אמידת האוכלוסייה החרדית בישראל באמצעות למידת מכונה

אביעד קלינגר

נובמבר 2020

תוכן עניינים

1. מבוא 3

2. רקע תיאורטי 4

3. מתודולוגיה 7

3.1. מקור הנתונים 7

3.2. ניקוי ושיפור הנתונים 7

3.3. ניתוח נתונים 10

3.4. בחירת ואימון האלגוריתם 13

3.5. הערכת האלגוריתם 18

3.6. פירוש האלגוריתם 20

4. תוצאות 23

5. סיכום 27

6. ביבליוגרפיה 28

1. מבוא

האוכלוסייה החרדית היא קבוצה בעלת מאפיינים גיאו-דמוגרפיים וחברתיים-כלכליים ייחודיים בחברה הישראלית. מסיבה זו היא מסקרנת ומעניינת קובעי מדיניות, חוקרים ואת הציבור הרחב כבר שנים ארוכות. חלקה ההולך וגדל באוכלוסייה רק מעצים את הצורך באמידתה בצורה נרחבת ומפורטת עד כמה שניתן.

הקושי להגדיר מיהו חרדי עומד בלב אתגר זיהוי ואמידת קבוצת האוכלוסייה הזו. הקשיים בהגדרת האוכלוסייה החרדית נובעים בין היתר מהעובדה שבשונה מדת אין הגדרה מוסכמת של חרדיות אלא ישנו מנעד של רמות דתיות שונות אשר פעמים רבות ניתנות לזיהוי רק על ידי הגדרה עצמית סובייקטיבית.

בלשכה המרכזית לסטטיסטיקה נעשו בעבר מספר ניסיונות להשתמש במידע מנהלי ובסקרים על מנת להפיק אומדני אוכלוסייה לאוכלוסייה החרדית. קיימות ארבע שיטות לאמידת האוכלוסייה החרדית – זיהוי לפי דפוסי הצבעה למפלגות חרדיות, זיהוי לפי סוג בית ספר אחרון בסקר כוח אדם, זיהוי לפי הגדרה עצמית בסקר החברתי וזיהוי לפי סוג פיקוח של מוסדות הלימוד בקבצים מנהליים.[1]

בעבודה זו מוצע מודל חדש לאמידת גודלה והרכבה של האוכלוסייה החרדית בישראל. המודל החדש מבוסס על למידת מכונה (Machine Learning) ומטרתו לבצע חיזוי עבור כל פרט באוכלוסייה שייקבע מה ההסתברות שהוא חרדי. באמצעות שימוש בלמידה מונחית (Supervised Learning) המודל "לומד" את המשתנים המאפיינים אנשים לפי תיוג של חרדי/לא-חרדי שהתקבל מנתוני הסקר החברתי. לאחר שהמודל למד את המאפיינים של הקובץ ניתן להזין אליו רשומות פרט חדשות ולקבל לכל פרט הסתברות להיות חרדי. בסופו של התהליך מתקבל אומדן שנתי ברמת פרט שמכסה את כלל האוכלוסייה היהודית אותה ניתן לפלח לפי גיל, מין, יישוב, אזור סטטיסטי ועוד, ואשר ניתן לקשר לשלל קבצים מנהליים.

[1] פרידמן י', שאול-מנע, נ', פוגל, נ', רומנוב, ד', עמדי, ד', פרידמן, מ', סחייק, ר', שיפריס, ג', ופורטנוי, ח', (2011). *שיטות מדידה ואמידת גודלה של האוכלוסייה החרדית בישראל*. סדרת ניירות טכניים מס' 25. ירושלים: הלשכה המרכזית לסטטיסטיקה.

