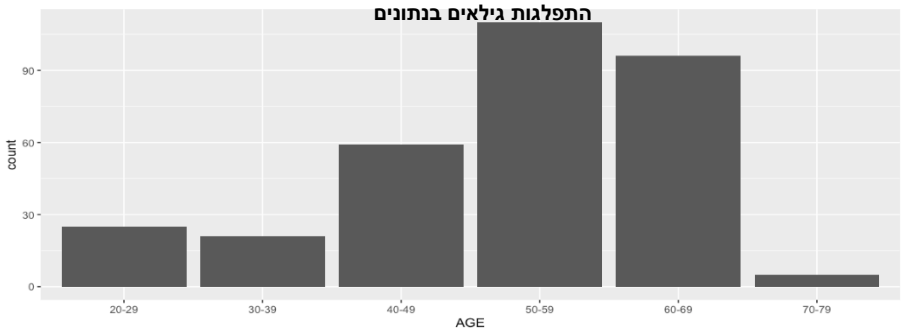
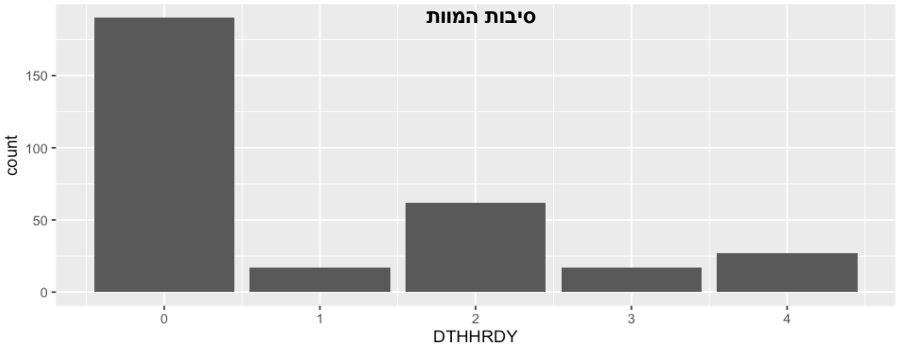


Figure 5
התפלגות גילאים בנתונים

- התפלגות סיבות המוות



0 – ventilator cases

1 – violent death

2 – fast death

3 – intermediate death

4 – slow death

Figure 6
התפלגות סיבת מוות בנתונים

- התפלגות גילאים לכל סיבת מוות

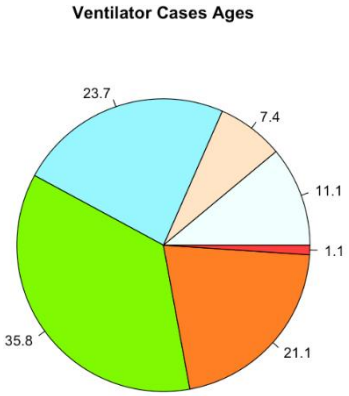
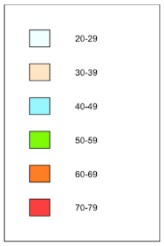


