פרויקט גמר בתואר B.Sc בהנדסת תעשייה וניהול, התמחות מדעי הנתונים, הפקולטה להנדסה

דוח מסכם - סמסטר ב'

הגיאוגרפיה של מחלות נשימה כרוניות

מנחה : ד"ר רוני ערמון

מגישים:

אלמוג סיסו 204307516

אלמוג בורה 206295115

פלג שובל 316485549

תוכן עניינים

הקדמה

מחקר זה עוסק במחלות נשימה כרוניות, בדגש על שתי מחלות עיקריות: אסתמה ו-COPD (מחלת ריאות חסימתית כרונית). תקופת המחקר מתמקדת בשנים 2013-2019.

שאלת המחקר המרכזית היא:

האם וכיצד ניתן לנבא את רמת התחלואה של מחלות נשימה כרוניות על בסיס משתנים חברתיים?

לצורך בחינת שאלה זו, ביצענו את הפעולות הבאות:

1. ארגון וסיווג הנתונים לקטגוריות על פי סוגי המשתנים.
2. המרת כלל המידע לתאי שטח אחידים.
3. יצירת פאנל נתונים מאוחד הכולל את כלל המשתנים.

באמצעות שימוש במודלים סטטיסטיים, החל מרגרסיה לינארית ועד למודלים מתקדמים יותר, אנו שואפים לחשוף תובנות משמעותיות על הקשר בין גורמים חברתיים לתחלואה במחלות נשימה כרוניות. מחקר זה עשוי לתרום להבנה מעמיקה יותר של גורמי הסיכון החברתיים ולשפר את יכולת הניבוי והמניעה של מחלות אלו.

שיטות

בפרק הנוכחי נעמיק בשיטות המחקר והניתוח שיושמו בעבודה זו. נבחן את התפלגות המשתנים, ונדון בשיטות חלופיות לניתוח הנתונים לאור ממצאינו. נציג טכניקות לטיפול בערכים חסרים וחריגים ונסקור את המודלים הנבחרים לחיזוי רמות התחלואה של מחלות נשימה כרוניות. נתאר את תהליך הערכת המודלים, כולל שיטות לבחירת היפר-פרמטרים ומטריקות להערכת ביצועים. לבסוף, נציג את ה- Pipeline שפותח למחקר זה, המאפשר ניהול יעיל של ניסויים מרובים. גישה זו מקיפה את כל שלבי המחקר, מבחירת תכונות רלוונטיות ועד לניתוח התוצאות, ומאפשרת גמישות רבה בתכנון וביצוע הניסויים, תוך שמירה על תיעוד מדויק ומקיף.

התפלגות המשתנים

במסגרת הניתוח שנערך, בוצעה בדיקת התפלגות נורמלית על כל המשתנים המופיעים במסד הנתונים. עבור כל אחד מהמשתנים, הוחל מבחן אנדרסון-דרלינג לבדיקת נורמליות התפלגותו, תוצאות הבדיקה הן כדלקמן:

לא מצאנו שום משתנה במסד הנתונים שמתפלג נורמלית באופן מובהק, כשרמת המובהקות נקבעה ל .P = 0.05 ממצא זה מצביע על הצורך בשימוש בשיטות סטטיסטיות חלופיות.

שיטות להשלמת ערכים חסרים

טיפול על ידי שימוש בשיטות השלמה שונות כגון השלמה על ידי חציון או ממוצע. להלן פירוט השיטות:

<u>השלמה על ידי ממוצע</u> - הערך החסר מוחלף בממוצע של כלל הערכים באותו משתנה. היתרון בשיטה זו הוא שהיא קלה ליישום ומקטינה את הסיכון להטיות משמעותיות. עם זאת, היא עלולה לטשטש את הפיזור הטבעי של הנתונים.

<u>השלמה על ידי חציון</u> - בשיטה זו, הערך החסר מוחלף בחציון של כלל הערכים באותו משתנה. החציון הוא הערך שמחלק את המדגם לשני חלקים שווים, כאשר מחצית מהערכים נמוכים ממנו ומחציתם גבוהים ממנו. שיטה זו עמידה יותר בפני ערכים קיצוניים ואינה מושפעת מריכוזי ערכים חריגים, מה שהופך אותה למתאימה במקרים של נתונים לא סימטריים.

שיטות לטיפול בערכים חריגים

שיטת סטיית התקן (Standard Deviation)

גישה זו מתבססת על פיזור הנתונים סביב הממוצע. ערכים המרוחקים מהממוצע מעבר למספר מוגדר של סטיות תקן נחשבים לחריגים.

שיטת הפרשי הרבעונים (Interquartile Range)

שיטה זו, המתאימה במיוחד לנתונים שאינם מתפלגים נורמלית, משתמשת ברבעונים לזיהוי חריגים. ערכים הנמצאים מחוץ לטווח המוגדר על ידי הרבעון הראשון (Q1) והשלישי (Q3) בתוספת מכפלה של ה IQR- מסווגים כחריגים.

<u>ערך הסף הרצוי</u>

ערך הסף מהווה את הגבול בין ערכים רגילים לחריגים. בשיטת סטיית התקן, הוא נקבע כמספר סטיות תקן מהממוצע, ובשיטת IQR כמכפלה של הטווח הבין-רבעוני. בחירת ערך הסף היא קריטית ותלויה בהקשר הספציפי של המחקר. ערך סף נמוך מדי עלול לזהות יותר מדי חריגים, בעוד ערך סף גבוה מדי עלול להחמיץ חריגים משמעותיים.

בחירת המודלים

אנו מתמודדים עם שאלת חיזוי מסוג רגרסיה.

משתני המטרה: רמות התחלואה של מחלות נשימה כרוניות (Asthma/COPD)

הסבר על 3 המודלים שאיתם נעבוד :

1. **רגרסיה ליניארית** היא שיטה פשוטה ויעילה לחיזוי ערכים רגרסיים בהתבסס על קשר לינארי בין המשתנים הסבירים לרמת התחלואה של מחלות נשימה כרוניות. היא מנסה למצוא את הקו הטוב ביותר שמתאר את הקשר בין קלטים (משתנים חברתיים וסביבתיים) לבין הפלט (רמת התחלואה).
2. **יערות רנדומיים** הוא אלגוריתם שבונה על ידי כמה עצי החלטה, שכל אחד מהם נבנה באופן עצמאי והם מכונים יחדיו לחיזוי ערך סופי,בזכות ריבוי עצי ההחלטה, המודל יכול לשפר את הדיוק של החיזויים.
3. **XGBoost** משתמש בטכניקת למידה חזקה שבה עצי החלטה בנויים בשלבים וכל עץ נוסף מתמקד בהשגת שגיאה קטנה יותר מהקודם. המודל עמיד לעומס ויכול לעבד נתונים גדולים בצורה יעילה ומספק חישובים מהירים

Cross-Validation

קרוס-ולידציה היא טכניקה להערכת ביצועי מודל על נתונים חדשים. היא מחלקת את הנתונים לקבוצות, מאמנת על רוב הקבוצות ובודקת על אחת, וחוזרת על התהליך עד שכל קבוצה שימשה לבדיקה. על ידי הרצת המודל עם פרמטרים שונים על קבוצות שונות, ניתן לזהות את הפרמטרים שמביאים לביצועים הטובים ביותר. הביצועים הסופיים מחושבים כממוצע על פני כל הסיבובים. הבחירה בטכניקה היא חיונית להבטחת אמינות, דיוק ועמידות של המודל הלומד מנתוני האימון.

היפר פרמטרים:

את המודלים נבחן עם שינוי בהיפר פרמטרים על מנת למצוא את התוצאה הטובה ביותר.

- n_estimators: מספר העצים במודל יערות רנדומיים. יותר עצים מובילים לדיוק גבוה יותר.
- max_depth: העומק המקסימלי של העצים. עומק גדול מדי עלול לגרום ל-overfitting.
- learning_rate: קצב העדכון של משקלי המודל במהלך האימון.
- min_samples_split: מספר הדגימות המינימלי הנדרש לפיצול צומת בעץ.
- min_samples_leaf: מספר הדגימות המינימלי הנדרש בעלה של העץ.
- colsample_bytree: אחוז העמודות הנבחרות אקראית בכל עץ. ערך גבוה יותר מגדיל את המגוון.
- reg_lambda: פרמטר לענישת L2. ערך גבוה מוביל לענישה חזקה יותר.
- cv: מספר הקיפולים בקרוס-ולידציה.

מטריקות להערכת המודל

- **Rבריבוע :(R-squared)** אחוז **השונות** במשתנה המטרה המוסברת ע"י המשתנים המנבאים.
  ערך 1 מציין התאמה מושלמת, 0 מציין חוסר התאמה.
- **MSE (Mean Squared Error)**: ממוצע ריבועי השגיאות בין הערכים החזויים לאמיתיים.
  ערך נמוך יותר מצביע על ביצועים טובים יותר.
- **MAE (Mean Absolute Error)**: ממוצע הערך המוחלט של השגיאות.
  ערך נמוך יותר מצביע על ביצועים טובים יותר.
- **RMSE (Root Mean Squared Error):** שורש ריבועי של MSE.
  נותן מידה של השגיאה ביחידות המקוריות של הנתונים שבמקרה שלנו משתנה המטרה נע בין 0 ל 100. כאשר הערך המקסימלי ב COPD הוא 30 ו ב אסתמה הוא 25.

השוואת ביצועי המודלים תתבצע באמצעות שילוב מטריקות יעילות, כגון R^2 וRMSE,

שילוב זה יאפשר לנו להשוות בצורה מדויקת את ביצועי המודלים השונים ולבחור את המודל הטוב ביותר עבור המחקר שלנו לדוגמא, שילוב R^2 גבוה ו RMSE-נמוך יצביע על התאמה מיטבית בין הערכים החזויים של המודל לערכים בפועל, תוך צמצום השגיאות. בחירת שתי מטריקות אלו התבססה על קלות ההבנה והאינטואיטיביות שלהן, המאפשרות פרשנות ברורה של תוצאות השוואת המודלים.

Feature Importance

לאחר שנקבל את המודל שמספק את התוצאות הטובות ביותר, נרצה לבחון את המשתנים שבהם המודל משתמש לצורך החיזוי. לשם כך, נשתמש בטכניקה של feature importance (חשיבות תכונות).

שיטת feature importance היא טכניקה חיונית בניתוח נתונים ולמידת מכונה. מטרתה העיקרית היא לזהות ולדרג את התכונות (משתנים) החשובות ביותר מתוך מערך נתונים גדול, על פי תרומתן למודל הניבוי או הסיווג. שיטה זו מאפשרת לנו להבין אילו משתנים משפיעים ביותר על התוצאות, ובכך מספקת תובנות חשובות לגבי המנגנונים העומדים בבסיס התופעה הנחקרת.

במקרה שלנו, נשתמש במדד ה-'gain' לחישוב חשיבות התכונות. מדד זה מודד את התרומה היחסית של כל תכונה לשיפור הדיוק של המודל. הוא מחושב על ידי בחינת כמה כל תכונה מפחיתה את פונקציית האובדן (loss function) של המודל.

חשיבות התכונות מוצגת באחוזים, כאשר הסכום של כל האחוזים מגיע ל-100%. לדוגמה, אם משתנה מסוים מחזיק ב-20% מהחשיבות, משמעות הדבר היא שהוא אחראי ל-20% מיכולת החיזוי של המודל.

שימוש בשיטה זו מאפשר לנו לקבל תובנות עמוקות יותר על הגורמים המשפיעים ביותר על שכחיות המחלות.

Pipeline

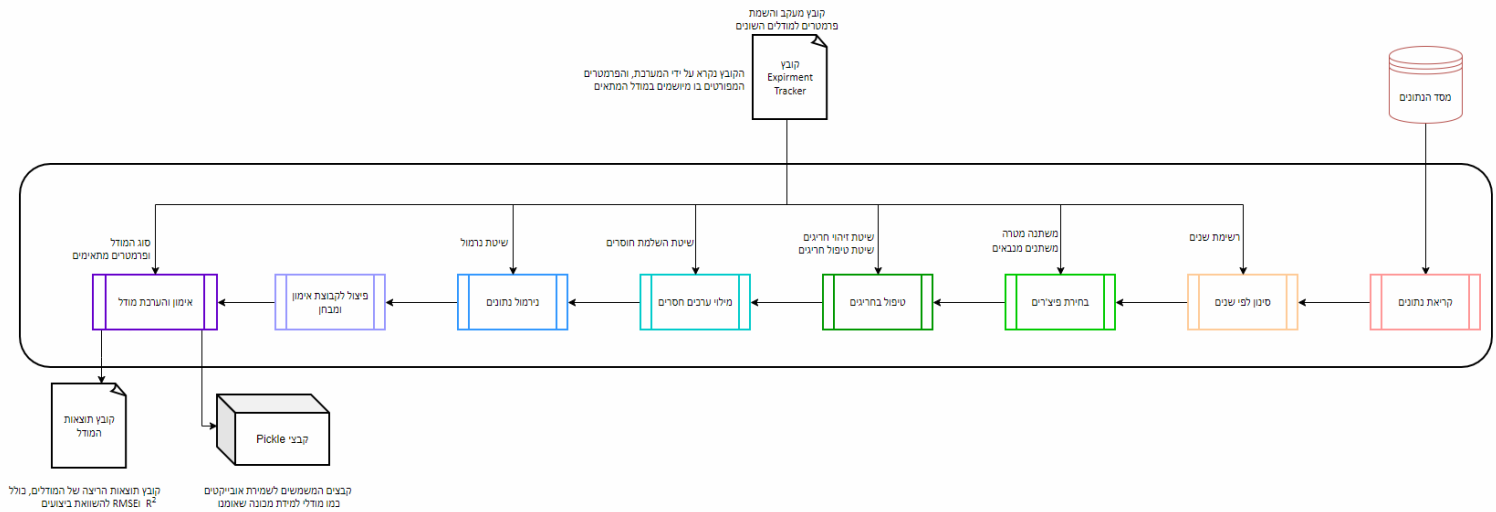
ה - Pipeline שפיתחנו מהווה גישה מקיפה ומובנית לניתוח נתונים. התחלנו בבניית פונקציה לבחירת תכונות רלוונטיות מתוך מסד הנתונים הגדול, המאפשרת לנו להתמקד במשתנים החשובים למחקר הספציפי שלנו. לאחר מכן, טיפלנו באיכות הנתונים: יצרנו פונקציות לזיהוי וטיפול בנקודות קיצון ובערכים חסרים , שני צעדים קריטיים לשמירה על דיוק הניתוח שלנו.

