ניתוח נתונים מתקדם ב- R: פרויקט גמר

קישור לקוד של הפרויקט כולל הוראות ברורות בקובץ README.md נמצאים בקישור הבא: קוד הפרויקט.

מבוא:

הפרויקט יעסוק במחקר של ד"ר טליה שוורץ, העוסק בהתפתחות נערים בסיכון. המחקר נעשה לאורך שנים רבות בחייו של הנער והוא אחד המחקרים הראשונים והיחידים שנעשו בצורה זו, ובוחן את השינויים שעובר אדם בסיכון במהלך שלבים משמעותיים בחייו. השלבים כוללים את הפנימייה (גילאים 16-18), השירות הצבאי / לאומי (גילאים 20-21) ותקופת הבגרות הצעירה (גילאים 25-26). שלושת השאלונים בשלבים השונים בחיים נותנים לנו את היכולת לבחון את התקדמות הנער לאורך תקופה מאוד חשובה בחייו.

שאלת המחקר שלנו היא: **"האם קיימים אספקטים של רמת הפתיחות הרגשית והמוכנות לקבל עזרה חיצונית, שעבורם יש קשר להתקדמות האישית (ההישגיות) של הנער בסיכון בחייו הבוגרים?".**

היפותזה:

H_0: לא קיימים אספקטים של פתיחות רגשית של נער בצעירותו עבורם יש קשר להישגיות בחייו הבוגרים.

H_1: קיימים אספקטים של פתיחות רגשית של נער בצעירותו עבורם יש קשר להישגיות בחייו הבוגרים.

ילדים בסיכון הם אוכלוסייה מוחלשת ביחס לאוכלוסיית הילדים הכללית. בפרויקט ננסה להראות ששיפור האספקטים של רמת הפתיחות הרגשית של הנער ומוכנותו לקבל עזרה / תמיכה חיצונית, יגדילו את סיכוייו לחיים עתידיים טובים יותר בתחומים רבים. השאיפה שלנו הינה להראות שאכן קיים קשר כזה, ובעזרת ממצאים אלו להציע לפנימיות לפתח שיטות הנוגעות להמלצות אלה, אשר בתקווה ישפרו את מסלול חייו של הנער. בעזרת שיטות אלו נשאף לגשר על הפערים הקיימים בין אוכלוסייה זו לאוכלוסיית הנערים הכללית.

כדי להעמיק את הידע שלנו בנושא וכדי לחזק את שאלת המחקר שלנו, נעזרנו במקורות חיצוניים. רוב המחקרים בנושא סבבו סביב תיאוריית התפתחות הילד של אריק אריקסון, מומחה בפסיכו-אנליזה אשר טען כי לסביבה החברתית שבה ילדים גדלים, מתפתחים ולומדים יש השפעה רבה על מידת ההישגיות שלהם. החוקר ניק דאפל טוען במחקרו בשנת 2000, כי עצם המעבר של הילד לפנימייה גורם לו להיכנס למצב מנטלי של הישרדות רגשית יומיומית אשר נדרשים משאבים רבים כדי לפצות על כך אחר כך.

"in order to cope with their loss of family and to adapt to their school environment, children unconsciously construct a strategic survival personality and that such personality structure invariably becomes counter-productive in adult life". (p. 51 - International Journal of Education and Research Vol. 4 No. 6 June 2016).

שאלת המחקר שלנו מבוססת בעיקרה על הפתיחות הרגשית של הנער. הקושי לאבחן ולהגדיר תכונה זו הופך את ניתוח הנתונים למורכב אך מרתק. המחקר שהוזכר לעיל הוא יחיד במינו באופן שבו הוא מתפרש על שנים רבות. לפיכך, יסודיות הניתוח שלנו תהיה ייחודית בהשוואה לשאלות מחקר דומות אחרות שבוצעו בעבר, כמו גם משתנה הפתיחות הרגשית של הנער, אשר יהווה מדד שאין כמותו בסוג הנתונים הרלוונטיים. נושא זה הינו מאתגר למחקר משום שפנימיות וכפרי נוער משתנים מאוד לפי אזורים, אוכלוסיות ותרבויות שונות. לפיכך, שינוי והתאמת גישה לנערים נדרש באופן רציף (המידע על קושי זה מבוסס על מחקרו של Stout משנת 2012, אשר מתמקד בשוני בין פנימיות באזורים שונים ברחבי העולם).

הגישה שלנו בבדיקת ההיפותזה היא התמקדות במדד הפתיחות הרגשית והמוכנות לקבל עזרה חיצונית, אותם נגדיר באמצעות שאלות רלוונטיות עליהן ענו הנערים בשאלונים. בהתאם לתוצאות המחקר שלנו, נרצה להסיק האם מומלץ לבצע שינוי בתהליכי הפנימיות לשיפור הפתיחות הרגשית של הנערים ו/או עידודם לפנות לתמיכה מסביבתם.