Figure 7



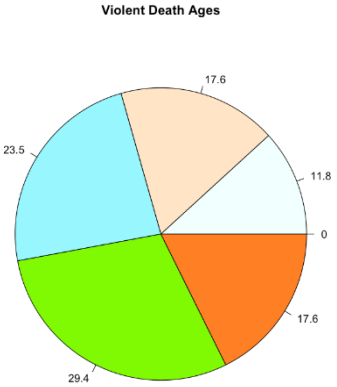


Figure 8

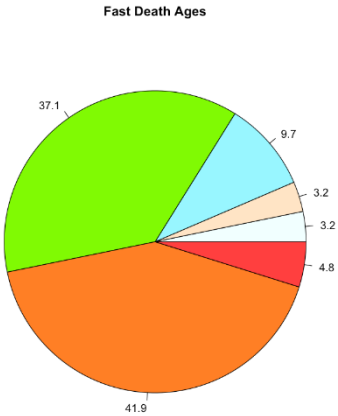


Figure 9

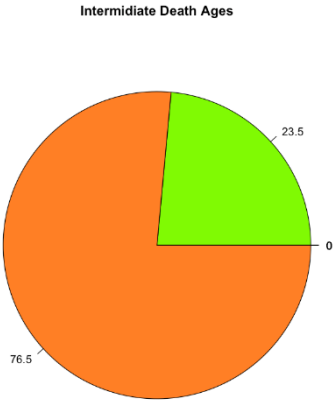


Figure 10

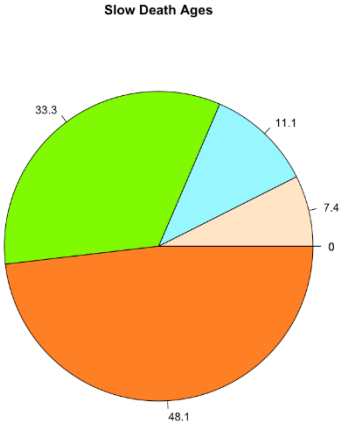
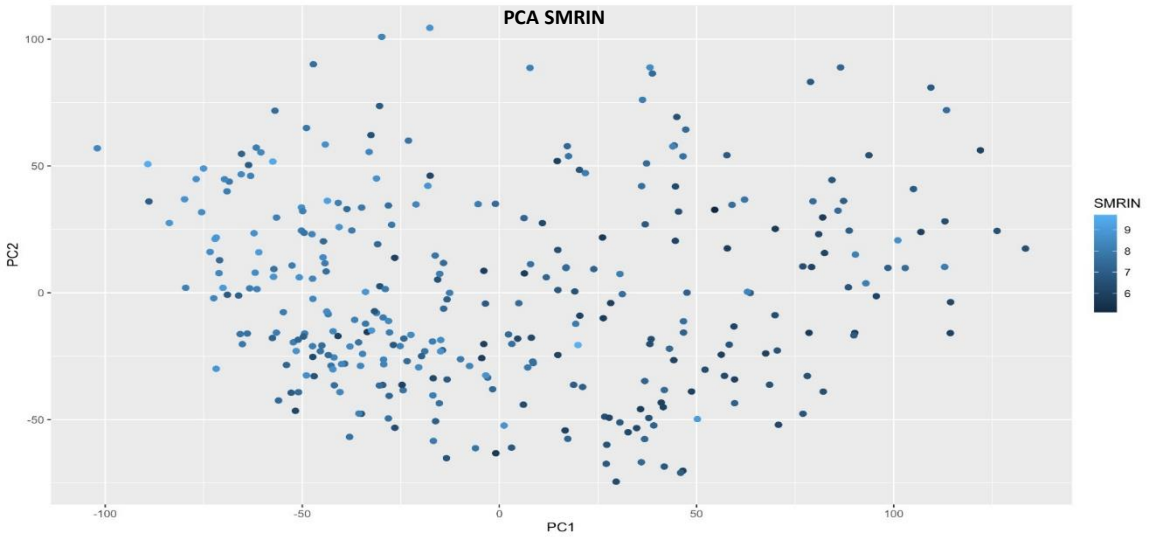


Figure 11

שלב שני – זיהוי תכונות שמוסיפות "רעשים" לנתונים

- **PCA לכל חמשת הרעשים**

1. SMRIN



אנחנו יכולים לראות שיש השפעה של פיצ'ר זה על הנתונים בצד שמאל של הגרף נמצאים דגימות יותר בהירות (SMRIN גבוה) .