המשכנו עם פונקציה שמאפשרת לנו לסנן את הנתונים לפי שנים ספציפיות , מה שמאפשר לנו לבחון מגמות לאורך זמן או להתמקד בתקופות מסוימות. לבסוף הנתונים עוברים נרמול , שלב חיוני להכנתם למודלים של למידת מכונה. בשלב הבא, פיתחנו פונקציה שמחלקת את הנתונים לקבוצות אימון ובדיקה , ואז מפעילה עליהם מודלים שונים של למידת מכונה .

הפונקציה הסופית שלנו, מדגימה את הגישה הגמישה והיעילה שלנו לניהול ניסויים מרובים:

1. **קריאת פרמטרים:** בחרנו לקרוא את פרמטרי הניסויים שלנו מקובץ אקסל חיצוני. זה מאפשר לנו להגדיר מגוון רחב של תרחישי ניסוי מבלי לשנות את הקוד עצמו.
2. **ביצוע ניסויים**: עבור כל שורה בקובץ האקסל (כל ניסוי), הפונקציה שלנו מבצעת את כל שלבי העיבוד והניתוח, תוך שימוש בפרמטרים שהגדרנו.
3. **שמירת תוצאות**: בסוף כל ניסוי, אנו כותבים את התוצאות לקובץ אקסל. זה מאפשר לנו לעקוב אחרי התקדמות הניסויים שלנו ולשמור על תיעוד מלא של כל התוצאות.

שיטת העבודה שפיתחנו מאפשרת לנו גמישות רבה בתכנון וביצוע הניסויים שלנו, מקלה עלינו להשוות בין תרחישים שונים, ומספקת לנו תיעוד מסודר של כל הניסויים. היא מתאימה במיוחד למחקרים המורכבים שלנו הדורשים ריבוי ניסויים עם וריאציות שונות של פרמטרים, ומאפשרת לנו לנהל את הניסויים שלנו ביעילות ושקיפות.



סיכום פרק שיטות

בפרק זה הצגנו את המתודולוגיה המקיפה שנשתמש בה לחקירת הקשר בין גורמים חברתיים לתחלואה במחלות נשימה כרוניות. בדיקה ראשונית של הנתונים הראתה כי אף אחד מהמשתנים אינו מתפלג נורמלית, מה שמדגיש את הצורך בשיטות סטטיסטיות מתאימות לנתונים שאינם פרמטריים.

אנו מתכננים להתמודד עם אתגרים מרכזיים בניתוח הנתונים, כולל טיפול בערכים חסרים וחריגים, נבחן מספר מודלים, החל מרגרסיה ליניארית ועד מודלים מתקדמים כמו Random Forest ו XGBoost ונשתמש ב Cross-Validation-ומדדי ביצוע מקובלים להערכתם.

פיתחנו Pipeline ייחודי שיאפשר לנו ניהול יעיל של ניסויים מרובים, כולל שלבים של בחירת תכונות, טיפול בנתונים, ואימון והערכת מודלים. נדגיש את חשיבות ניתוח חשיבות התכונות כדי להבין את הגורמים המשפיעים ביותר על התוצאות.

באמצעות גישה מקיפה זו, אנו מתכוונים לחקור ביסודיות את הקשר בין גורמים חברתיים-כלכליים לבריאות הנשימתית. תהליך זה יאפשר לנו לבחון את שאלת המחקר מזוויות שונות.

הוספת משתנים

במסגרת המחקר שלנו הוחלט להוסיף שני פיצ'רים מרכזיים: השמנה ועישון. החלטה זו מבוססת על הסיבות הבאות:

- **השפעת השמנה על מחלות נשימה** :השמנת יתר ידועה כגורם סיכון משמעותי לתחלואה במחלות נשימה כרוניות כמו אסתמה ו .COPD-מחקרים רבים מצביעים על כך שהשמנת יתר יכולה להחמיר את התסמינים ולהוביל להחמרה במצב הבריאותי הכללי של חולים עם מחלות נשימה.
- **השפעת העישון על מחלות נשימה** :עישון הוא אחד הגורמים המרכזיים להתפתחות ולהחמרה של מחלות נשימה כרוניות. הוספת משתנה עישון מאפשרת לנו להבין טוב יותר את הקשר בין עישון לתחלואה ולשפר את מודלי החיזוי שלנו.

הוספת משתנים אלו מאפשרת לנו לשפר ולהגיע לחיזוי בצורה מדויקת יותר.

משתני הרגלי עישון

קובץ Lifestyle_FG להלן "הרגלי חיים"

חבילת קבצים משנים 2014-2019 ,מדובר בנתונים הקשורים לעישון והשמנה והם מחולקים למספר מדדים שונים תחת כותרות שונות.
פירוט האינדיקטורים בקבצים:

- SMOK002: אחוז האנשים שהפחיתו את מספר הסיגריות שהם מעשנים ביום לאחר קבלת ייעוץ רפואי, בניכוי מקרים יוצאי דופן.
- SMOK004: אחוז האנשים שעברו בדיקות רפואיות הקשורות לעישון (כגון בדיקות תפקוד ריאות) לאחר קבלת ייעוץ רפואי, בניכוי מקרים יוצאי דופן.
- SMOK005: אחוז האנשים שהצטרפו לתוכניות גמילה מעישון לאחר קבלת ייעוץ רפואי, בניכוי מקרים יוצאי דופן.

לא היה משתנה של Prevalence לעישון ולכן לא נלקח.

משתני השמנה

- Obesity_Prevalence (per_cent) : אחוז האנשים באוכלוסייה עם BMI מעל 30.

להשמנה היה רק אינדיקטור אחד שערכו 100 לאורך כל השנים ולכן התעלמנו ממנו.

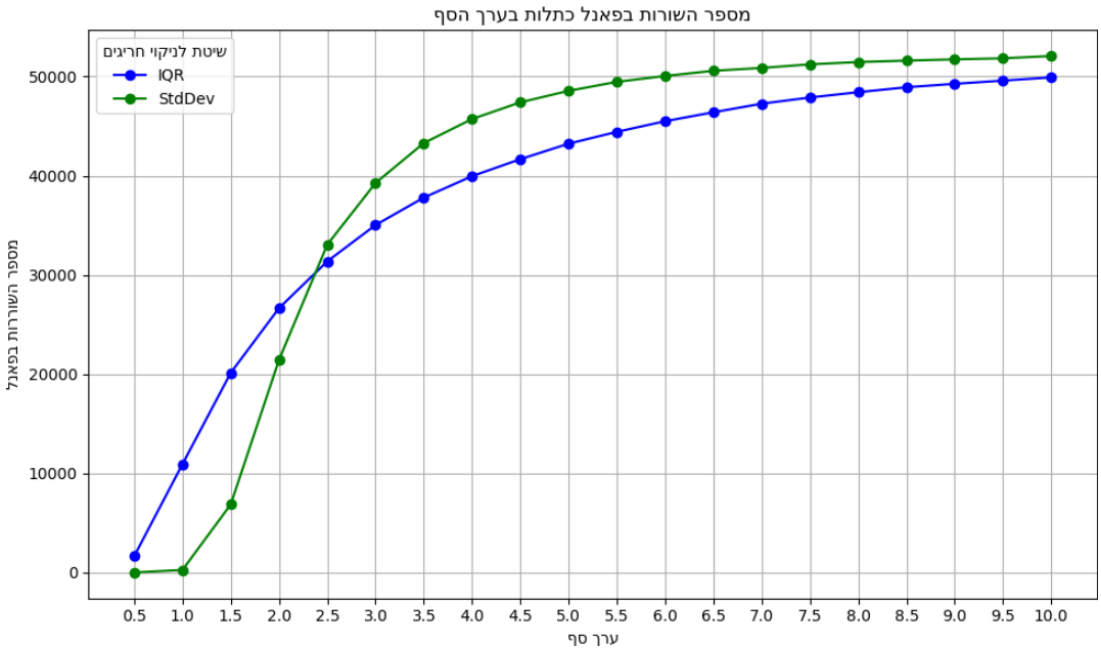
טיפול בערכים חריגים

הקדמה

טיפול בערכים חריגים הינו מרכיב קריטי בתהליך ניתוח נתונים, המשפר את איכות המידע ואמינות המודלים האנליטיים. נתמקד בשתי שיטות מרכזיות לזיהוי וטיפול בחריגים כפי שהצגנו בפרק שיטות.

חקירת ערכים חריגים

ראשית, לפני שנסתכל על המשתנים מקרוב, נרצה להבין כיצד גודל הנתונים שלנו משתנה בהתאם לערך הסף שנבחר לניכוי החריגים.



בגרף שלהלן ניתן לראות את כמות השורות בסט הנתונים לאחר ניכוי חריגים כתלות בערך הסף. הגרף ממחיש כיצד גודל סט הנתונים משתנה בהתאם לערך הסף שנבחר. ניתן להבחין כי גודל סט הנתונים מושפע באופן משמעותי מערך הסף שנבחר. ככל שערך הסף גבוה יותר, כך נהיה סלחניים יותר לגבי הנתונים שלנו, וניכוי החריגים יהיה פחות מחמיר. בהתאם, כאשר ערך הסף נמוך יותר, ייכללו פחות חריגים בסט הנתונים, וכתוצאה מכך גודל הסט יהיה קטן יותר.

מסקנה זו מדגישה את החשיבות בבחירת ערך הסף המתאים, שכן הוא משפיע על איזון בין שמירה על איכות הנתונים לבין גודל סט הנתונים שנשאר לעיבוד וניתוח. אנו מבינים כי אין ערך סף אולטימטיבי, ונרצה למצוא פשרה שמצד אחד תכלול כמות נתונים מספקת, ומצד שני עדיין תוודא שהמודל שלנו מבוסס על נתונים איכותיים. לשם כך, נחפש את הנקודה שממנה העלייה בגרף נעשית מתונה יותר, מה שמעיד על איזון מיטבי בין סילוק החריגים לשמירה על כמות הנתונים.

במקרה שלנו אנו נרצה לעבוד עם ערכי הסף הבאים:

IQR: ערך הסף עומד באזור 2.5

STDEV: ערך הסף עומד באזור 3.

מעתה שאר הניתוחים שיוצגו יהיו בעל אותו ערך סף שהוגדר למעלה.

המשתנים עם אחוז החריגים הגדול ביותר

IQR

להלן המשתנים בעל אחוז החריגים הגבוה ביותר עם ערך סף 2.5

שם המשתנה	ערך מקסימום	ערך מינימום	גבול עליון	גבול תחתון	אחוז חריגים	מספר אפסים	מספר ערכים ריקים
אינדיקטור עישון 005	100	3.157	106.931	88.777	7.819	0	8140
אינדיקטור COPD 007	100	0	107	90.200	7.549	143	8042
אינדיקטור COPD 003	100	0	106.82	78.260	7.060	164	8043

ניתן לראות כי שלושת המשתנים המובילים בטבלה, הם משתנים מסוג אינדיקטורים.
הטווח שלהם נע בין 0 – 100, אולם ניתן להניח כי ישנן מרפאות אשר לא ביצעו מדידה כלל ורשמו במדד 0, דבר אשר יכול להשפיע על התוצאות .

לשם כך ביצענו בדיקה נוספת של אחוז החריגים לאחר הצבת ערך 'NULL' בתא שבו יש את הערך 0 .
בסיום הבדיקה גילינו כי **התוצאות לא השתנו כלל ולכן** ניתן להניח כי משתנים שערכם הוא 0 אינם משפיעים על תוצאות החישוב של ערכי הקיצון, בשל ההשפעה המזערית שלהם.

STDEV

להלן המשתנים בעל אחוז החריגים הגבוה ביותר עם ערך סף 3.

שם המשתנה	ערך מקסימום	ערך מינימום	גבול עליון	גבול תחתון	אחוז חריגים	מספר אפסים
מרחק ממוצע מפארק קרוב	13929	111	3893.974	-2028.109	2.665	0
מספר מיקודים בטווח 900 מ' מפארק	1	0.03	1.248	0.633	2.533	0
אחוז כתובת עם שטח חצוני פרטי	0.98	0.14	1.113	0.663	2.157	0

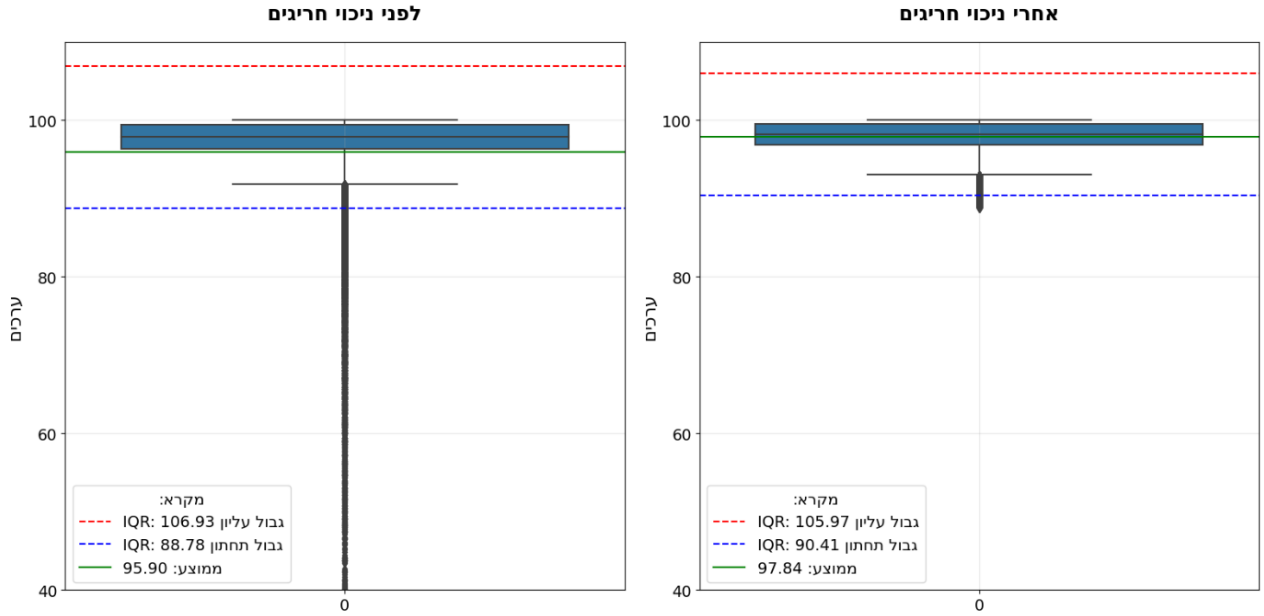
נראה שיש הבדלים בין המשתנים בין השיטות שבהם השתמשנו. בנוסף, ניתן לראות שהמשתנים הנ"ל לא מכילים ערך של 0 בכלל.