2. רקע תיאורטי

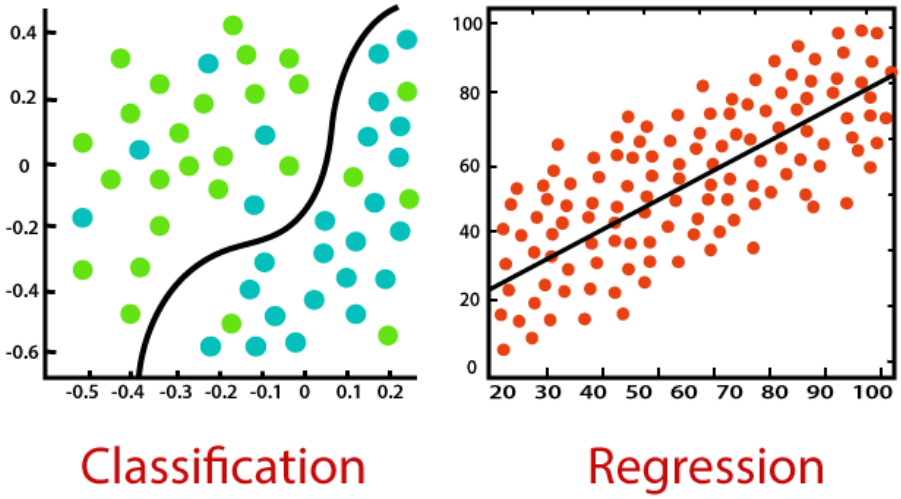
בשנים האחרונות נעשה שימוש הולך וגדל בשיטות מתקדמות על מנת לפתור בעיות בתחומים שונים כגון עסקים, רפואה, טכנולוגיה ועוד. לשיטות אלו מתייחסים בשמות שונים כמו למשל מדע הנתונים, למידת מכונה, ביג דאטה, אינטליגנציה מלאכותית ועוד, אך הרעיון הכללי מאחוריהם זהה: לשלב סטטיסטיקה, ניתוח נתונים, למידת מכונה ושיטות נוספות כדי להבין ולנתח תופעות אמיתיות בעזרת נתונים. נתונים (או דאטה) נחשבים כגורם שמאפשר קבלת החלטות טובה יותר והגידול בכמות ובמגוון הנתונים הוביל לכך שיותר ויותר ארגונים עושים שימוש במקורות המידע שלהם כדי להגיע לתובנות ולפתור בעיות שלא יכלו לפתור בעבר.[2]

עולם למידת המכונה מורכב משתי קטגוריות עיקריות: למידה מונחית (Supervised learning) ולמידה בלתי מונחית (Unsupervised learning). בלמידה מונחית המודל מאומן באמצעות מידע מתוייג, למשל נתונים של דירות ומחיר המכירה של כל אחת מהן. לאחר האימון המודל יבצע חיזוי למחיר דירה על בסיס הנתונים גם לדירות שלא ידוע מחיר מכירתן. בלמידה בלתי מפוקחת הנתונים אינם מתוייגים והמודל ינסה לזהות קבוצות בנתונים הגולמיים על פי מאפיינים שונים.

ניתן לחלק את הלמידה המונחית לשתי קטגוריות עיקריות - רגרסיה (Regression) וסיווג (Classification). בעיות מסוג רגרסיה הן בעיות בהן משתנה המטרה הוא רציף וערך החיזוי יכול להיות ערך שלא היה קיים בנתונים המקוריים, למשל כמו בדוגמא הקודמת, חיזוי מחיר מכירה של דירה. בעיות מסוג סיווג הן בעיות בהן משתנה המטרה הוא קטגוריאלי וערך החיזוי חייב להיות אחת מהקטגוריות האפשריות, למשל, האם המייל שקיבלנו הוא ספאם או לא. בעבודה זו נעשה שימוש בלמידה מונחית ובסיווג – שימוש במידע מתוייג של רמת דתיות על בסיס הגדרה עצמית של משיבים לסקר כדי לבצע חיזוי של שתי קטגוריות: חרדי/ לא-חרדי.