2. SMGEBTCH

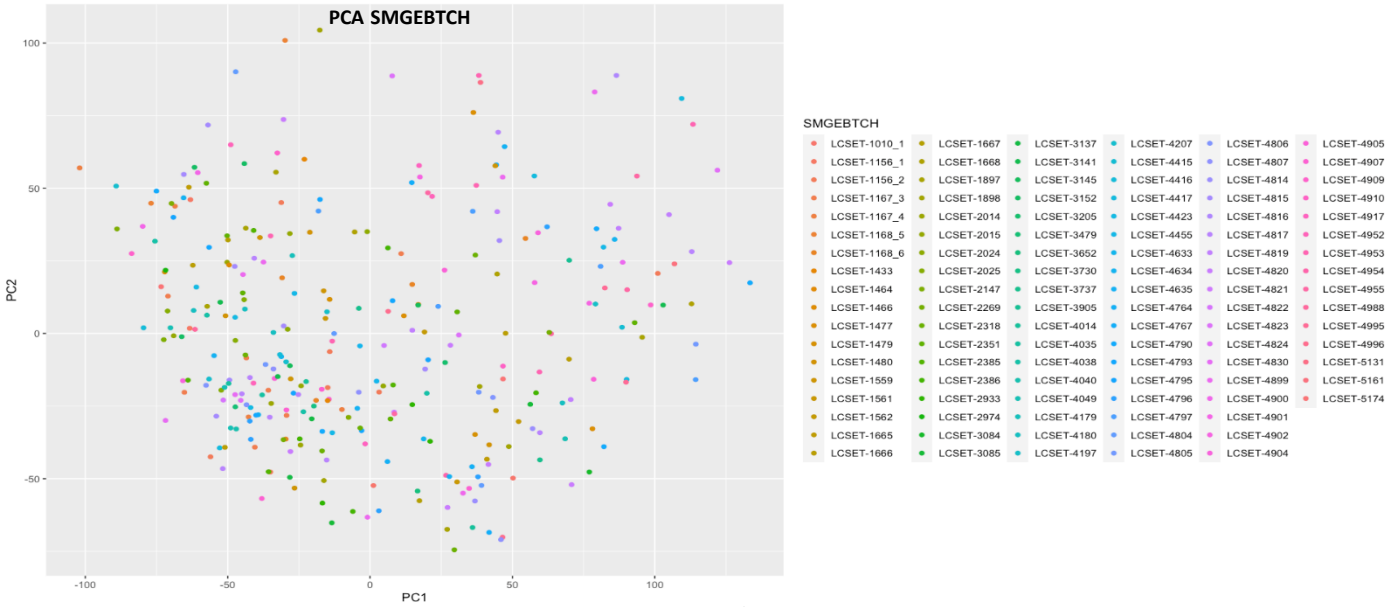


Figure 14

במקרה זה ה PCA היה לא כל כך יעיל מסיבה פשוטה שיש כל כך הרבה BATCH ID שקשה לקבץ את הנתונים לפיהם ולהציגם בצורה שנראית טוב לעין ומובנת. (מצריך הרבה מגוון של צבעים)

3. DTHHRDY

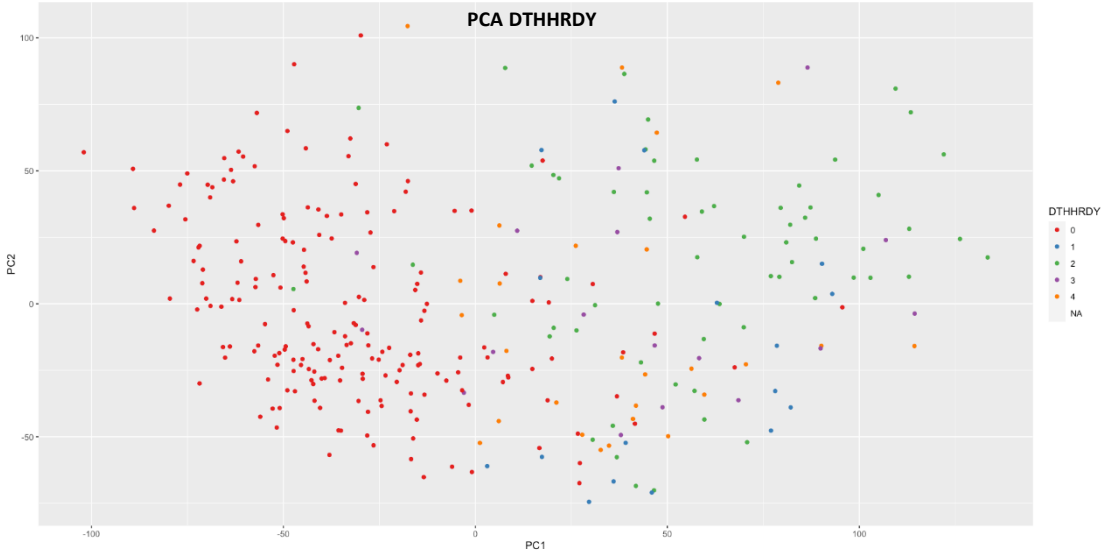


Figure 15

אנחנו יכולים לראות בבירור את הצבע האדום שמסמל את הסיבה שאנשים מתים **ממקרי הנשמה** מתאחד בצידו השמאלי של התרשים.