תצוגה גרפית

IQR

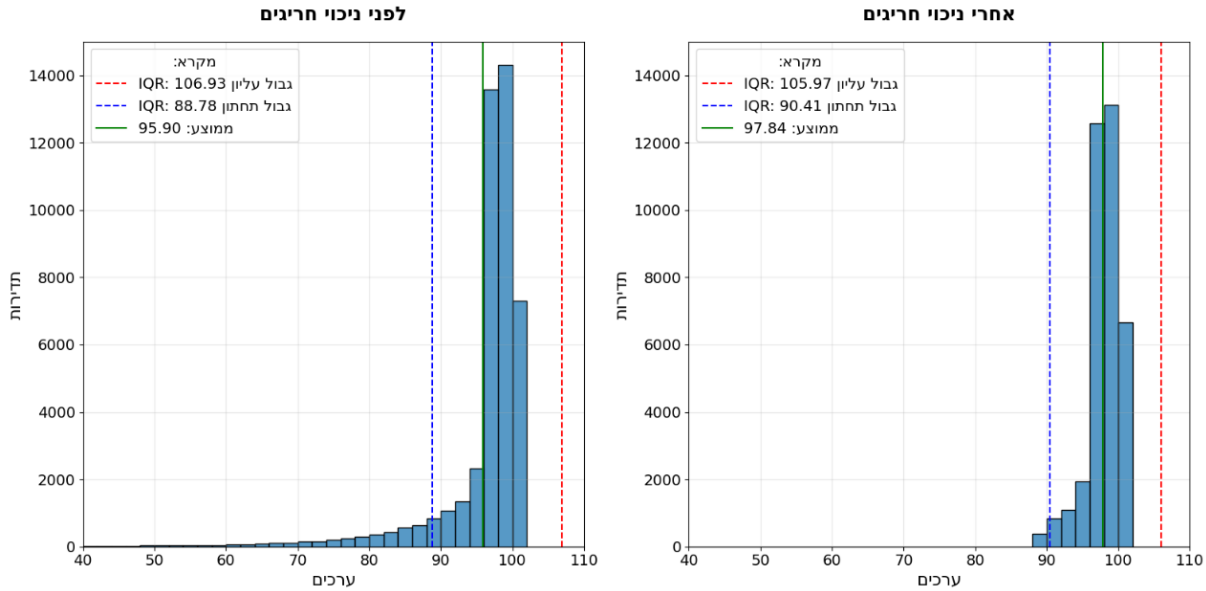
משתנה : "אינדיקטור עישון 005" , עם ערך סף = 2.5

השוואת הטיפול בערכים חריגים בעזרת Boxplot



ניתן לראות כיצד טווח הערכים מצטמצם לאחר הטיפול בחריגים, אך יחד עם זאת הגבולות והממוצע לא השתנו באופן משמעותי. הקופסאות בשני התרשימים, המייצגות את מרבית הנתונים, נשארו דומות למדי.

התפלגות הערכים של 'אינדיקטור עישון 005' :

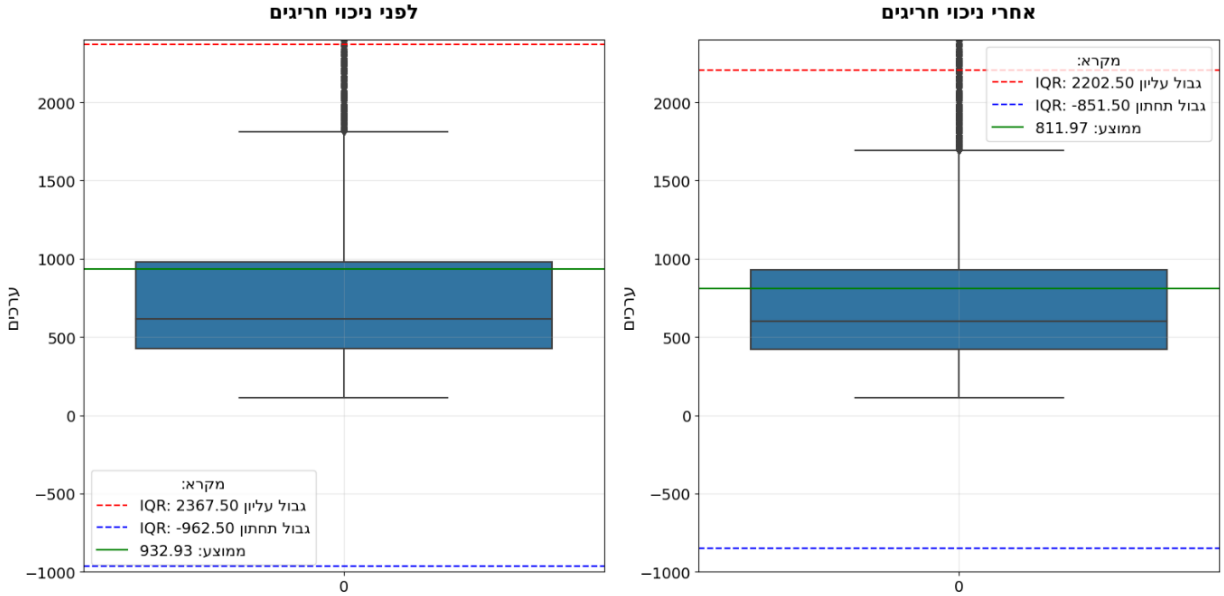


ניתן לראות כי התפלגות תדירות הנתונים לא השתנתה באופן משמעותי לאחר ניכוי החריגים רוב הערכים מרוכזים בטווח הגבוה, בין 90 ל-100. זה מעיד על כך שהטיפול בחריגים לא שינה את המגמה הכללית של הנתונים. ההבדל העיקרי נראה בקצוות ההתפלגות, בעיקר בערכים הנמוכים יותר, שם נראית הפחתה קלה במספר המקרים לאחר ניכוי החריגים.

STDEV

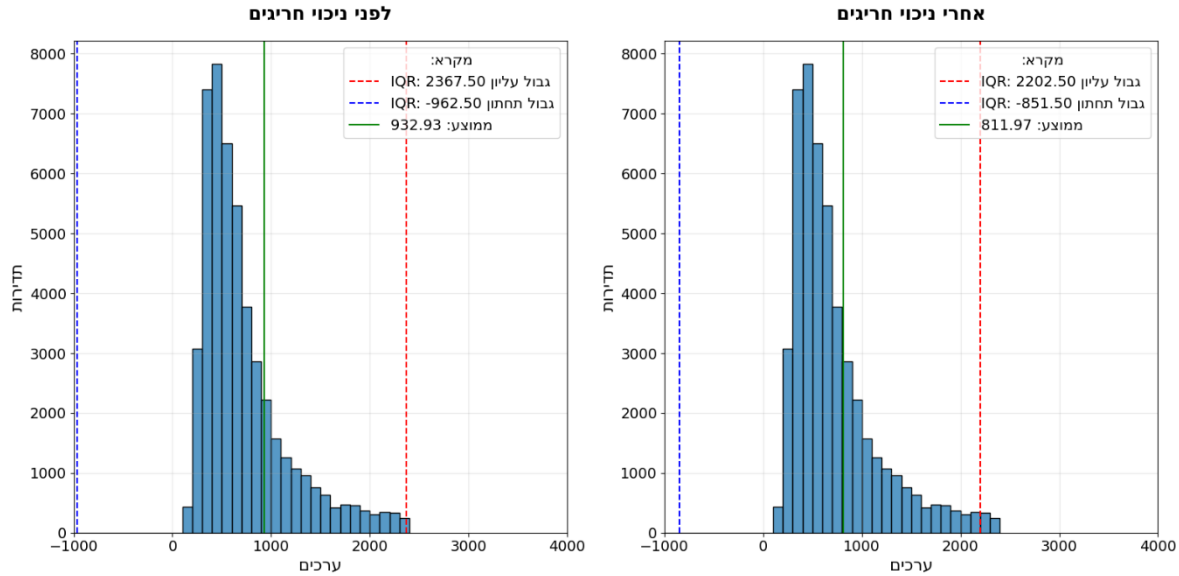
משתנה : " מרחק ממוצע מפארק קרוב" , עם ערך סף = 3.

השוואת ערכים חריגים בעזרת Boxplot



מהגרף הנ"ל ניתן לראות שהחריגים באים מהגבול העליון. כמו כן ניתן לראות איך הממוצע (הקו הירוק) קטן והגבולות הצטמצמו לאחר ניכוי החריגים. הקופסה, המייצגת את הרבעון השני והשלישי של הנתונים, נשארה יחסית דומה בגודלה, אך הגבול העליון שלה ירד מעט. זה מצביע על כך שהטיפול בחריגים השפיע בעיקר על הערכים הגבוהים ביותר.

התפלגות הערכים של 'מרחק ממוצע מפארק קרוב'



ניתן לראות כי השינוי בין שתי ההיסטוגרמות הוא קטן יחסית, וזאת בגלל שאחוז החריגים בנתונים המקוריים היה מאוד נמוך. עיקר ההתפלגות נשארת דומה בשני המקרים, עם רוב הערכים מרוכזים באותו טווח. ניתן לראות הפחתה קלה במספר הערכים הגבוהים מאוד בהיסטוגרמה השנייה, אך צורת ההתפלגות הכללית במעט ולא השתנתה.

החלפת חריגים בערך ריק

לאחר חקירה של סט הנתונים, הבנו כי עלינו למצוא פשרה בין ערך סף לכמות השורות בסט הנתונים. הבנו כי ככל שנעלה את ערך הסף נקבל יותר שורות אך מהלך זה עלול לפגוע באיכות הנתונים. לכן עלינו לבחון את האפשרות שבמקום מחיקת השורות עם ערך חריג נחליף את הערך החריג בערך ריק, ונטפל בו כמו שאנחנו מטפלים בערכים חסרים (ראה בהמשך פרק ערכים חסרים).

לצורך כך ביצענו בדיקה על מנת לראות את האחוז השורות הריקות בסט הנתונים בשתי השיטות ללא שנת 2013.

1. מחיקת השורה בעל ערך חריג
2. החלפת הערך החריג בערך ריק

בטבלה למטה בחנו את שלושת המשתנים בעלי אחוז החריגים הגבוה ביותר. בדקנו כיצד שתי שיטות טיפול בערכים חריגים משפיעות על מספר השורות בסט הנתונים.

		החלפת חריג בערך ריק		מחיקת ערך חריג	
שיטה	שם המשתנה	מספר השורות בסט הנתונים	אחוז השורות עם ערך ריק	מספר השורות בסט הנתונים	אחוז השורות עם ערך ריק
IQR ערף סף = 2.5	אינדיקטור עישון 005	44682	9.491518	25606	0.007811
	אינדיקטורCOPD 007	44682	8.954389	25606	0.011716
	אינדיקטורCOPD 003	44682	8.379213	25606	0.015621
StDev ערך סף = 3	גודל ממוצע של גינה פרטית בדירות	44682	4.191845	27421	0.590788
	מרחק ממוצע מפארק קרוב	44682	4.176178	27421	0.590788
	אינדיקטור עישון 005	44682	3.726333	27421	0.007294

נתמקד בתוצאות עבור אינדיקטור עישון 005:

- שיטת המחיקה:
  - מספר השורות הצטמצם מ-44,682 ל-25,606. הפחתה של כ-42.7 % מהנתונים המקוריים.

- שיטת ההחלפה לערך ריק:
  - נקבל כ-9.49% שורות ריקות עבור אותו אינדיקטור.
  - שורות אלו ניתנות להשלמה בשלב הבא באמצעות ממוצע או חציון.

<u>ניתוח התוצאות:</u>

יש מקום לשקול את אופציית החלפת ערכים חריגים לחסרים. גישה זו עשויה לאפשר שימור של יותר מידע, מה שעשוי לתרום לאיכות הניתוח הסטטיסטי.
עם זאת, ההחלטה הסופית תלויה בשיקולים נוספים כגון מטרות המחקר, אופי הנתונים, והשפעת הערכים החסרים על המודל הסופי.

סיכום ערכים חריגים

לאחר ניתוח מעמיק של המשתנים וחקירת שתי שיטות לזיהוי ערכים חריגים, הגענו למסקנה כי נדרש איזון עדין בין שמירה על איכות הנתונים לבין שימור מספר מספק של רשומות במסד הנתונים. כדי להשיג איזון זה, החלטנו לבחון את המודל תוך שימוש בשלושה ערכי סף שונים: 2, 2.5, ו-3. בחירת ערכי סף אלה מאפשרת לנו לבדוק מגוון רחב של תרחישים.

בנוסף לבחינת ערכי הסף השונים, נחקור את האפקטיביות של החלפת ערכים חריגים בערכים חסרים (NULL). שיטה זו מציעה מספר יתרונות על פני מחיקה מוחלטת של רשומות:

1. שמירה על מבנה הנתונים הכולל.
2. מניעת אובדן מידע בשדות אחרים באותה רשומה.

באמצעות בדיקה מקיפה זו, נוכל להעריך את ההשפעה של כל גישה על איכות הנתונים, כמות המידע הזמין, ולבסוף על ביצועי המודל. זה יאפשר לנו לקבל החלטה מושכלת לגבי הגישה המיטבית לטיפול בערכים חריגים בנתונים הרפואיים שלנו, תוך איזון בין הצורך בדיוק לבין הרצון לשמור על מירב המידע הרלוונטי.