[2] *Hayashi, Chikio (1998). "What is Data Science? Fundamental Concepts and a Heuristic Example". Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan.*

תרשים 1. רגרסיה מול סיווג בלמידה מונחית



תהליך העבודה של פיתוח מודל למידת מכונה מונה מספר שלבים אשר ייסקרו כאן בקצרה ויפורטו בהרחבה בפרק המתודולוגי:

א. הגדרת הבעיה: מה הבעיה שצריך לפתור ומהם הכלים שעומדים לרשות הארגון על מנת לפתור אותה. הגדרה חכמה וממוקדת יכולה לעזור לכוון את כל הפרוייקט למקום הנכון.

ב. איסוף הנתונים: לא תמיד איסוף הנתונים הוא תהליך קל, לעיתים מדובר בנתונים גולמיים שיש צורך לעבד אותם לפני שניתן לעשות בהם שימוש במודל. כמו כן ישנה חשיבות מכרעת לאסוף כמות גדולה ככל האפשר של נתונים ושהם יהיו איכותיים ככל האפשר, כמות ואיכות הנתונים ישפיעו ישירות על טיב מודל החיזוי. כפי שאמר חוקר האינטליגנציה המלאכותית פיטר נורוויג – "More data beats clever algorithms, but better data beats more data".

ג. ניקוי ושיפור הנתונים: ברוב המקרים גם לאחר איסוף ועיבוד ראשוני לנתונים יהיו בעיות איכות שונות. הפעולות שנדרשות הן בדרך כלל מחיקת ערכים שגויים, השלמת ערכים חסרים ותיקון אי התאמות שונות. כמו כן בשלב זה ניתן להשתמש בנתונים הקיימים על מנת לבנות משתנים חדשים ואיכותיים יותר שפעמים רבות יכולים לשפר את טיב חיזוי המודל בצורה משמעותית. שלב זה של הכנת הנתונים צורך את רוב זמן העבודה על הפרויקט כולו.

ד. ניתוח נתונים ויזואליזציות: שלב זה מכונה EDA (Exploratory Data Analysis) ובו מבוצעת בחינה מעמיקה של המשתנים המסבירים (או פיצ'רים, Features), התפלגותם ומידת הקשר שלהם עם משתנה המטרה. בשלב זה ניתן לזהות אילו משתנים ככל הנראה לא יועילו למודל וליכולת החיזוי, אילו משתנים מסבירים ניתן לקבץ או לשנות ועוד.

ה. בחירה ואימון האלגוריתם: בחינת מספר מודלים אפשריים לפי סוג הבעיה ובחירת המודל שהשיג את התוצאות הטובות ביותר מול המטריקה שנבחרה לאמוד את איכות החיזוי של

המודל. הנתונים מפוצלים למדגם אימון שישמש לבניית המודל, ולמדגם אימות שישמש לבחינת טיב ודיוק המודל.

ו. הערכת המודל: בשלב זה מריצים את המודל הנבחר על מדגם האימות כדי לתקף את התוצאות שהתקבלו ממדגם האימון וכדי לבחון ולהעריך מה צפויה להיות מידת הדיוק שלו בעולם האמיתי.

ז. חיזוי: שימוש במודל בצורה שוטפת על מנת לבצע חיזוי על בסיס נתונים חדשים.

3. מתודולוגיה

3.1. מקור הנתונים

בעבודה זו נעשה שימוש בשלושה מקורות מידע: הסקר החברתי, אומדן רמת דתיות על פי זיקה למוסדות חינוך ואומדני אוכלוסייה ברמת פרט. מקור הנתונים המרכזי הוא בנתוני הסקר החברתי לשנים 2008-2017. הסקר החברתי הוא סקר שנתי המבוצע מאז שנת 2002 על אוכלוסיית בני 20 ומעלה. כל שנה נדגמים כ-7,500 אנשים חדשים. המטרה העיקרית של הסקר היא לספק מידע עדכני על תנאי החיים ורווחת האוכלוסייה בישראל. שאלון הסקר בנוי משני חלקים עיקריים: גרעין קבוע, המכיל כ-200 שאלות קבועות במספר רב של תחומי חיים, ובניהן גם שאלות על רמת הדתיות, וחלק נוסף המוקדש בכל שנה לנושא אחר כדי להעמיק את הידע בו. נחקרים יהודים מתבקשים להגדיר באופן סובייקטיבי את מידת דתיותם לפי השתייכות לאחת מחמש הקבוצות הבאות: חרדי, דתי, מסורתי דתי, מסורתי לא כל כך דתי, לא דתי/חילוני.[3]