4. AGE

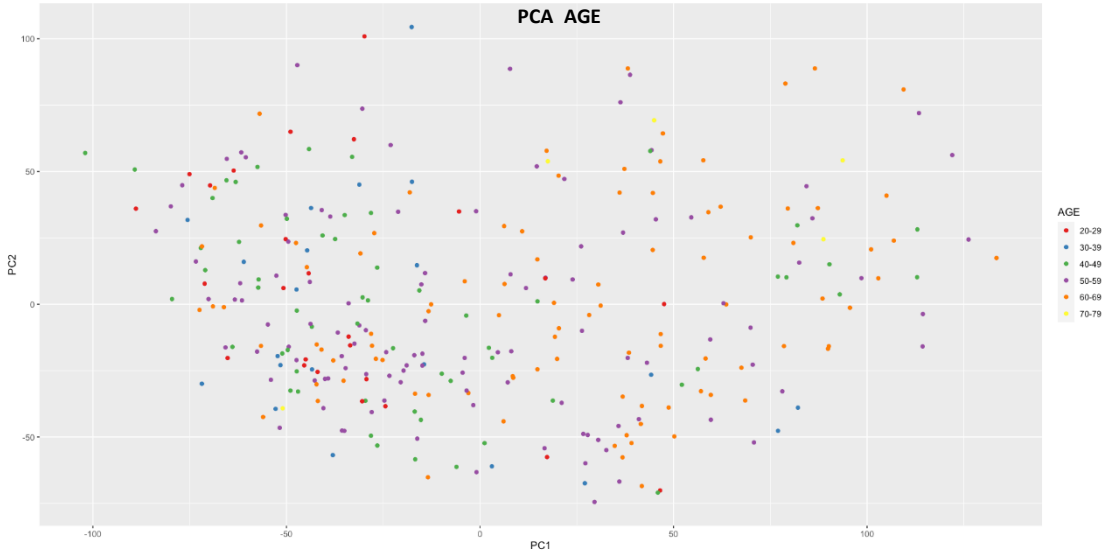
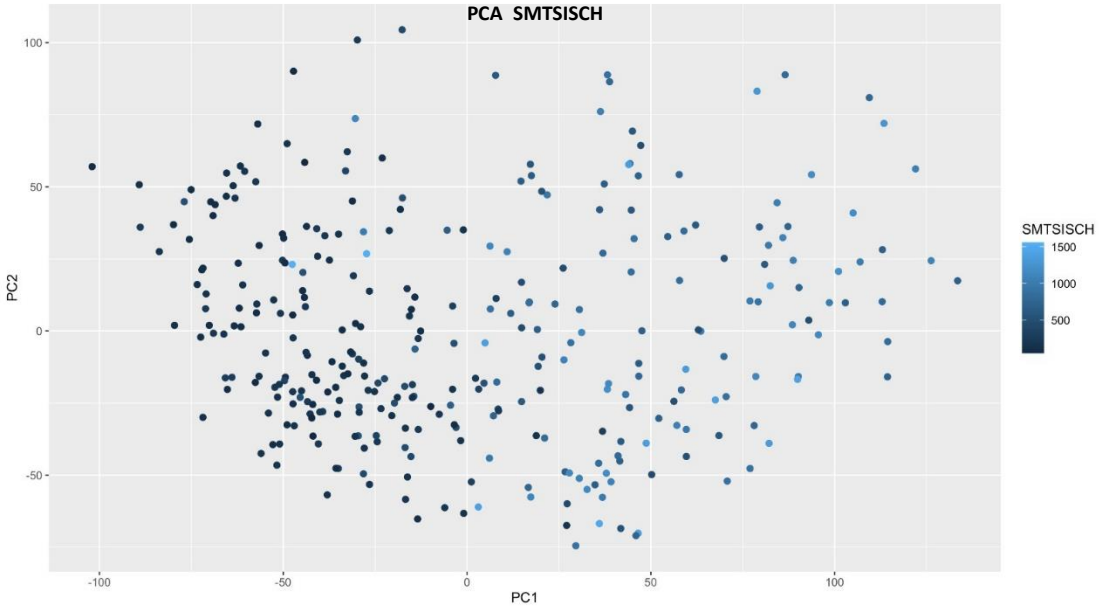


Figure 16

אין pattern או התנהגות ממסוימת מעניינת בהתפלגות הגילאים .