טיפול בערכים חסרים

כפי שהוצג בסמסטר א' להלן טבלה של ערכים חסרים :

משתנה	כמות חוסרים	אחוז מכלל הפאנל	הערות	דרכי טיפול
שביחות AST	1	0.019%	משתנה מטרה	מחיקת שורות
שביחות COPD	10	0.0192%	משתנה מטרה	מחיקת שורות
אינדיקטור - כללי	8073	15.32%	בשנת 2013 לא קיים	הפרדת מודלים
אינדיקטור – COPD006	44796 (2013 ללא 36723)	85% (2013 ללא 69.68%)	במהלך השנים השתנה ל AST007	מיזוג האינדיקטורים / הפרדת מודלים
אינדיקטור – COPD007	15950 (2013 ללא 7877)	30.26% (2013 ללא 14.95%)		
גישה לפארקים וגינות	464	0.88%		השלמת ערכים
קיפוח	464	0.88%		השלמת ערכים
הכנסה	464	0.88%		השלמת ערכים
גיל ומגדר	109	0.20%		השלמת ערכים

<u>להלן הפירוט איך התמודדנו עם כל נושא</u>

ערכים חסרים במשתנה המטרה – ערכים אלו לא הושלמו מכיוון שקיימת סכנה להטיית הנתונים בניתוח. לכן, הם נמחקו. ניתן לראות כי הם מהווים פחות מעשירית האחוז.

<u>איחוד אינדיקטורים</u>

בשלב זה, אינדיקטורים 6 ו-7 אוחדו לאינדיקטור אחד, לאחר בדיקה שאין ערכים כפולים באותה שנה. מטרת האיחוד היא לפשט את הנתונים ולצמצם את מספר המשתנים במודל האנליטי.

<u>בידוד שנת 2013</u>

שנת 2013 בודדה מתוך הנתונים מאחר שבשנה זו לא נאספו אינדיקטורים מסוימים של עישון והשמנה. על מנת להימנע מבעיות בניתוח הנתונים שנגרמות מהיעדר מידע, הוחלט להריץ את המודל לפי שנים ולהוציא את שנת 2013 מהנתונים הכלליים או להריץ עם 2013 ללא האינדיקטורים המצוינים למעלה.
פעולה זו מאפשרת לבצע ניתוח מדויק ואמין יותר, תוך שמירה על שלמות הנתונים בשנים אחרות.

<u>זוהי תמונת מצב של הפיצ'רים לאחר סינון של שנת 2013:</u>

שם המשתנה	מספר השורות	אחוז השורות עם ערך ריק
קיפוח, הכנסה, גישה	257	**0.0057%**
אינדיקטורים	171	**0.0388%**
גיל ומגדר	109	**0.0024%**

ניתן לראות כי ללא 2013 אחוז הערכים החסרים שואף לאפס, ולכן אין חשיבות לשיטת הטיפול בחוסרים.

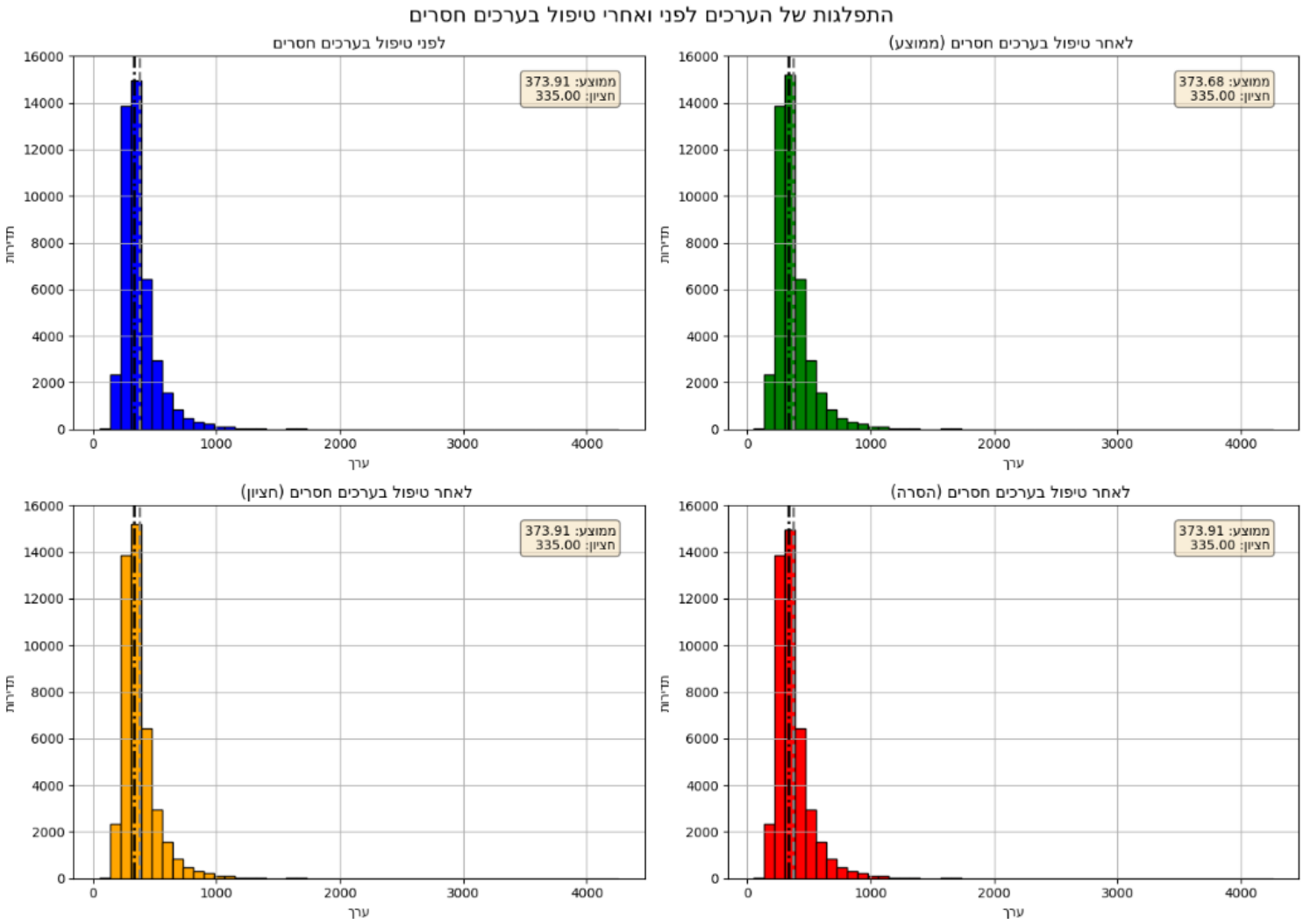
ניתוח דרכי טיפול בערכים חסרים

לצורך בחינה של שיטת הטיפול בערכים החסרים אנו נבחן שלוש חלופות:

- **חלופה 1 –** השלמת חוסרים לאחר סינון של שנת 2013
- **חלופה 2 –** השלמת חוסרים עם שנת 2013

חלופה 1

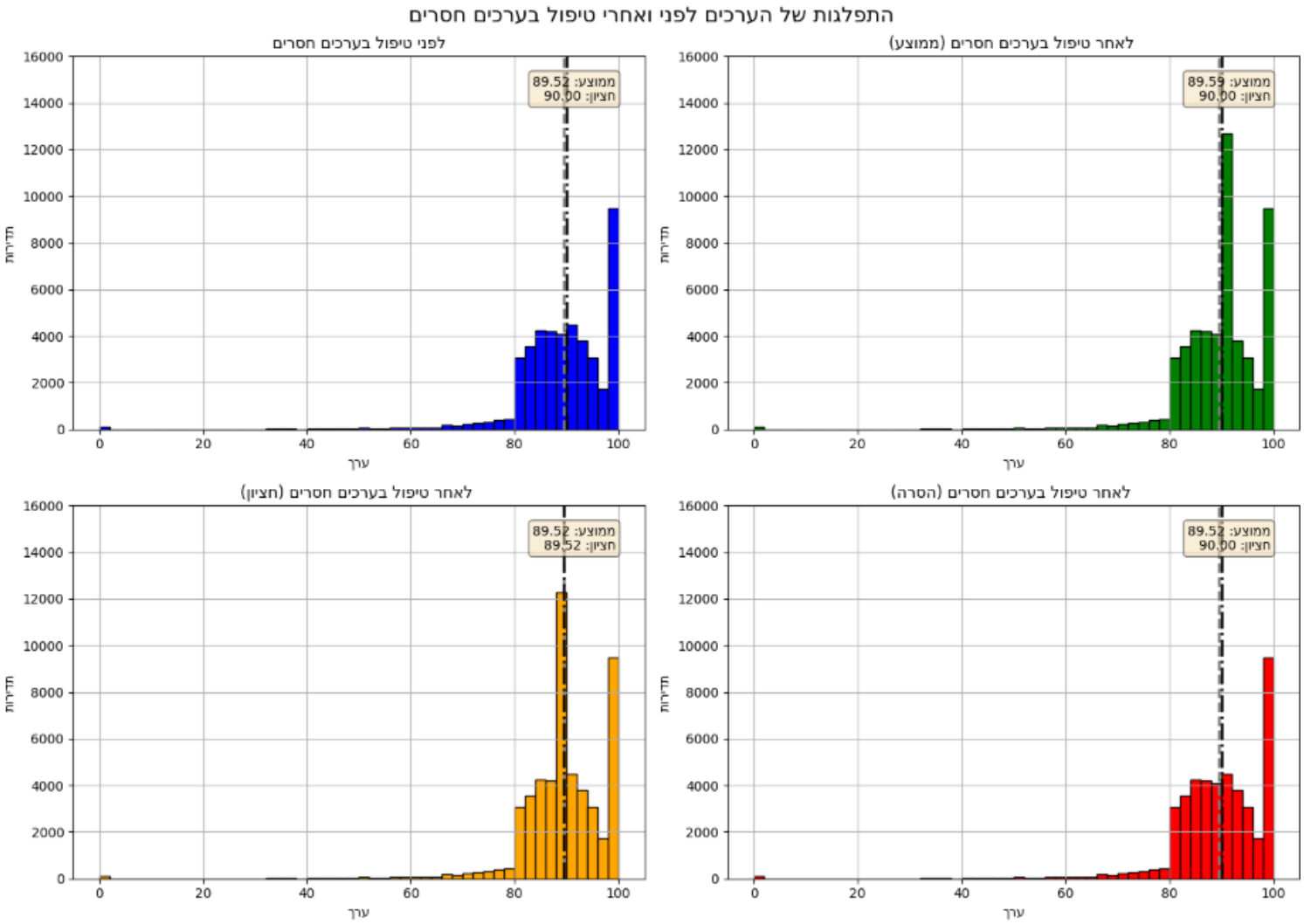
הגרפים הבאים מתארים את כמות החוסרים ואת דרכי הטיפול בערכים החסרים במשתנה – " מרחק ממוצע מהפארק הקרוב " משתנה זה נבחר כדוגמא מתוך קבוצה של משתנים בעל אותו אחוז חוסרים.



כפי שניתן לראות, מכיוון שאחוז החסרים מאוד נמוך, כל דרך פעולה שנבצע לא תשפיע כלל על הדאטה סט, ולכן אין חשיבות לדרך הפעולה.

חלופה 2

הגרפים הבאים מתארים את האינדיקטור – ” Asthma 004” ואת דרכי הטיפול בערכיה החסרים .



אם נשאיר את 2013 נקבל אחוז חוסרים מאוד גבוה בתחום האינדיקטורים - כ 15.4% . ניתן לראות בבירור על עמודה חדשה שמתווספת לגרפים כתלות בבחירה דרך הפעולה. לגבי שאר המשתנים, אחוז החוסרים שם עומד על פחות מ 1%

נוסף על כך בדקנו את המרחק האבסולוטי בין הממוצע לחציון בקטגוריית האינדיקטורים וראינו שהשוני בין השניים הוא לכל היותר 2% לכן ההבדל הוא מינורי אם נבחר בין ממוצע לחציון.

סיכום ערכים חסרים

1. הסרת 2013 מסט הנתונים תביא לאחוז חוסרים מאוד קטן ולכן ניתן למחוק את השורות.
2. קיימת אפשרות להחליף ערכים חריגים בערכים חסרים ולבצע השלמת ערכים.

אנו מעוניינים לבחון את המודל בשתי אפשרויות:

1. ללא נתוני 2013, עם כל האינדיקטורים.
2. עם נתוני 2013, ללא אינדיקטורים.

גילינו כי עבור האינדיקטורים אין הבדל משמעותי בין החציון לממוצע. לכן, אין חשיבות לטכניקה ספציפית של השלמת הערכים החסרים עבור משתנים אלה.

תהליך החקירה

הקדמה

במהלך תהליך זה, אנו מתמקדים בחקירה ובשיפור ביצועי המודלים השונים בעזרת מספר פרמטרים ושלבי עיבוד נתונים. אנו מתמקדים בסט נתונים הכולל שתי מחלות נשימה שונות: אסתמה ומחלת ריאות נשימתית כרונית (COPD) החלוקה הזו מבטיחה שאין התנגשויות בין המודלים שלנו ומאפשרת ניתוח מעמיק של מספר רחב של קומבינציות במסגרת הנתונים. אנו מאתחלים את המודלים שלנו עם פרמטרים כמו שנים לבדיקה, שיטות חישוב ממוצע או חציון, טיפול בנתונים חריגים, משתנה סף ועוד.

החקירה תתחיל על ידי בדיקת היכולת של רגרסיה לינארית להתאמה לנתונים הקיימים. רק לאחר מכן, יתכן שנעבור לחקירה של מודלים יותר מורכבים, בהתאם לתוצאות של רגרסיה לינארית.