מקור הנתונים המשני בו נעשה שימוש הוא אומדן רמת דתיות על פי זיקה למוסדות חינוך לשנים 2008-2017. בשיטה זו נבנה מאגר המסווג את האוכלוסייה היהודית לפי רמת דתיות בהתבסס על קישור רשומות ומיזוג של קבצים מנהליים כגון: קובץ תלמידים (משרד החינוך), קובץ ישיבות, קובצי מורים, קובץ גני ילדים ועוד. משתנה רמת הדתיות נבנה על בסיס הקבצים הנ"ל ובאמצעות אלגוריתם נבנה סיווג לפי סוג הפיקוח – לא דתי, דתי, חרדי.[4]

בנוסף נעשה שימוש גם במשתנים דמוגרפיים שנלקחו מנתוני האוכלוסייה ברמת פרט לשנים 2008-2017 ששימשו לחישוב אומדני האוכלוסייה. אומדני האוכלוסייה מבוססים על תוצאות מפקד האוכלוסין שנערך ב-2008 ועל השינויים שחלו באוכלוסייה מאז כפי שנרשמו במרשם האוכלוסין. בכל שנה מחושבת האוכלוסייה בהתאם לשינויים הדמוגרפיים שהתרחשו באותה שנה ואשר נרשמו במרשם התושבים של רשות האוכלוסין וההגירה.

3.2. ניקוי ושיפור הנתונים

ישנה הסכמה רחבה כי בעת שימוש בנתונים, איכות הניתוחים והתובנות תלויה משמעותית באיכות הנתונים עליהם הם מבוססים. ניקוי נתונים, המכונה גם טיוב נתונים, הוא אחד הצעדים החשובים ביותר סביב קבלת החלטות מבוססות נתונים ושלב זה צורך בדרך כלל את רוב זמן העבודה על הפרויקט כולו.

[3] נדגמים ערבים נחקרים לפי סולם דרגות שונה – דתי מאוד, דתי, לא כל כך דתי, לא דתי.
[4] פורטנוי, חיים. (2007). אפיון רמת הדתיות באוכלוסייה היהודית על פי זיקה למוסדות חינוך. סדרת ניירות טכניים מס' 19. ירושלים: הלשכה המרכזית לסטטיסטיקה.

ניקוי נתונים הוא ברובו תהליך של תיקון או הסרה של נתונים שגויים, פגומים או כפולים, אך גם תהליך של הבנת הנתונים בו ניתן לתקן ערכים חסרים, להוסיף משתנים ממקורות אחרים וליצור משתנים חדשים שפעמים רבות יכולים לשפר את טיב חיזוי המודל בצורה משמעותית.

בעבודה על נתוני הסקר החברתי בוצע ניקוי נתונים וסינון כך שיישארו בו רק נתונים מתאימים של יהודים (כ-55,000 רשומות). כמו כן בוצעה לכל נדגם השלמת כתובת המגורים מהמרשם בשנת הסקר, כולל עיגון כתובות, כדי להתאים את כל הכתובות מהשנים השונות לאזורים סטטיסטיים מעודכנים ל-2017. בנוסף נוספו משתנים דמוגרפיים של ארץ לידה וארץ לידת אב לכל רשומה מקובץ האוכלוסייה ברמת הפרט.