SMTSISCH .5



זה הפיצ'ר שעד כה ניתן לחלק דרכו את הנתונים/דגימות לשני חלקים מאוד ברורים ! מעיד על רעש מאוד בולט (נבדוק ונדייק בהמשך דרך מודלים ML)

- שיטות ML והעמקת ניתוח

בחלק זה התאמנו את האלגוריתם לסוג השדה שאנחנו בודקים (קטגוריאלי או נומרי) , ובחרנו להשתמש בשני אלגוריתמים ובנינו 4 מודלים ל 4 רעשים שונים לצורך ניבוי :

פיצלנו את הנתונים ל 20% test ו 80% train (השתמשנו ב 1PCA - 20PCA)

- Random Forest – בחרנו בו כי הוא נחשב אלגוריתם סובלני לשדות שהם קטגוריאליים . ולאחר פרדיקציה הצגנו את ה ACTUAL LABEL ואת ה PREDICTED LABEL ב CONFUSION MATRIX ומדדנו את איכות המודל ודיוקו לפי ה ACCURACY .

***לתכונת AGE**

Confusion Matrix to Age Prediction

Predicted Label

Actual Label	20-29	30-39	40-49	50-59	60-69	70-79
20-29	0	0	0	4	0	0
30-39	0	0	0	4	0	0
40-49	0	0	0	9	1	0
50-59	0	0	0	17	3	0
60-69	0	0	0	18	5	0
70-79	0	0	0	3	0	0

Accuracy : 0.3438

מ Figure 5 (התפלגות הגילאים בנתונים) אנחנו יכולים לראות שמספר הסמפלים בגילאים 50-69 הוא גבוה ביחס לשאר הטווחים ,הסטייה בתוצאות הניבוי בדיוק בטווחים אלו שיקפה את הפער בכמות הסמפלים .

***לתכונת DTHHRDY**

Confusion Matrix to DTHHRDY Prediction

Predicted Label

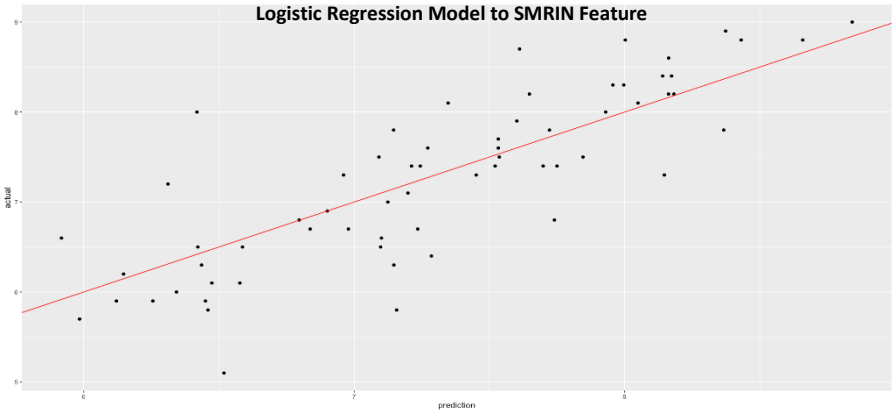
Actual Label	0	1	2	3	4
0	32	0	0	0	0
1	1	0	0	0	0
2	12	0	4	0	0
3	7	0	0	0	0
4	8	0	0	0	0