נתונים:

הנתונים מכילים שאלות לנערים מכפרי נוער ופנימיות. הנתונים מחולקים ל- 3 שאלונים אשר יפורטו בהמשך:

בשאלון הראשון (גילאים 16-18) ישנן שאלות העוסקות ב: זיהוי הנער, בירור טיב היחסים בין הנער למשפחתו כמו גם עם צוות הפנימייה, בדיקת מיומנויות לחיים שהשיג בפנימייה, בחינת נכונותו לקבלת עזרה מאיש צוות מומחה, וסיכוייו העתידיים.

בשאלון השני (גילאים 20-21) ישנן שאלות העוסקות ב: מצבו הנוכחי של הנער, תכנונים לעתיד הקרוב מבחינת לימודים אקדמיים, תמיכת הסביבה החברתית, רצון לבקש עזרה מהממסד ממנו הגיע, וכישורי החיים החברתיים והכלליים כיום ובעתיד.

בשאלון השלישי (גילאים 26-25) ישנן שאלות העוסקות ב: מצב החיים הנוכחי ושביעות הרצון ממנו, טראומות עבר והטיפולים שנעשו כדי לטפל בהן, בדיקת חידוש הקשרים עם אנשי צוות כפר הנוער / הפנימייה והשפעת יחסים אלו על חייו, הערכה עצמית ותקוות ותכניות עתידיות.

המשתנים (השאלות) שבהם השתמשנו הם אלה הרלוונטיים לשאלת המחקר שלנו. בחרנו את השאלות הקשורות לרמת הפתיחות הרגשית והמוכנות של הנער לקבלת עזרה מכל שלושת השאלונים, ואת השאלות על הישגיו של הנער משאלון 3 בלבד. בסך הכל, השתמשנו ב- 48 משתנים כאשר 33 מהם משתנים מסבירים (משתני x) ו- 15 מהם משתנים תלויים (משתני y).

שיטות וממצאים:

לאחר שביצענו ניקיון, תיקון ונרמול של הנתונים, בחרנו את הפיצ'רים (המשתנים) הרלוונטיים המתוארים למעלה. המשתנה המוסבר הוא מדד שיצרנו להישגיות וייקרא משתנה y. יצרנו את מדד זה בעצמנו באמצעות התחשבות במספר תחומים בחייו הבוגרים של הנער (המורכבים מ- 15 המשתנים התלויים) וקביעת משקל לכל תחום. ערכו יהיה בטווח [0,1]. היות וכמות הפיצ'רים שאיתם עבדנו הייתה רבה, קיים סיכוי שחלקם לא תרמו לבדיקת ההשערות. לכן, החלטנו לבנות מודלים המתמקדים בצמצום פיצ'רים ובחירת הרלוונטיים מבינם. ביצענו מספר מודלים על הפיצ'רים הרלוונטיים כדי להגיע לתמונה בהירה יותר על הנתונים, ולהשיג ממצאים שיקדמו אותנו במציאת התשובה לשאלת המחקר.

ראשית, השתמשנו ב- Lasso Regression: שיטת ניתוח רגרסיה המבצעת בחירת פיצ'רים ורגולריזציה על מנת לשפר את דיוק הניבוי. המודל בחר 4 משתנים שמספקים את הקורלציה הגבוהה ביותר, אך ניתן לראות כי ה- R-Squared של המודל עבור משתנים אלה הוא בקירוב 0.1485 (מוצג בפלט כ- dev.ratio). נתון זה מראה על קורלציה נמוכה מאוד.

"term"	"step"	"estimate"	"lambda"	"dev.ratio"
"(Intercept)"	1	0.667553658766626	0.018966991306339	0.148511087819537
"T3_ACE_1"	1	0.00134061510161529	0.018966991306339	0.148511087819537
"T3_ACE_5"	1	0.0562364776884276	0.018966991306339	0.148511087819537
"PnimiaTrust"	1	0.00510386083156609	0.018966991306339	0.148511087819537
"B.SUPPORT_4"	1	0.0608664833821079	0.018966991306339	0.148511087819537