<u>יצירת משתנים חדשים</u>

משתנה אחוז חרדים (Supervised Ratio) – האוכלוסייה החרדית מאופיינת במגורים קהילתיים בשכונות ובערים הומוגניות, בעלות רוב חרדי. מכאן שמקום המגורים הינו בעל פוטנציאל ניבוי גבוה לזיהוי אוכלוסייה זו. אולם שימוש בנתוני יישוב המגורים ואזור סטטיסטי יוצרים משתנה קטגוריאלי בעל קרדינליות גבוהה, כלומר משתנה בעל מספר רב מאוד של קטגוריות, מה שמקשה על עיבוד הנתונים בשל רב המימדיות הנלווית למצב זה. לכן נעשה שימוש בשיטת Supervised Ratio כדי ליצור משתנה נומרי רציף שיפתור את בעיית הקרדינליות. בשיטה זו מחושב אחוז החרדים עבור כל שילוב של יישוב ואזור סטטיסטי (מתוך נתוני הסקר) וכך מתקבלת אינדיקציה כמותית למידת ההומוגניות החרדית של כל אזור.

משתנה רמת דתיות על פי מוסדות חינוך – משתנה מחושב המבוסס על נתוני אומדן רמת דתיות על פי זיקה למוסדות חינוך. קישור נתוני האומדן עם נתוני הסקר החברתי איפשר הוספת משתנה חדש המעריך לכל פרט את רמת דתיותו על פי מוסד הלימודים האחרון בו למד. למשתנה זה בוצעה השלמת ערכים חסרים כפי שיורחב בהמשך. הערכים החסרים נובעים בחלקם מחוסר תיעוד של מוסדות לימוד של האוכלוסייה המבוגרת.

משתנה אחים – האוכלוסייה החרדית מאופיינת במשפחות מרובות ילדים וחישוב גודל המשפחה של פרט יכולה להיות בעלת פוטנציאל ניבוי גבוה. משתנה אחים הוא משתנה מחושב אשר באמצעות קישור של כל רשומה למרשם האוכלוסין לפי מספר תעודת זהות של אם ואב מזהה את מספר הפרטים החולקים את אותם הורים. בצורה זו ניתן לכל פרט מספר משוער של אחים. גם במשתנה זה בוצעה השלמת ערכים חסרים. הערכים החסרים נובעים מאוכלוסייה מבוגרת שלא מקושרים לה הורים במרשם, וכן מאוכלוסייה צעירה של עולים, בעיקר אלו שלא הגיעו יחד עם הוריהם לארץ.

השלמת ערכים חסרים (זקיפה)

כחלק מתהליך טיוב הנתונים בוצעה השלמת ערכים חסרים עבור המשתנים מספר אחים ורמת דתיות על פי מוסדות חינוך, להם זוהו ערכים חסרים ברמה של כ-30% וכ-15%, בהתאמה. השלמת הערכים החסרים מאפשרת לא למחוק רשומות מקובץ הנתונים. כמו כן, כפי שנראה בתהליך ניתוח הנתונים, למשתנים אלו יש קשר חיובי עם משתנה המטרה ולכן השלמה איכותית של הערכים החסרים יכולה לשפר את טיב המודל. לאחר בחינת מספר שיטות להשלמת ערכים חסרים כגון: ממוצע, K-nearest neighbor, Hot deck imputation נבחרה שיטת Hot deck שהניבה תוצאות של 84% דיוק עבור משתנה אחים ו-94% דיוק עבור משתנה רמת דתיות על פי מוסדות חינוך, מתוך מדגם בדיקה מיוחד שנבנה לצורך כך.