Accuracy : 0.5625

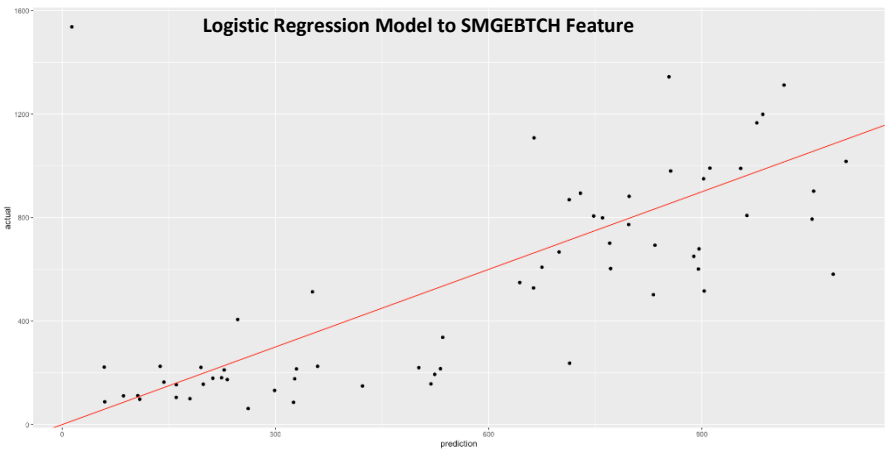
במקרה הזה ה ACCURACY שקיבלנו הרבה יותר גבוה וזה מעיד על יותר דיוק בניבוי אבל שוב אנחנו באותה בעיית התפלגות סמפלים שניתן לראות ב Figure 6 שהקלאס 0 יש לו פער בכמות הסמפלים .

- Logistic Regression – בחרנו במודל זה כדי לטפל בבעיית קלסיפיקציה לערכים נומריים .

***לתכונת SMRIN**



***לתכונת SMGEBTCH**



הערכים שנמצאים על קו ה hyper plane האדום הם הערכים שסווגו באופן נכון . כדי להשוות בין שני הביצועים של המודלים ודיוקם הצלחנו למדוד את מדד ה ERROR לשתי התכונות.

נגדיר מה זה ERROR ? מרחק כל נקודה מה HYPER PLANE . אם קיימת נקודה (X,Y) אז ה ERROR שלה הוא המרחק שלה מהנקודה (Y,Y) – ה Y הוא ה ACTUAL .

בגלל שהערכים בשני הפיצ'רים נמצאים בטווחים שונים לחלוטין , בהתחלה לא הצלחנו להשוות בין ערכי ה **ERROR הממוצע** של שני המודלים , לכן עשינו עוד שלב , והוא שלב הנרמול (לאותו טווח) .

וקיבלנו את התוצאות הללו :

*למודל **SMTSISCH** קיבלנו **MSE** (minimum squared error) 0.2338556

* למודל **SMRIN** קיבלנו **MSE** (minimum squared error) 0.09029161

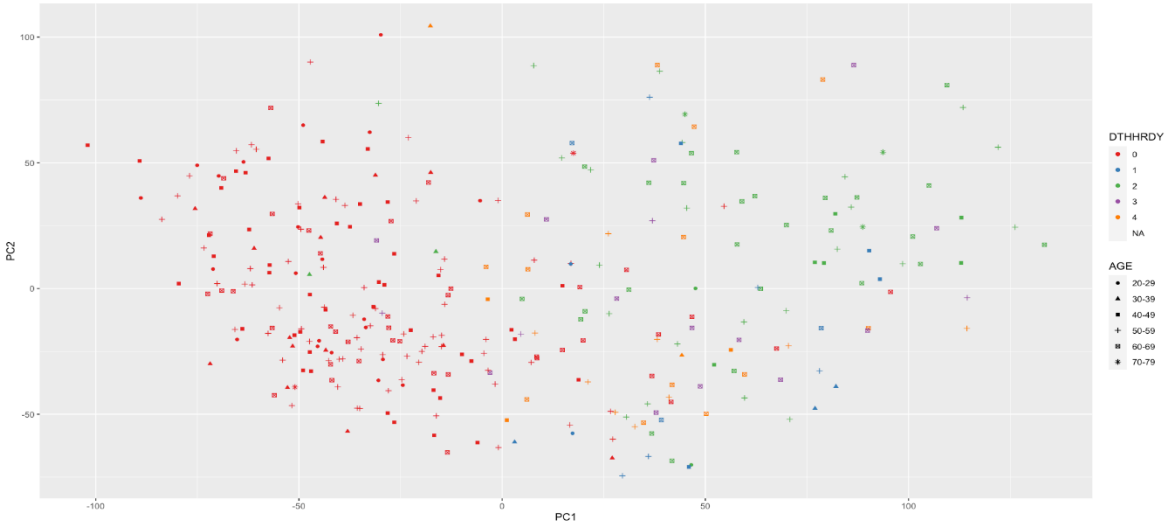
ולכן המודל ש"ניצח" בדיוק שיש לו ערך MSE יותר קטן והוא SMRIN .