שלב ראשון - שינוי הרכב המשתנים

נחקור את השפעת המשתנים השונים על תוצאות המודל. מטרתנו היא להבין כיצד ניתן לשפר את המודל ובאילו היבטים. נבחן את כל הצירופים האפשריים של שינויים בקבוצות המשתנים. לכל אחת משתי המחלות הנבדקות, ישנם כ-127 צירופים אפשריים. כתוצאה מכך, בשלב זה נבצע סך הכל 254 ריצות שונות של המודל. ניתוח מקיף זה יאפשר לנו לזהות את הגורמים המשפיעים ביותר על ביצועי המודל ולכוון את מאמצי השיפור בהתאם.

אתחול ראשוני:

- **מודל רגרסיה ליניארית**
- בחרנו בנתונים המתייחסים לשנים 2014-2019 ע"מ לכלול את האינדיקטורים, עישון והשמנה
- שיטת זיהוי ערכים חריגים – IQR עם ערך סף של 2.5
- שיטת טיפול בערכים חריגים – הסרת השורות
- שיטת השלמת ערכים חסרים – השלמה בממוצע
- שיטת נרמול – Min-Max

להלן התוצאות הטובות ביותר לכל משתנה מטרה :

משתנה מטרה	מקטגוריות המשתנים	RMSE	R^2
COPD	אינדיקטורים, מגדר וגיל, הכנסה ,קיפוח ,גישה לפארקים ,השמנה ,עישון	0.568	0.631
AST	מגדר וגיל ,הכנסה ,קיפוח ,השמנה ,עישון	1.025	0.3480

שלב שני - שינוי שיטת הטיפול בחריגים וערכי הסף

כפי שראינו בפרקים הקודמים , בחירה בין ממוצע לחציון אינה משפיעה על ההתפלגות הנתונים.
לקחנו את עשר ההרצות מכל משתנה מטרה, בהן ה R^2 היה הכי גבוה ובדקנו 3 אופציות שונות בהן כל פעם שינינו משהו אחד בלבד (כלומר הרצנו עוד 60 מודלים)

<u>שלוש האופציות הן:</u>

1. טכניקה לזיהוי חריגים סטיית התקן (STDEV) ערך סף של 3, טיפול בערכים חריגים - מחיקה
2. טכניקה לזיהוי חריגים IQR עם ערך סף של 2.5, טיפול בערכים חריגים - החלפה לערך ריק
3. טכניקה לזיהוי חריגים סטיית התקן (STDEV) ערך סף של 3, טיפול בערכים חריגים - החלפה לערך ריק

בחרנו לבצע את ההשוואה הראשונית בין המודלים לפי R^2 מכיוון להגדיל את אחוז השונות המוסברת , להלן תוצאות המודלים מהמובילים עבור כל מחלה לפי R^2

אופציה	משתנה מטרה	RMSE	R^2	הרכב משתנים
1	COPD	0.553	0.612	אינדיקטורים ,מגדר וגיל, הכנסה ,קיפוח, גישה לפארקים ,השמנה
	AST	1.010	0.344	אינדיקטורים ,מגדר וגיל ,הכנסה ,קיפוח
2	COPD	0.581	0.596	אינדיקטורים, מגדר וגיל, הכנסה, קיפוח, גישה לפארקים, השמנה, עישון
	AST	1.048	0.360	מגדר וגיל, הכנסה, קיפוח, השמנה, עישון
3	COPD	0.582	0.578	אינדיקטורים, מגדר וגיל, הכנסה, קיפוח, גישה לפארקים, השמנה, עישון
	AST	1.024	0.364	קיפוח, גישה לפארקים , השמנה , עישון

בתוצאות הנ"ל לא נמצא התאמה מוחלטת בין ה RMSE לבין R^2. כלומר ה R^2 הכי גבוה לא בהכרח יהיה עם ה RMSE הכי נמוך.

ביחס לשלב הקודם:

- COPD - במודל הנוכחי ניתן לראות ירידה קלה בערך ה R^2-מ-0.631 בשלב הקודם ל-0.612 בשלב הנוכחי. לעומת זאת, ה RMSE-השתפר במעט, יורד מ-0.568 ל-0.5533. למרות הבדלים אלו ניתן לומר כי השינוי הוא זניח.
- אסתמה - (AST) במקרה זה, יש שיפור קל ב R^2-מ-0.3480 ל-0.364, אך ה RMSE-נשאר ללא שינוי מ-1.025 ל-1.024.

<u>סיכום שלב שני</u>

השינויים בשיטות הטיפול בחריגים ובערכי הסף הובילו לשינויים מזעריים בלבד במודלים.
לאחר בחינת התוצאות, הבחנו כי בין ההרצות השונות התוצאות כמעט ואינן משתנות ביחס לתוצאות של השלב הראשון. משמעות הדבר היא שבמודל הרגרסיה, שינוי שיטת הטיפול בערכים חריגים וחסרים, וכן שינוי ערכי הסף, כמעט ואינם משפיעים על התוצאה הסופית. ממצאים אלו מחזקים את המסקנה כי בחירת המשתנים המתאימים למודל היא קריטית יותר מאשר שינויים בטכניקות עיבוד הנתונים.

שלב שלישי - מודלים לא לינאריים

לאחר ההרצה על רגרסיה לינארית, רצינו להחיל את אותן המודלים מהשלב הראשון על מודלים לא לינאריים כגון Random Forest ו-XGBoost במהלך הבדיקה, גילינו כי זמן הריצה של המודל היה ממושך ולא סביר, במיוחד עבור נתונים גדולים ומורכבים.
בשל המורכבות והזמן הדרוש להרצת מודלים לא לינאריים על כל הדגימות, החלטנו לשנות את האסטרטגיה שלנו. במקום להפעיל את כל המודלים על כל קבוצות הצירופים השונים בדומה לשלב הראשון, בחרנו להתמקד בעשר הדגימות הטובות ביותר על פי מדד ה-R^2 שהתקבל ממודל הרגרסיה הלינארית. גישה זו מאפשרת לנו לשמור על איזון בין יעילות ודיוק ולהפחית את זמן הריצה והמשאבים הנדרשים.

<u>חשיבות בדיקת מודלים הלא לינאריים</u>

- **מורכבות הקשרים בין המשתנים:**

במקרים רבים ,הקשרים בין המשתנים המסבירים למשתנה המוסבר אינם לינאריים .המשמעות היא שהשפעת שינוי במשתנה מסביר אחד על המשתנה המוסבר עשויה להשתנות בהתאם לערכים של משתנים אחרים או בהתאם לערכו של המשתנה עצמו.

- **יכולת למידה גבוהה יותר:**

מודלים לא לינאריים מסוגלים ללמוד קשרים מורכבים ובלתי לינאריים בנתונים, בניגוד לרגרסיה לינארית שבה הקשרים נבנים לפי קווים ישרים. גישות לא לינאריות כמו המודלים שציינו למעלה יכולות לתפוס אינטראקציות מורכבות בין המשתנים ולספק תחזיות מדויקות יותר.

<u>2 המודלים נבחנו על אותו קובץ של נתונים בשלב זה</u>

- בחרנו בנתונים המתייחסים לשנים 2014-2019 ע"מ לכלול את האינדיקטורים, עישון והשמנה
- שיטת זיהוי ערכים חריגים – IQR עם ערך סף של 2.5
- שיטת טיפול בערכים חריגים – הסרת השורות
- שיטת השלמת ערכים חסרים – השלמה בממוצע
- שיטת נרמול – Min-Max

<u>Random Forest</u>

משתנה מטרה	מקטגוריות המשתנים	RMSE	R^2
COPD	מגדר וגיל, הכנסה, קיפוח, גישה לפארקים, עישון	0.4418	0.7707
COPD	מגדר וגיל, הכנסה, קיפוח, גישה לפארקים, השמנה ,עישון	0.4513	0.7598
AST	אינדיקטור ,קיפוח ,גישה לפארקים ,השמנה ,עישון	0.957	0.495
AST	אינדיקטור ,מגדר וגיל ,הכנסה ,קיפוח ,השמנה	0.969	0.451

<u>XGBoost</u>

משתנה מטרה	מקטגוריות המשתנים	RMSE	R^2
COPD	מגדר וגיל, הכנסה, קיפוח, גישה לפארקים, השמנה	0.3057	0.8951
COPD	מגדר וגיל, הכנסה, קיפוח, גישה לפארקים, השמנה, עישון	0.3112	0.8879
AST	אינדיקטור ,קיפוח ,גישה לפארקים ,השמנה ,עישון	0.6245	0.7585
AST	אינדיקטור ,מגדר וגיל ,קיפוח ,השמנה	0.7597	0.6499

סיכום השינויים ביחס לשלב הקודם :

COPD	שלב קודם		שלב נוכחי	
	RMSE	R^2	RMSE	R^2
Random Forest	0.5533	0.612	0.4418	0.7707
XGBoost	0.5533	0.612	0.3057	0.8951

AST	שלב קודם		שלב נוכחי	
	RMSE	R^2	RMSE	R^2
Random Forest	1.057	0.3503	0.957	0.495
XGBoost	1.057	0.3503	0.6245	0.7585

סיכום שלב שלישי

ניתן לראות שה RMSE-הנמוך ביותר אכן מתאים ל R^2-הגבוה ביותר

המודלים המורכבים (XGBoost ו Random Forest) מציגים שיפור משמעותי בביצועים לעומת מודל רגרסיה לינארית מהשלב השני. במיוחד עבור XGBoost המציג את התוצאות הטובות ביותר עבור שתי המחלות, עם R^2 גבוה מאוד של 0.8951 עבור COPD ו-0.7585 עבור אסתמה.

השיפור הגדול ביותר נצפה במחלת האסתמה (AST) עם מודל הXGBoost , במודל זה ה - R^2 עלה מ-0.3503 במודל של הרגרסיה לינארית ל-0.7585 במודל של XGBoost .

יש לציין כי בהשוואה בין מודל הרגרסיה הלינארית ל XGBoost נמצא כי הרכב המשתנים שגרם לתוצאה הטובה ביותר במודל הרגרסיה הלינארית לא תואם את הרכב המשתנים שגרמו לתוצאה הטובה ביותר ב-XGBoost. זה מדגיש את החשיבות של בחירת המודל המתאים ואת היכולת של מודלים מורכבים יותר לזהות יחסים לא ליניאריים בין המשתנים.

שלב רביעי – הרצת המודל עם 2013

כפי שצוין בפרק השלמת החוסרים, נפנה כעת לבחינת ביצועי המודל עם נתונים משנת 2013. בשלב זה:

1. לא נכלול במודל את האינדיקטורים ומשתני העישון וההשמנה, עקב נתונים חסרים.
2. נתבסס על התובנה שלאורך כל השלבים הקודמים, המודלים שהניבו את תוצאות החיזוי הגבוהות ביותר השתמשו בכל המשתנים הזמינים. לפיכך, נאתחל את המודל עם כל המשתנים הקיימים: מגדר וגיל, הכנסה, קיפוח, גישה לפארקים
3. נבחן את המודל תוך שימוש באותן אופציות שנבדקו בשלב השני

מטרת שלב זה היא לבחון האם הגדלת סט הנתונים יעזרו ביציבות המודל וביכולת לחזות באופן מדויק

אתחול המודל:

- בחרנו בנתונים המתייחסים לשנים 2013-2019 ללא האינדיקטורים, עישון והשמנה
- שיטת נרמול - Min-Max

שלוש ההרצות:

1. טכניקה לזיהוי חריגים IQR עם ערך סף של 2.5, טיפול בערכים חריגים - מחיקה
2. טכניקה לזיהוי חריגים IQR עם ערך סף של 2.5, טיפול בערכים חריגים - החלפה לערך ריק
3. טכניקה לזיהוי חריגים סטיית התקן (STDEV) עם ערך סף של 3, טיפול בערכים חריגים – מחיקה

להלן התוצאות למודל XGBOOST :

משתנה מטרה	שיטת זיהוי חריגים	טיפול בערכים חריגים	ערך סף	RMSE	R^2
COPD	IQR	מחיקה	2.5	0.2834	0.9066
	StdDev	מחיקה	3	0.2873	0.8967
	IQR	החלפה	2.5	0.3105	0.8836
AST	IQR	מחיקה	2.5	0.5616	0.8077
	IQR	החלפה	2.5	0.5824	0.8036
	StdDev	מחיקה	3	0.5642	0.8001

התוצאות מראות כי המודלים מצליחים לחזות COPD טוב יותר מאשר אסתמה, עם ערכי R^2 גבוהים יותר (0.88-0.90 לעומת 0.80-0.81) ו RMSE-נמוכים יותר ברוב המקרים.

עבור COPD :

1. ניתן לראות שינוי קל בין ההרצות, אך לא משמעותי.
2. ההרצה המובילה היא זו שהתחלנו בה בתחילת שלב החקירה: טכניקה לזיהוי חריגים IQR עם ערך סף של 2.5, טיפול בערכים חריגים - מחיקה

עבור אסתמה :

1. השינוי בתוצאות של R^2 הוא מזערי, כמעט שואף לאפס.
2. ניתן לראות שינוי גדול יותר באופן יחסית ב RMSE-בין השיטות השונו , אך גם הוא אינו משמעותי.

שיפור ביחס לשלב הקודם:

	COPD		AST	
	RMSE	R^2	RMSE	R^2
שלב קודם	0.3057	0.8951	0.6245	0.7585
שלב נוכחי	0.2834	0.9066	0.5615	0.8077

סיכום שלב רביעי

הגדלת סט הנתונים לשנת 2013 הביאה לשיפור בשני המודלים. מודל האסתמה הראה את השיפור הגדול ביותר עם עלייה של 5% ב -R^2 וירידה של 0.6 בRMSE. מודל ה-COPD השתפר גם באופן מתון יותר אך עדיין מציג ביצועים טובים יותר. בשני המודלים יש התאמה ברורה בין ערכי הR^2 וRMSE כאשר הערכים הטובים ביותר של שניהם מופיעים יחד. זה מחזק את אמינות התוצאות ומראה שהגדלת מסד הנתונים תרמה לשיפור יכולת החיזוי של שתי המחלות

שלב חמישי - בחינה סופית של המודל המוביל

לאחר שחקרנו לעומק את תוצאות המודלים השונים וניסינו לאתר את המודל שמספק את הניבוי המיטבי לכל אחת מהמחלות הנבדקות, אנו מתקדמים לשלב הבחינה הסופית. בשלב זה, נתמקד במודל המוביל של XGBoost שזוהה בשלבים הקודמים ונבצע הערכה מעמיקה של ביצועיו לניבוי של **COPD**

במקום להשתמש בשיטת ה Cross-Validation-כפי שעשינו בשלבים הקודמים, נבצע הרצה בודדת ומדויקת של המודל. לצורך כך, נשתמש בהיפר-פרמטרים האופטימליים שנשמרו מתוצאות המודל המוביל גישה זו תאפשר לנו לבחון את ביצועי המודל בתנאים אידיאליים ולהעריך את יכולת ההכללה שלו.