Hot Deck Imputation[5]

שיטה לזקיפת ערכים חסרים ברשומה נתונה ("הנתרם") על ידי העתקת הערכים הנצפים המקבילים ברשומה אחרת ("התורם"). המונח Hot Deck מתייחס למצב בו התורם מגיע מאותו קובץ נתונים כמו הנתרם ולא מקובץ נתונים אחר (Cold Deck). ניתן לנסח את המשוואה הכללית של זקיפה בשיטת Hot Deck כך:

$$\tilde{y}_i = y_d$$

כאשר $\tilde{y}_i$ משקף את הערך הזקוף עבור הרשומה ה-i במשתנה הרלוונטי ו-y_d משקף את הרשומה שנבחרה כתורם. באופן כללי, נרצה למצוא תורם שדומה ככל הניתן לנתרם לפי מספר משתני עזר, לרוב אלה יהיו המשתנים המסבירים האחרים בקובץ הנתונים. ישנן דרכים שונות לבחירת תורם, מה שמוביל לגרסאות שונות של זקיפה לפי Hot Deck למשל: זקיפה אקראית, זקיפה סדרתית, זקיפת השכן הקרוב ביותר (K-nearest neighbor) ועוד.

בעבודה זו נעשה שימוש בשיטת הזקיפה האקראית שנתנה תוצאות מדויקות יותר מהשיטות האחרות. קובץ הנתונים מחולק לקבוצות על בסיס מאפיינים זהים של המשתנים המסבירים. כל הרשומות המלאות בקבוצה מסויימת מהוות תורמות פוטנציאליות לרשומות באותה קבוצה להן יש ערך חסר במשתנה המטרה. מבין הרשומות הפוטנציאליות נבחרת אחת באופן אקראי והערך החסר נזקף ברשומת הנתרם.

בשיטת הזקיפה הסדרתית קובץ הנתונים לא מחולק לקבוצות אלא האלגוריתם עובר על הרשומות בקובץ לפי הסדר וזוקף לכל רשומה חסרה את ערך הרשומה הקודמת שנצפתה שהתאימה בדיוק בכל

[5] Cranmer, S.J. and Gill, J.M.. (2013) "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science* 43:2

ערכי המשתנים המסבירים. בשיטת השכן הקרוב אין תנאי של זהות בערכי המשתנים המסבירים אלא נעשה בהם שימוש כדי להגדיר פונקציית מרחק. התורם נבחר כרשומה עבורה פונקציית המרחק היא מינימלית מבין מספר ה"שכנים" שהוגדרו מראש.

3.3. ניתוח נתונים

ניתוח נתונים מתייחס לתהליך של ביצוע חקירות ראשוניות על נתונים כדי לגלות דפוסים, לזהות חריגות, לבדוק השערות ולבדוק הנחות בעזרת סטטיסטיקה תיאורית וייצוגים גרפיים. לפני שעוברים לשלב בחירת מודל נהוג לנסות להבין את הנתונים ולאסוף מהם כמה שיותר תובנות.