שנית, כדי להעמיק ולהבין בצורה ברורה יותר את בחירת הפיצ'רים של Lasso, ביצענו מודל של רגרסיה לינארית מרובת משתנים עם כלל הפיצ'רים (33 משתני x) החוזים את משתנה y. בתוצאות המודל התמקדנו בערך המובהקות (p-value) של כל פיצ'ר. המשתנים המסבירים שבחרנו הם המשתנים בעלי רמת מובהקות הקטנה מ- 5%, כלומר שהרגרסיה החזירה עבורם p-value הקטן מ- 0.05, ואכן קיבלנו כי 75% מהפיצ'רים שנבחרו על ידי ה- Lasso עמדו בתנאי זה. חמשת הפיצ'רים בעלי ערך המובהקות הנוקשה (הנמוך) ביותר שעמדו ב- threshold הנ"ל הם:

```
Coefficients:          Estimate Std. Error t value Pr(>|t|)
PnimiaTrust            0.089939   0.042520   2.115  0.03613 *
B.SUPPORT_4            0.102509   0.042827   2.394  0.01798 *
B.INTERNET.SUPPORT_3 -0.062527   0.030284  -2.065  0.04075 *
T3_ACE_5               0.097745   0.029937   3.265  0.00137 **
B.HELP.MENTAL          0.063799   0.029074   2.194  0.02981 *
```

גם במודל זה קיבלנו קורלציה נמוכה כאשר ערך ה- R-Squared הוא 0.3379 וה- Adjusted R-Squared הוא 0.1862.

```
Multiple R-squared:  0.3379,    Adjusted R-squared:  0.1862
```

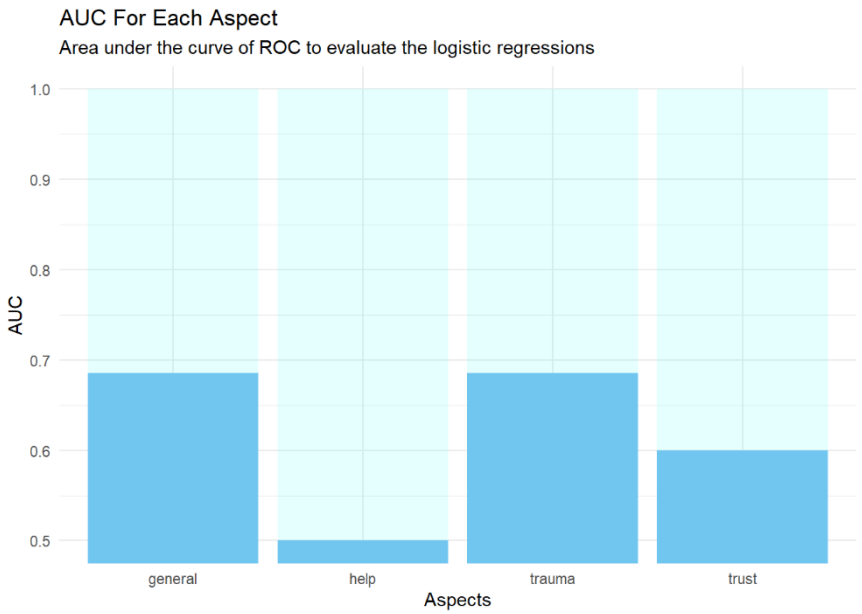
לאחר מכן, ביצענו רגרסיה לינארית מרובת משתנים המשתמשת אך ורק ב- 6 המדדים האיכותיים ביותר עבורנו, שהם איחוד הפיצ'רים שהחזיר Lasso עם הפיצ'רים שעמדו ב- threshold שקבענו ברגרסיה הקודמת. במודל רגרסיה זה קיבלנו שוב Adjusted R-Squared נמוך, אם כי יותר גדול מהמודל המכיל את כלל משתני x.

```
Multiple R-squared:  0.2534,    Adjusted R-squared:  0.2317
```

מודל נוסף שביצענו הוא Logistic Regression - מודל סטטיסטי לקלסיפיקציה (ניבוי של משתנה y וסיווג למחלקות). הגדרנו את המשתנה התלוי להיות בינארי: הישגי / לא הישגי. נער הישגי הוגדר כנער שמדד ההישגיות שלו גדול מ- 0.7. בחרנו את ערך זה מתוך שיקולים סטטיסטיים של ערכי עמודת ההישגיות וסף שקבענו על דעת עצמנו להגדרת נער הישגי / לא הישגי באופן בינארי. במודל זה השתמשנו שוב באיחוד הפיצ'רים שהחזיר Lasso עם הפיצ'רים שעמדו ב- threshold שקבענו ברגרסיה מרובת המשתנים.