מטרות עיקריות בשלב זה:

1. הערכת דיוק: נבחן את מידת הדיוק של המודל בניבוי שכיחות המחלה על סמך המשתנים שנבחרו.
2. בדיקת Overfitting :נבצע ניתוח מעמיק כדי לזהות האם המודל סובל מ- Overfitting כלומר האם הוא מותאם יתר על המידה לנתוני האימון ועלול להתקשות בהכללת נתונים חדשים.
3. ניתוח ביצועים: נערוך ניתוח מקיף של תוצאות המודל, כולל מדדי ביצוע כגון R^2, RMSE ומדדים נוספים שיכולים לספק תובנות לגבי איכות החיזוי.
4. השוואה לשלבים קודמים: נשווה את התוצאות של ההרצה הבודדת לתוצאות שהתקבלו בשלבים הקודמים עם Cross-Validation כדי להעריך את עקביות הביצועים.
5. זיהוי נקודות חוזק וחולשה: ננתח את הביצועים של המודל בהקשר של משתנים ספציפיים או תת-קבוצות של הנתונים, כדי לזהות היכן המודל מצליח במיוחד והיכן הוא עשוי להזדקק לשיפור.

באמצעות בחינה מעמיקה זו, נוכל להעריך את האפקטיביות הכוללת של המודל שפיתחנו ולקבל החלטות מושכלות לגבי יישומו בפועל או הצורך בשיפורים נוספים.

להלן התוצאות:

מדד	ביצועי מבחן	ביצועי אימון
RMSE	0.2937	0.1390
R^2	0.9008	0.9777
StdDev	0.8253	0.8629

בחינת התוצאות מראה כי הגענו לביצועים דומים מאוד לאלו שהושגו בשלב הרביעי, מה שמעיד כי אכן מצאנו את ההיפר-פרמטרים האופטימליים למודל הנוכחי. כמו כן ניתן לראות כי הסטית התקן של תוצאות המבחן גדולה משמעותית מ RMSE מה שמעיד על מודל חזק.

עם זאת, ניכר פער משמעותי במדדי הביצוע בין נתוני האימון לנתוני המבחן. נשים לב כי ערך ה-R^2 בנתוני האימון (0.9777) גבוה משמעותית מערכו בנתוני המבחן (0.9008). כמו כן ,ערך ה- RMSE בנתוני האימון (0.1390) נמוך באופן ניכר מה- RMSE בנתוני המבחן (0.2937). פערים אלו מצביעים על כך שהמודל מתחיל לשנן את דוגמאות האימון ונכנס למצב של התאמת יתר(Overfitting)

התמודדות עם התאמת יתר (Overfitting)

כדי להתמודד עם בעיית התאמת יתר נבצע חלוקה מחודשת של הנתונים אך הפעם לשלושה חלקים - אימון, מבחן, וולידציה . חלוקה זו תאפשר לנו להעריך את ביצועי המודל על נתונים שלא נחשף אליהם כלל בתהליך האימון והוולידציה.

להלן התוצאות העדכניות:

מדד	אימון	וולידציה	מבחן
RMSE	0.1336	0.3155	0.3116
R^2	0.9796	0.8834	0.8885
STD	0.8872	0.8100	0.8210

למרות המאמצים הראשוניים לטפל בOverfitting באמצעות חלוקה מחודשת של הנתונים, התוצאות מראות כי הבעיה עדיין קיימת במידה מסוימת. הפער המשמעותי בין ביצועי האימון לביצועי הוולידציה והמבחן, במיוחד במדדי RMSE ו- R^2 מעיד על כך שהמודל עדיין מתקשה להכליל היטב.

לפיכך, נמשיך לטפל בOverfitting באמצעות טכניקות מתקדמות נוספות:

1. Grid Search: נבצע חיפוש מקיף על פני ההיפר-פרמטרים לאיתור הקומבינציה האופטימלית.
2. היפר-פרמטרים מותאמים: נבחן 8 היפר-פרמטרים שונים המתמקדים בהגבלת גודל העץ ומניעת Overfitting. פרמטרים אלו כוללים ח לק מהאלמנטים שהוצגו בפרק השיטות.
3. Cross-validation: נשתמש בשיטת K-fold עם K=5 שתאפשר הערכה מדויקת יותר של ביצועי המודל.

לכל היפר-פרמטר נבחן 3 ערכים שונים, מה שמוביל ל-3^8 = 6,561 קומבינציות אפשריות.
בשילוב עם ה -Cross-Validation (K=5) המודל ירוץ בסך הכול 5 * 6,561 = 32,805 פעמים.

תהליך זה, אף שהוא אינטנסיבי מבחינה חישובית, יאפשר לנו לזהות את התצורה האופטימלית של המודל, תוך התמודדות יסודית עם אתגר התאמת יתר ושיפור משמעותי ביכולת ההכללה. התוצאות של תהליך זה יספקו לנו תובנות מעמיקות לגבי הביצועים האמיתיים של המודל ויאפשרו בחירה מושכלת של ההיפר-פרמטרים המתאימים ביותר למשימה הנוכחית.

להלן התוצאות:

מדד	אימון	ולידציה	מבחן
R^2	0.9420	0.8567	0.8616
RMSE	0.2249	0.3498	0.3470
StdDev	0.8580	0.8040	0.8137

ניתוח התוצאות:

1. **פער ב-R^2:** קיים פער בין ביצועי האימון (0.94202) לבין ביצועי הוולידציה (0.856718) והמבחן (0.861672). זהו פער של כ-8%, המצביע על מידה מסוימת של התאמת יתר, אך לא חמורה במיוחד
2. **RMSE:** ערכי ה-RMSE גבוהים יותר בוולידציה (0.349826) ובמבחן (0.347068) לעומת האימון (0.224981). זה מחזק את הסימן להתאמת יתר מסוימת
3. **עקביות בין ולידציה למבחן** :הערכים בין סט הוולידציה לסט המבחן דומים מאוד, מה שמעיד על יציבות טובה של המודל
4. **שיפור לעומת התוצאות הקודמות** :בהשוואה לתוצאות שהוצגו קודם לכן, נראה שיש שיפור בהתמודדות עם התאמת יתר, אם כי הבעיה לא נפתרה לחלוטין.

לסיכום, המודל עדיין מראה סימנים מסוימים של התאמת יתר, אך במידה מתונה יחסית. הפער בין ביצועי האימון לביצועי הוולידציה והמבחן קטן יותר מאשר בגרסאות קודמות של המודל, מה שמעיד על שיפור ביכולת ההכללה. עם זאת, ייתכן שיש מקום לשיפור נוסף בכיוון זה.

<u>סיכום שלב חמישי</u>

בשלב זה, התמקדנו בהערכה מעמיקה של המודל המוביל שזוהה בשלבים הקודמים. מטרתנו הייתה לבחון את ביצועי המודל בתנאים אידיאליים ולהעריך את יכולת ההכללה שלו.

תחילה, ביצענו הרצה בודדת ומדויקת של המודל עם ההיפר-פרמטרים האופטימליים שנשמרו מהשלבים הקודמים. בהמשך, התמודדנו עם אתגר הOverfitting באמצעות מספר טכניקות:

1. חלוקה מחודשת של הנתונים לסטים של אימון, ולידציה ומבחן.
2. יישום שילוב של Grid Search, Cross-Validation, והתאמת היפר-פרמטרים לצמצום Overfitting.

התוצאות הראו שיפור בהתמודדות עם Overfitting בהשוואה לגרסאות הקודמות של המודל. הצלחנו לשחזר ואף לשפר את התוצאות שהושגו בשלבים הקודמים, תוך צמצום הפער בין ביצועי האימון לביצועי הוולידציה והמבחן.

למרות המאמצים הרבים, נראה כי הגענו למיצוי של הטכניקות הנוכחיות בשלב החמישי והאחרון של תהליך החקירה. אף על פי שהשגנו שיפור משמעותי, עדיין יש מקום לשיפור נוסף. הפתרון שעשוי לסייע בהמשך הוא הוספת דגימות למסד הנתונים. הגדלת מספר הדוגמאות יכולה לספק למודל מידע נוסף ומגוון יותר, מה שעשוי לשפר את יכולת ההכללה שלו ולהפחית עוד יותר את התאמת היתר.

לסיכום, השלב החמישי והאחרון בתהליך החקירה הציג התקדמות משמעותית בביצועי המודל ובהתמודדות עם Overfitting . עם זאת, הוא גם חשף את הצורך בהמשך מחקר ושיפור לקראת יישום מעשי של המודל. הרחבת מסד הנתונים מהווה כיוון מבטיח להמשך פיתוח ושיפור המודל בעתיד, מעבר לתהליך החקירה הנוכחי.

סיכום פרק חקירה

החקירה שלנו התפתחה דרך חמישה שלבים מרכזיים, כל אחד העמיק את הבנתנו ושיפר את מודל החיזוי. גישה מדורגת זו אפשרה לנו לבחון בקפידה כל היבט של הנתונים והמודלים.

התחלנו בבחינת השפעת הרכב המשתנים, תהליך שחשף את חשיבותם היחסית של גורמים שונים בחיזוי מחלות נשימה כרוניות. בהמשך, בדקנו שיטות שונות לטיפול בערכים חריגים, שלב קריטי בהבטחת איכות הנתונים ובהבנת השפעת ערכים קיצוניים על המודלים. גילינו כי בחירת המשתנים הנכונים הייתה משמעותית יותר מאשר שיטת הטיפול בחריגים.

המעבר למודלים מתקדמים כמו יערות רנדומיים ו XGBoost-היווה נקודת מפנה בחקירה. שלב זה הדגים את יתרונות המודלים המורכבים בתפיסת קשרים לא-ליניאריים, במיוחד עבור אסתמה.

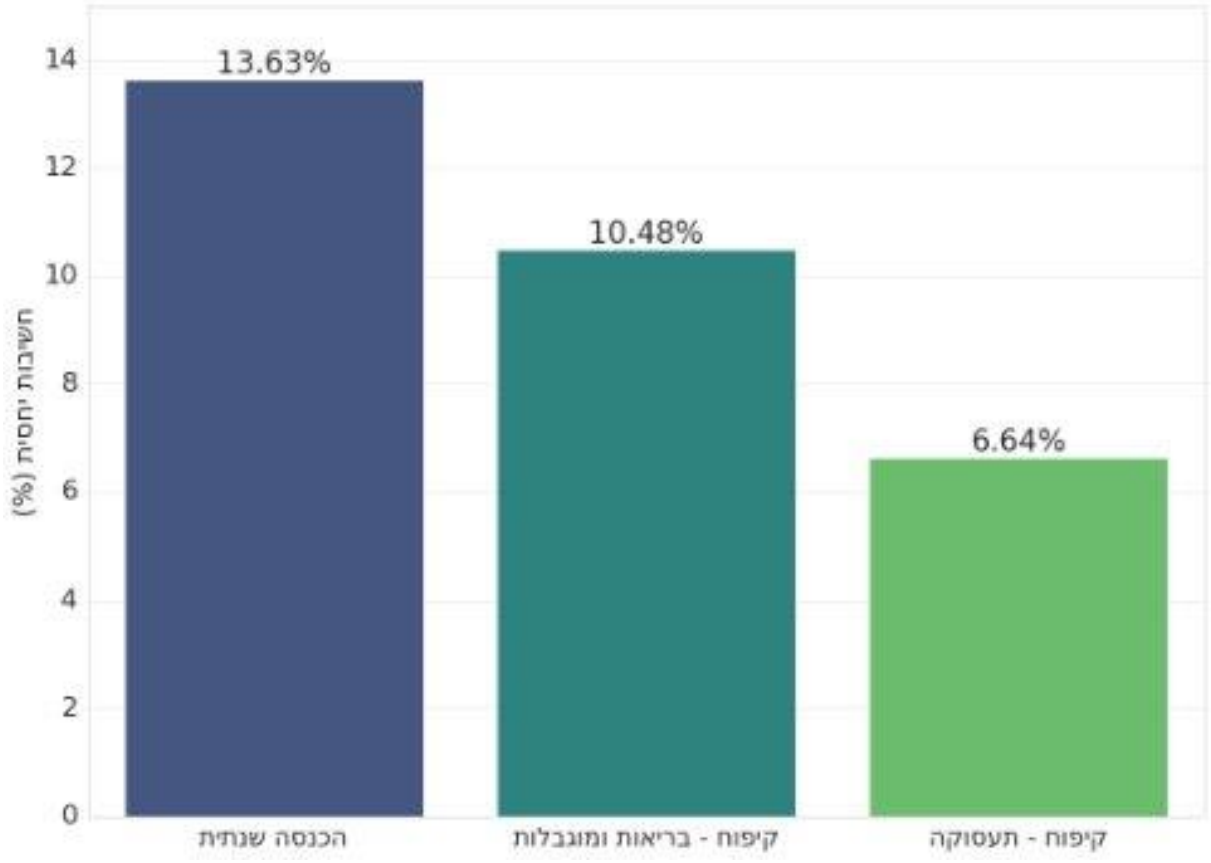
בניסיון להגדיל את מספר הדגימות, הרחבנו את בסיס הנתונים לכלול את שנת 2013. אולם, צעד זה דרש ויתור על משתנים חשובים כמו אינדיקטורים, השמנה ועישון. למרות צמצום העמודות, שלב זה הדגיש את החשיבות של איזון בין כמות הנתונים לאיכותם בבניית מודלים אמינים.