<u>מסקנות מהניתוח וסיכום תוצאות :</u>

- SMRIN כמו שראינו במודל ה PCA הצלחנו לראות התפלגות נתונים ברורה לפי פיצ'ר זה ובתוצאות ה Logistic regression הוא היה יותר מדויק בניבוי לכן אנחנו מחשיבים אותו **כרעש בולט** . מבחינה לוגית זה הסתדר לנו עם ההשערות הראשוניות וההתרשמות הראשונית בנתונים .
- DTHHRDY כמו שראינו במודל ה PCA הצלחנו לראות התפלגות נתונים ברורה לפי פיצ'ר זה ובתוצאות ה Random Forest הוא היה יותר מדויק בניבוי לכן הוא גם נחשב **כרעש בולט**.
- SMTSISCH כמו שראינו במודל ה PCA הצלחנו לראות התפלגות נתונים ברורה לפי פיצ'ר זה , גם מבחינה לוגית זה מסתדר לנו שהזמן כן ישנה תוצאות של דגימות (למשל אנשים שכבר מאושפזים ונפתרו סביר להניח שהדגימה שלהן תילקח באופן מידי מאשר דגימה ממישהו שנפתר בתאונת דרכים) לכן נחשיב אותו גם **כרעש בולט** .

נספחים

- **DTHHRDY and AGE in PCA**



אנחנו יכולים לראות בבירור את הצבע האדום שמסמל את הסיבה שאנשים מתים ממקרי הנשמה מתאחד בצידו השמאלי של התרשים וניסינו לראות אם יש השפעה של הגיל על מקרי המוות אבל אין התפלגות מיוחדת (אין השפעה) .

- **אינטגרציה בין גיל ל איכות ה RNA**

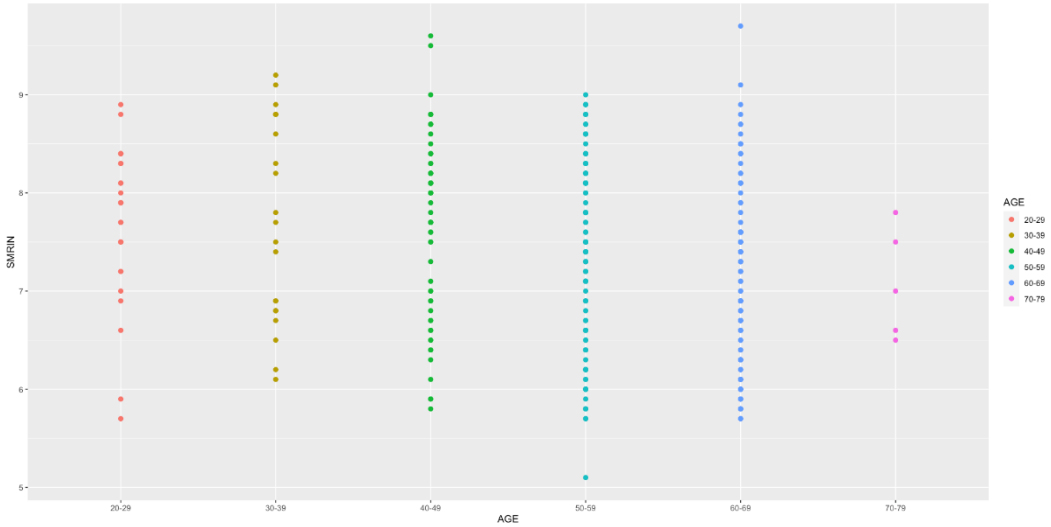


Figure 12