את המשתנים הללו חילקנו ל- 3 זוגות של פיצ'רים המגדירים כל אחד אספקט מחייו של הנער:

- עזרה: 2 שאלות הנוגעות בנכונותו של הנער לבקש עזרה ולחפש פתרונות, ומוכנותו להתייעץ בהתמודדות עם קשיים רגשיים (הפיצ'רים B.HELP.MENTAL + B.INTERNET.SUPPORT_3).
- טראומות: 2 שאלות הנוגעות באירועים משפחתיים קשים אותם חווה / לא חווה הנער בצעירותו (הפיצ'רים T3_ACE_1 + T3_ACE_5).
- אמון: 2 שאלות העוסקות במידת הסמך של הנער על סביבתו בנוגע לבקשת עזרה כאשר הוא זקוק לה (הפיצ'רים PnimiaTrust + B.SUPPORT_4).



על כל אספקט (זוג) ביצענו Logistic Regression לחיזוי הישגיות או חוסר הישגיות של הנערים. באמצעות מדד ה- AUC (השטח מתחת לעקומת ROC), נוכל לקבוע את איכות הרגרסיה עבור כל אספקט. בנוסף, ביצענו את המודל גם על ששת הפיצ'רים יחדיו (general) כדי לבדוק האם הקומבינציה שלהם מניבה תוצאות מובהקות יותר.

קיבלנו כי ה- AUC המקסימלי מתקבל עבור אספקט הטראומות ועבור קומבינציה של כל הפיצ'רים יחד (general) וערכו שווה ל- 0.686, תוצאה נמוכה באופן יחסי. כלומר, כל ארבעת האספקטים לא הניבו רגרסיה לוגיסטית איכותית (המתקרבת לערך AUC שהוא 1) לחיזוי הישגיות.

לסיכום, כל המודלים שבנינו החזירו מדדים המעידים על קורלציה נמוכה ולא מספקת עבורנו. לכן, לא נוכל לטעון בצורה מובהקת כי ישנם אספקטים של פתיחות רגשית ומוכנות של נער לקבלת עזרה, עבורם יש קשר להישגיותו בחייו הבוגרים. למרות זאת, נוכל לציין שעשוי להיות קשר מסוים (אך לא חזק או מובהק) כפי שקיבלנו עבור אספקט הטראומות.

הגבלות ועבודה עתידית:

הגבלה מרכזית ראשונה שהיוותה גורם משמעותי בקושי לבחון את הדאטה, נובעת משיטת התאמה סטטיסטית שאינה מבטיחה דמיון מלא בין קבוצת המחקר לקבוצת ההשוואה. כלומר, יש לקחת בחשבון כי קיימת האפשרות שההבדלים בין הנערים נובעים ממשתנים שלא נכללו במחקר. משמע, משתנים שלא הופיעו עליהם שאלות, עשויים להשפיע על משתני התוצאה הנבחנים (הישגיות הנער). בעקבות כך, אנו ממליצים להמשיך ולפתח את מאגר הנתונים, לבצע עוד שאלונים לנערים בסיכון הכוללים יותר שאלות, וזאת בכדי להמשיך את קו הלמידה על הישגיות הנערים ועל מדד הפתיחות הרגשית שלהם.

קושי נוסף הוא הניסיון להפיק ממצאים מנתונים שסופקו לנו מגורם חיצוני ולא נאספו או בוצעו על ידינו. בהינתן הדאטה שקיבלנו, התבקשנו לנתח את הנתונים ולהסיק עליהם מסקנות מעניינות. אנו מאמינים שבמידה והיו עוד שאלות המתמקדות בנושא של פתיחות רגשית, היינו יכולים להגיע למסקנות מובהקות יותר, בין אם לקבל את היפותזת המחקר או בין אם לדחות אותה.

במידה והיו לנו עוד מספר חודשים של מחקר, היינו מתעמקים יותר במשמעות המודלים שאותם בנינו כדי להגיע למסקנות ובכך להיות בטוחים יותר בתוצאות שקיבלנו. בנוסף, היינו מתשאלים אנשים לגבי מה הם היו מגדירים כהישגיות. כלומר, אילו תחומים עבורם הם החשובים ביותר. מטרת פעולה זו היא בשביל הגדרת מדד ההישגיות בצורה כללית יותר, המשקפת דעות של כמות רחבה של אנשים, ופחות מסתמכת רק על דעותינו בעניין.

כמו כן, במידה והיה לנו יותר זמן לפרויקט, היינו מבצעים בדיקה מקדימה ומקיפה של הנושא לפני העלאת השערות. הייתה נפתחת בפנינו האפשרות לחפש יותר חומר / לקרוא מחקרים על נערים בסיכון ובכך להעמיק בתחום. היינו מתייעצים עם גורמים מקצועיים שעובדים או מתעסקים בתחום, ובעזרתם מנסים להגיע לרמת הבנה טובה יותר של הנושא כדי לדעת מה ואיך מומלץ לחקור.