לבסוף, התמקדנו במודל המוביל של COPD וניסינו לצמצם את ה Overfitting- באמצעות מניפולציות על המודל עצמו, הצלחנו לשפר במעט את הביצועים. תהליך זה הוביל אותנו לתובנה כי סט נתונים גדול יותר היה עשוי לסייע בצמצום משמעותי יותר של הOverfitting

ניתוח תוצאות

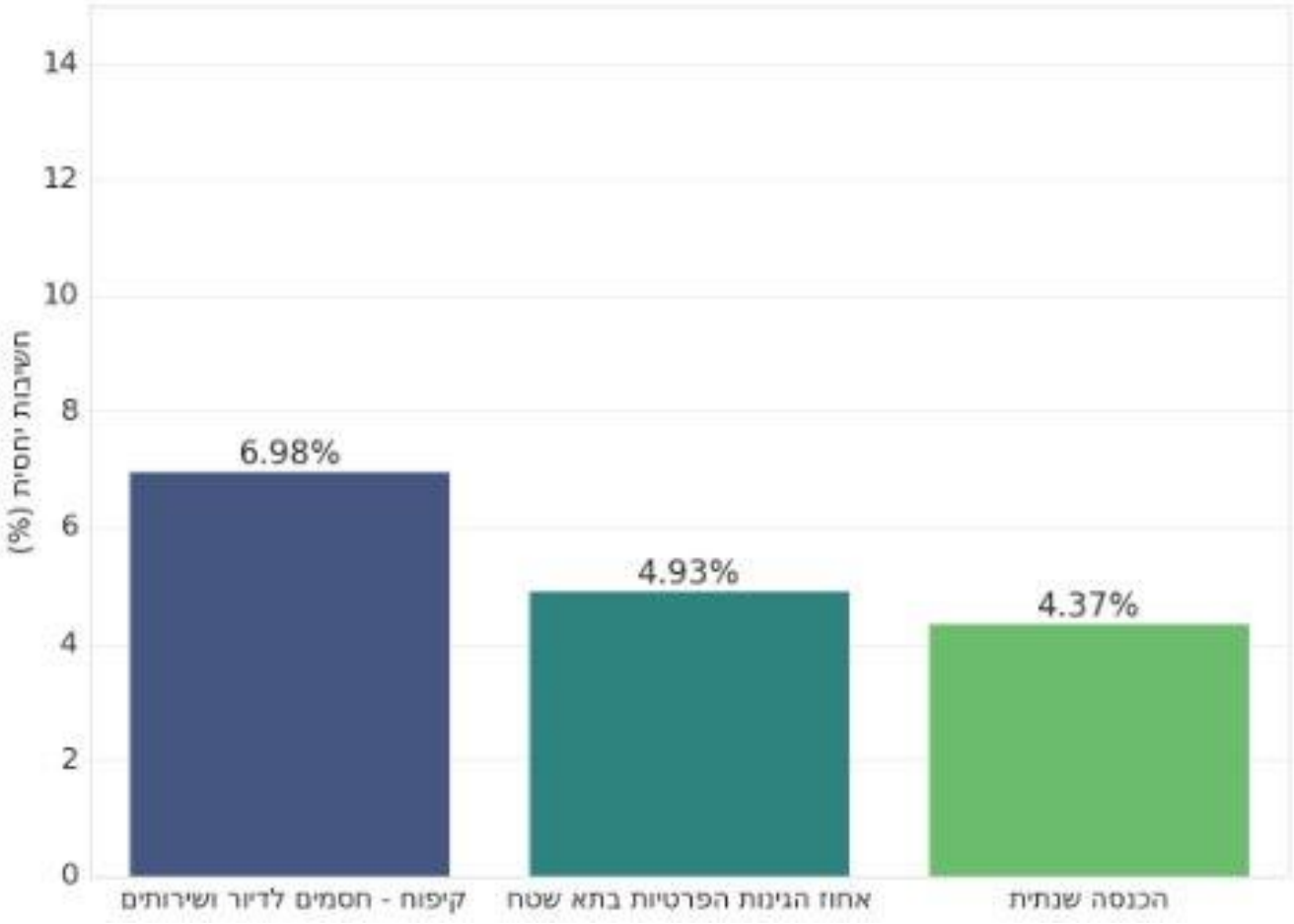
בפרק זה נציג את הממצאים העיקריים שעלו מתהליך החקירה והניתוח של הנתונים. ממצאים אלו מספקים תובנות חשובות לגבי הגורמים המשפיעים על שכיחות מחלות הנשימה שנחקרו, ומדגישים את המורכבות של הקשרים בין המשתנים השונים לבין התוצאות הבריאותיות. לצורך זה, נשתמש בניתוח חשיבות המשתנים (feature importance), כפי שהוצג בפרק השיטות. ניתוח זה מאפשר לנו לזהות את המשתנים בעלי ההשפעה הגדולה ביותר על תוצאות המודל, ובכך לספק תובנות מעמיקות על הגורמים המרכזיים המשפיעים על שכיחות מחלות הנשימה באוכלוסייה

חשיבות יחסית של תכונות למודל חיזוי ה COPD



ניתן לראות כי שלושת המשתנים המובילים הם גורמים סוציו-אקונומיים, המשקפים היבטים שונים של מצב כלכלי וחברתי. יחד, שלושת המדדים הללו מהווים 30.75% מיכולת הניבוי של המודל, מה שמדגיש את החשיבות המכרעת של גורמים חברתיים-כלכליים בהקשר של מחלת COPD.

חשיבות יחסית של תכונות למודל חיזוי אסטמה



שלושת המדדים הללו מהווים 16.28% מיכולת הניבוי של המודל, מה שמדגיש את החשיבות של גורמים סביבתיים כלכליים וחברתיים בהקשר של מחלת אסתמה.

<u>סיכום ניתוח התוצאות</u>

בשתי המחלות, המשתנים המובילים מדגישים את החשיבות של גורמים חברתיים, כלכליים וסביבתיים בהקשר של בריאות הנשימה. זה מחזק את ההבנה כי מחלות נשימה מושפעות באופן משמעותי מתנאי החיים הכוללים של האדם, ולא רק מגורמים רפואיים ישירים.

במקרה של COPD, המשתנים המובילים מראים השפעה חזקה יותר על המודל. הכנסה שנתית, קיפוח בבריאות ומוגבלות, וקיפוח בתעסוקה מהווים יחד 30.75% מיכולת הניבוי של המודל. זה מצביע על קשר הדוק יותר בין גורמים סוציו-אקונומיים לבין הסיכון ל-COPD. לעומת זאת, במודל לאסתמה, ההשפעה של המשתנים המובילים מפוזרת יותר. קיפוח בנגישות לדיור ושירותים, אחוז הגינה הפרטית, והכנסה שנתית מהווים יחד 16.28% מיכולת הניבוי. זה מרמז על מורכבות גדולה יותר בגורמים המשפיעים על אסתמה, עם דגש על היבטים סביבתיים לצד כלכליים.

ההבדלים בחיזוי בין COPD ל AST

במהלך המחקר שלנו, בחנו את היכולת לחזות את שכיחות מחלת ריאות חסימתית כרונית (COPD) ואסתמה (AST) באמצעות מגוון מודלים סטטיסטיים ומודלים של למידת מכונה. תהליך החקירה חשף הבדלים ביכולת החיזוי של שתי המחלות, כאשר COPD הציגה באופן עקבי תוצאות טובות יותר מבחינת המדדים **R^2** ו**RMSE** לאורך רוב שלבי המחקר.

בשלב הראשוני של המחקר, התמקדנו במודל רגרסיה לינארית. מתוך 127 קומבינציות שונות של משתנים שנבדקו עבור כל מחלה, מצאנו כי 120 מודלים עבור COPD הציגו ביצועים טובים יותר בהשוואה לאסתמה, כאשר הרכב המשתנים היה זהה. זה מהווה עדות ליכולת הניבוי העדיפה של COPD בשלבים המוקדמים של המחקר.

עם זאת, חשוב לציין כי לאורך שלבי המחקר, ובמיוחד במעבר ממודלים לינאריים למודלים מורכבים יותר, ראינו שיפור גדול יותר בחיזוי של אסתמה בהשוואה ל-COPD, זה מרמז על כך שהקשרים בין המשתנים המסבירים לבין שכיחות האסתמה מורכבים יותר ודורשים מודלים מתוחכמים יותר כדי לתפוס אותם באופן מדויק.

אחד הממצאים המשמעותיים במחקרנו היה ההבדל בקורלציות ובחשיבות המאפיינים בין אסתמה ו-COPD. ניתוח מעמיק של הקורלציות בין המשתנים השונים לבין שכיחות המחלות חשף כי ערכי הקורלציות של COPD עם משתנים אחרים היו גבוהים במעט מאלו של אסתמה. עם זאת, הקורלציות עבור שתי המחלות היו נמוכות משמעותית, כאשר כל הקורלציות בין המשתנים לבין אסתמה ו-COPD היו נמוכות מ-0.6.

לסיכומו של דבר, ההבדלים בחיזוי בין COPD ל AST-מדגישים את המורכבות והייחודיות של כל מחלה. הם מצביעים על כיווני מחקר עתידיים, כגון זיהוי משתנים נוספים שעשויים לשפר את החיזוי של אסתמה, או חקירה מעמיקה יותר של הגורמים המשפיעים על התפתחות כל אחת מהמחלות. יתר על כן, הם מדגישים את החשיבות של שימוש במגוון שיטות לבניית מודל ואנליזה בעת חקירת תופעות רפואיות מורכבות.

סיכום ומסקנות

סיכום המחקר

מחקר זה התמקד בחקירת הגורמים המשפיעים על שכיחות מחלות נשימה כרוניות, ובפרט מחלת ריאות חסימתית כרונית (COPD) ואסתמה, תוך שימוש בשיטות מתקדמות של ניתוח נתונים ולמידת מכונה. המחקר התבסס על מסד נתונים מקיף הכולל מגוון רחב של משתנים חברתיים-כלכליים, סביבתיים ובריאותיים מהשנים 2013-2019.

תהליך המחקר כלל מספר שלבים מרכזיים. ראשית, בוצע טיפול מקיף בנתונים, כולל התמודדות עם ערכים חסרים וחריגים. שיטות שונות נבחנו לטיפול בנתונים אלו, כגון השלמת ערכים חסרים באמצעות ממוצע או חציון, וזיהוי וטיפול בערכים חריגים באמצעות שיטות כמו IQR ו .Standard Deviation-שלב זה היה קריטי להבטחת איכות הנתונים ואמינות הניתוחים הסטטיסטיים שבאו בעקבותיו.

לאחר הכנת הנתונים, יושמו מספר מודלים סטטיסטיים ומודלים מתקדמים של למידת מכונה. התחלנו עם רגרסיה לינארית כבסיס, ובהמשך התקדמנו למודלים מורכבים יותר כמו Random Forest ו .XGBoost-תהליך זה אפשר לנו להשוות בין ביצועי המודלים השונים ולזהות את המודל המתאים ביותר לחיזוי שכיחות המחלות.

ממצאי המחקר העיקריים הראו כי מודל ה XGBoost-הציג את הביצועים הטובים ביותר בחיזוי שכיחות הן של COPD והן של אסתמה. עם זאת, נצפו הבדלים משמעותיים ביכולת החיזוי בין שתי המחלות, כאשר COPD הציגה תוצאות טובות יותר באופן עקבי לאורך כל שלבי המחקר.

ניתוח חשיבות התכונות (feature importance) חשף תובנות מעניינות לגבי הגורמים המשפיעים ביותר על שכיחות כל מחלה. עבור COPD גורמים חברתיים-כלכליים כמו הכנסה וקיפוח בבריאות נמצאו כמשפיעים ביותר. לעומת זאת, עבור אסתמה, גורמים סביבתיים כמו נגישות לדיור ושטחים ירוקים הראו השפעה גבוהה.

במהלך המחקר, התמודדנו עם אתגרים מתודולוגיים שונים, במיוחד בהקשר של Overfitting, לשם כך יישמנו טכניקות מתקדמות כמו Cross-validation ו Grid Search-לבחירת היפר-פרמטרים אופטימליים למודלים. כמו כן, השתמשנו בטכניקות נרמול ובחירת תכונות לשיפור ביצועי המודלים.

לסיכום, מחקר זה תרם להבנה מעמיקה יותר של הגורמים המשפיעים על שכיחות מחלות נשימה כרוניות, תוך הדגשת ההבדלים בין COPD לאסתמה. הממצאים מדגישים את החשיבות של גורמים חברתיים-כלכליים וסביבתיים בהקשר של בריאות הנשימה, ומספקים בסיס למחקרים עתידיים ולפיתוח מדיניות בריאות ציבורית מבוססת ראיות.

מסקנות

מחקר זה התמקד במטרה מרכזית אחת: לבחון את היכולת לחזות את שכיחותן של מחלות נשימה כרוניות על בסיס משתנים חברתיים-כלכליים וסביבתיים. שאלת המחקר המרכזית שהנחתה אותנו הייתה" :**האם וכיצד ניתן לנבא את רמת התחלואה של מחלות נשימה כרוניות על בסיס משתנים חברתיים "?** במהלך המחקר, בחנו גם את ההבדלים בגורמי הסיכון וביכולת החיזוי בין COPD לאסתמה. הממצאים שלנו מספקים תובנות משמעותיות בנוגע לשאלת מחקר זו. ניתן לנבא ברמה גבוהה את רמת התחלואה במחלות שנבחנו על ידי משתנים חברתיים ודמוגרפיים, מה שמספק מענה חיובי לשאלת המחקר המרכזי.

1. **השפעת גורמים חברתיים-כלכליים:** משתני הקיפוח וההכנסה נמצאו כמשפיעים העיקריים על רמת התחלואה במחלות נשימה כרוניות. ממצא זה מדגיש את החשיבות של התנאים החברתיים-כלכליים בקביעת בריאות הנשימה של האוכלוסייה, ומצביע על הצורך בהתייחסות לגורמים אלו בתכנון מדיניות בריאות ציבורית.
2. **הבדלים בין COPD לאסתמה:** ביצועי החיזוי של COPD עלו באופן עקבי על אלו של אסתמה, במיוחד כאשר השתמשנו במודלים לינאריים. הבדל זה מרמז על מורכבות גדולה יותר בגורמים המשפיעים על אסתמה, ומדגיש את הצורך בגישות שונות לחקר וטיפול בשתי המחלות.
3. **יעילות מודלים מתקדמים:** שימוש במודלים מתקדמים כמו XGBoost הוביל לשיפור משמעותי בחיזוי רמת התחלואה של אסתמה. ממצא זה מדגיש את החשיבות של שימוש בטכניקות למידת מכונה מתקדמות בניתוח נתוני בריאות מורכבים, ומצביע על הפוטנציאל של גישות אלו בשיפור הבנתנו וחיזוי מחלות נשימה.
4. **עמידות המודל לשיטות טיפול בנתונים:** העובדה שלא נצפתה השפעה מהותית של שיטות שונות לטיפול בערכים חריגים על ביצועי המודל מצביעה על עמידות המודל ויציבות הקשרים שזוהו. זה מחזק את אמינות הממצאים ומדגיש את חוסנם של הגורמים המשפיעים שזוהו, ללא תלות בשיטת הטיפול בנתונים.
5. **השלכות על מדיניות בריאות הציבור:** הממצאים מדגישים את הצורך בגישה רב-ממדית לטיפול במחלות נשימה כרוניות. ההשפעה המשמעותית של גורמים חברתיים-כלכליים וסביבתיים מצביעה על כך שמדיניות בריאות אפקטיבית צריכה לכלול לא רק התערבויות רפואיות ישירות, אלא גם מאמצים לשיפור תנאי המחיה, הפחתת אי-שוויון כלכלי, ושיפור איכות הסביבה. זה מחזק את הצורך בשיתוף פעולה בין-תחומי בין מערכת הבריאות, קובעי מדיניות חברתית, ומתכנני ערים בפיתוח אסטרטגיות למניעה וטיפול במחלות נשימה כרוניות.

מסקנות אלו מספקות תובנות חשובות להבנת הגורמים המשפיעים על מחלות נשימה כרוניות ומציעות כיוונים מחקר עתידיים, כמו גם השלכות פוטנציאליות על מדיניות בריאות הציבור. הן מדגישות את המורכבות של הנושא ואת הצורך בגישה מקיפה לטיפול באתגרים הבריאותיים הללו.

נספחים

Pipeline:

שם הפונקציה: SelectFeatures

רציונל: הפונקציה נועדה לאפשר בחירה גמישה של משתנים מסבירים (תכונות) מתוך סט הנתונים, על בסיס קטגוריות של תכונות. זה מאפשר לבחור תת-קבוצה של התכונות הרלוונטיות למודל, על פי צורכי הניתוח או הבעיה הספציפית. (חלוקת הקטגוריות נמצאת בנספים)

קלט: DataFrame, שם מחלה (AST או COPD), רשימה של קטגוריות תכונות

פלט: DataFrame עם העמודות הרלוונטיות לתכונות שנבחרו.

חלוקת הקטגוריות של המשתנים :

חלוקה זו נעשתה על מנת להקל על הבנת סוגי הנתונים ולזהות את ההקשרים ביניהם.
כאשר הפיצ'רים מקובצים לקטגוריות ניתן לבחור באופן מדויק יותר את הפיצ'רים הרלוונטיים לכל משימה או מודל חיזוי ספציפי שנרצה. בחירת פיצ'רים רלוונטיים בלבד משפרת את דיוק המודלים ומקטינה את הסיכוי לרעש או נתונים לא רלוונטיים שיכולים להשפיע על תוצאות החיזוי.

feature_categories =

<u>Gender:</u>

description: Demographic data based on gender

columns: Total_All, MALE_All, FEMALE_All, Male_0_4, Male_5_9, Male_10_14, Male_15_19, Male_20_24, Male_25_29, Male_30_34, Male_35_39, Male_40_44, Male_45_49, Male_50_54, Male_55_59, Male_60_64, Male_65_69, Male_70_74, Male_75_79, Male_80_84, Male_85+, Female_0_4, Female_5_9, Female_10_14, Female_15_19, Female_20_24, Female_25_29, Female_30_34, Female_35_39, Female_40_44, Female_45_49, Female_50_54, Female_55_59, Female_60_64, Female_65_69, Female_70_74, Female_75_79, Female_80_84, Female_85+

<u>Asthma</u>

description: Asthma-related data

columns: AST002(percent), AST003(percent), AST004(percent)

<u>COPD</u>

description: Chronic Obstructive Pulmonary Disease (COPD) related data

columns : COPD002(percent), COPD003(percent), COPD004(percent), COPD005(percent), COPD007(percent)

<u>Income:</u>

description: Economic data related to income

columns: Weighted_Total_annual_income_(£), Weighted_Net_annual_income_(£), Weighted_Net_annual_income_before_housing_costs_(£), Weighted_Net_annual_income_after_housing_costs_(£)

<u>Deprivation:</u>

description: Data related to deprivation index

columns: Weighted_Index_of_Multiple_Deprivation_(IMD), Weighted_Income, Weighted_Employment, Weighted_Education,_Skills_and_Training, Weighted_Health_Deprivation_and_Disability, Weighted_Crime, Weighted_Barriers_to_Housing_and_Services, Weighted_Living_Environment

Parks_and_Housing:

description: Data related to parks and housing

columns: Weighted_ParksOnly_Average_distance_to_nearest_Park_or_Public_Garden_(m), Weighted_ParksOnly_Average_size_of_nearest_Park_or_Public_Garden_(m2), Weighted_ParksOnly_Average_number_of_Parks_or_Public_Gardens_within_1,000_m_radius, Weighted_ParksOnly_Number_of_built_up_area_postcodes__within_300m_of_a_Park_or_Public_Garden_(percentage), Weighted_ParksOnly_Number_of_built_up_area_postcodes__within_900m_of_a_Park_or_Public_Garden_(percentage), Weighted_ParksAndPlayingFields_Average_distance_to_nearest_Park_Public_Garden_or_Playing_Field_(m), Weighted_ParksAndPlayingFields_Average_size_of_nearest_Park_Public_Garden_or_Playing_Field_(m2), Weighted_ParksAndPlayingFields_Average_number_of__Parks_Public_Gardens_or_Playing_Fields_within_1,000_m_radius, Weighted_ParksAndPlayingFields_Number_of_built_up_area_postcodes__within_300m_of_a_Park_Public_Garden_or_Playing_Field_(percentage), Weighted_ParksAndPlayingFields_Number_of_built_up_area_postcodes__within_900m_of_a_Park_Public_Garden_or_Playing_Field_(percentage), Weighted_Houses_Percentage_of_adresses_with_private_outdoor_space, Weighted_Houses_Average_size_of_private_outdoor_space_(m2), Weighted_Houses_Median_size_of_private_outdoor_space_(m2), Weighted_Flats_Percentage_of_adresses_with_private_outdoor_space, Weighted_Flats_Average_size_of_private_outdoor_space_(m2), Weighted_Flats_Average_number_of_flats_sharing_a_garden, Weighted_Total_Percentage_of_adresses_with_private_outdoor_space

Obesity:

description: Obesity-related data

columns: Obesity_Prevalence(per_cent)

Smoking:

description: Smoking-related data

columns: SMOK001_Underlying_achievement_net_of_exceptions(per_cent), SMOK002_Underlying_achievement_net_of_exceptions(per_cent), SMOK004_Underlying_achievement_net_of_exceptions(per_cent), SMOK005_Underlying_achievement_net_of_exceptions(per_cent)

שם הפונקציה: OutliersRemoveOrReplace

רציונל: הפונקציה נועדה לנקות נקודות קיצון מתוך DataFrame באמצעות זיהוי נקודות קיצון בעזרת שתי שיטות אפשריות - IQR או StdDev ולטפל בנקודות קיצון אלו על ידי הסרתן או החלפתן בערך ריק בהתאם לפעולה שנבחרה. זה מאפשר להתמודד עם נתונים חריגים שעלולים להשפיע לרעה על תוצאות הניתוח או המודל.

קלט: DataFrame, השיטה לזיהוי נק' קיצון , ערך הסף ופעולה לביצוע על נק' קיצון

פלט: DataFrame עם הטיפול בנקודות קיצון (נקי מנקודות אלו או נק' קיצון שהוחלפו בערך NULL)

שם הפונקציה: FillMissingValues

רציונל: הפונקציה נועדה למלא ערכים חסרים ב-DataFrame באמצעות שלוש שיטות אפשריות: ממוצע, חציון או הסרת שורות עם ערכים חסרים. זה מאפשר לשמור על שלמות הנתונים ולהימנע מהטיות במודלים ובניתוחים סטטיסטיים שנבצע.

קלט: DataFrame, שיטת טיפול בערכים חסרים

פלט: DataFrame לאחר הטיפול בערכים החסרים

שם הפונקציה:FilterByYears

רציונל: אנו רוצים לבדוק את ההשפעה של תקופה מסוימת על המודל והאם הוא מראה שיפור עם השנים, ולכן יצרנו פונקציה המאפשרת סינון לפי שנים ספציפיות.

קלט: DataFrame, רשימה של שנים לסינון או שנה בודדת.

פלט: DataFrame מסונן עם רק השורות שעונות על השנים שהמשתמש ביקש.

שם הפונקציה:NormalizeData

רציונל: פונקציה זו נועדה לנרמל את הנתונים בדאטה פריים על פי שיטת נרמול שנבחרת, כדי להכין את הנתונים לשלב הלמידה של המודלים.

קלט: DataFrame ושיטת נרמול לביצוע

פלט: DataFrame עם הנתונים המנורמלים לפי השיטה שנבחרה.

שם הפונקציה: SplitDataAndPrepForModel

רציונל: הפונקציה נועדה לחלק את מסד הנתונים לקבוצות אימון (train) ובדיקה (test) לצורך בנייה והערכה של מודלים בלמידת מכונה.

קלט: DataFrame ושיעור הנתונים שיוקצו לקבוצת הבדיקה (test_size), ברירת המחדל היא 20%.

פלט: ארבעה מערכים המכילים את הנתונים המחולקים לקבוצות אימון ובדיקה : X_train, X_test, y_train, y_test.

שם הפונקציה:TrainAndEvaluateModel

רציונל: פונקציה זו נועדה לאמן ולהעריך מודלים שונים על סט נתונים מחולק לאימון ובדיקה לרגרסיה לינארית, יערות רנדומים ו-XGBoost.

קלט: X_train, X_test, y_train, y_test, שם המודל,מס' ניסוי ורשימת מילונים עם פרמטרים למודלים

פלט: רשימה עם תוצאות ביצועים של המודלים שהורצו.

שם הפונקציה:MakeFinalPanel

רציונל: הפונקציה נועדה להכין את מסד הנתונים לשלב הבא של בניית מודל לומד מכונה או ניתוח סטטיסטי. היא מפעילה בתוכה את כל הפונקציות הקודמות לטובת יצירת פאנל הנתונים. דרוש כתיבה מחדש

קלט:

- panel_path: נתיב לקובץ הפאנל המקורי.
- disease: מחלה או תופעה מסוימת לניתוח.
- feature_types: סוגי המאפיינים המבוקשים לכלול בפאנל הסופי.
- method_outliers: שיטת התמודדות עם ערכים חריגים.
- threshold: אחוזים או ערך מספרי המשמשים בשיטת הערכת ערכים חריגים.
- Outliersaction: פעולה לביצוע על ערכים חריגים (הסרה או החלפה).
- method_missing_values: שיטת טיפול בערכים חסרים.
- normalize_method: שיטת נורמליזציה להחלפת הערכים החסרים.
- years: שנים או טווח שנים לסינון הנתונים.

פלט: מסד נתונים סופי (DataFrame) שעבר את כל שלבי העיבוד והוכן לשלב הבא של בניית המודל.

שם הפונקציה:FinalFunction

רציונל: הפונקציה הזו מהווה את הממשק העליון לתהליך בניית המודל. היא מאפשרת להריץ את התהליך מספר פעמים עם פרמטרים שונים, ולקבל תוצאות משולבות עבור כל הריצות. בנוסף הפונקציה כותבת לקובץ את כל התוצאות של הריצות.

קלט:

- נתיב לקובץ אקסל עם הפרמטרים לריצה הנוכחית
- נתיב לקובץ המכיל את פאנל הנתונים

פלט: DataFrame המכיל את כל התוצאות של הריצות השונות, משולב עם הפרמטרים המקוריים.

פלט קורלציה

Top 10 features most correlated with COPD_Prevalence:

1 .Weighted_Net_annual_income_(£) 0.518942

2 .Weighted_Total_annual_income_(£) 0.511881

3 .Weighted_Health_Deprivation_and_Disability 0.490271

4 .Obesity_Prevalence(per_cent) 0.484711

5 .AST_Prevalence 0.465912

6 .Weighted_Education,*Skills_and_Training 0.464321

7 .Weighted_Employment 0.455885

8 .Weighted_Net_annual_income_before_housing_costs*(£) 0.441039

9 .Weighted_Barriers_to_Housing_and_Services 0.436192

10 .Weighted_Net_annual_income_after_housing_costs_(£) 0.349247

Top 10 features most correlated with AST_Prevalence:

1 .COPD_Prevalence 0.465912

2 .Weighted_Total_Percentage_of_adresses_with_private_outdoor_space 0.388399

3 .Weighted_Net_annual_income_(£) 0.344489

4 .Obesity_Prevalence(per_cent) 0.343648

5 .Weighted_Barriers_to_Housing_and_Services 0.340561

6 .Weighted_Total_annual_income_(£) 0.321165

7.Weighted_ParksOnly_Average_number_of_Parks_or_Public_Gardens_within_1,000_m_radius 0.321082

8 .Weighted_Net_annual_income_before_housing_costs_(£) 0.311499

9 .
Weighted_ParksOnly_Number_of_built_up_area_postcodes__within_900m_of_a_Park_or_Public_Garden_(percentage) 0.299220

10 .Male_70_74 0.281389

פרק חמישי

היפר-פרמטרים עיקריים של המודל הטוב ביותר: COPD

- max_depth: 7
- min_child_weight: 3
- learning_rate: 0.1
- n_estimators: 200
- colsample_bytree: 1.0
- subsample: 0.8
- reg_alpha: 0.5
- reg_lambda: 1