תרשים 2. סטטיסטיקה תיאורית של משתנים נבחרים

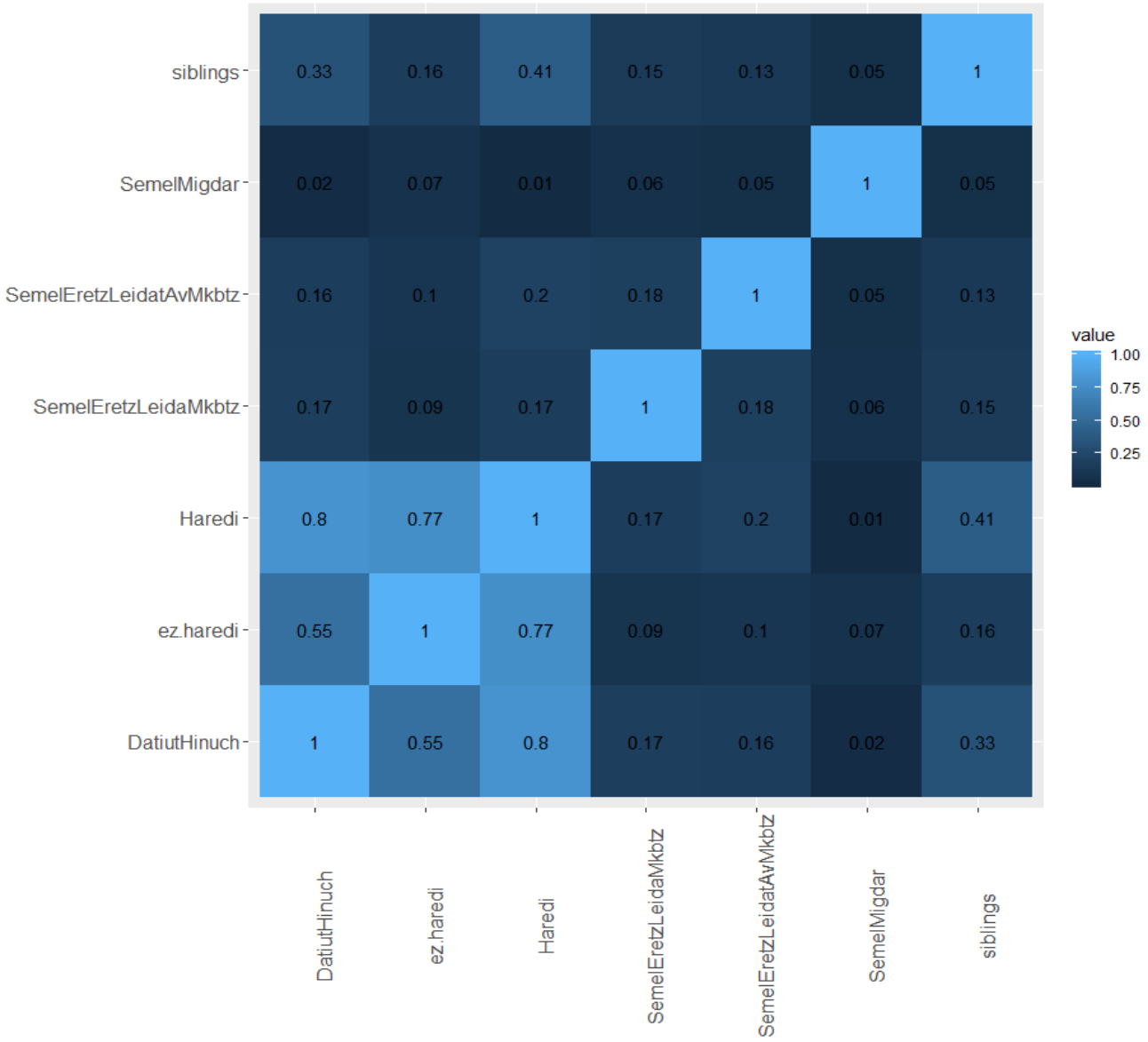
```
    Yishuv                ezor            DatiutYehudi          Haredi
 Min.   :   7        Min.   :   0.0        1  : 5041        0  :50088
 1st Qu.:2500        1st Qu.:   5.0        2  : 5871        1  : 5041
 Median :5000        Median :  52.0        3  : 7176
 Mean   :4763        Mean   : 227.6        4  :13010
 3rd Qu.:7400        3rd Qu.: 349.0        5  :24031
 Max.   :9800        Max.   :2911.0


SemelErertzLeidaMkbtz SemelErertzLeidatAvMkbtz    Siblings          SemelMigdar
 900    :34896         0      :20284          Min.   : 1.000        1  :26608
 300    : 8400         900    :13385          1st Qu.: 3.000        2  :28521
 200    : 2078         200    : 4061          Median : 3.000
 400    : 1174         50     : 2283          Mean   : 4.039
 700    :  907         310    : 2067          3rd Qu.: 5.000
 50     :  768         400    : 1602          Max.   :19.000
 (Other): 6906         (Other):11447          NA's   :18128


DatiutHinuch         ez.haredi
 1   :33964       Min.   :0.00000
 2   : 6694       1st Qu.:0.00000
 3   : 5268       Median :0.00000
 NA's: 9203       Mean   :0.09144
                  3rd Qu.:0.05000
                  Max.   :1.00000
```

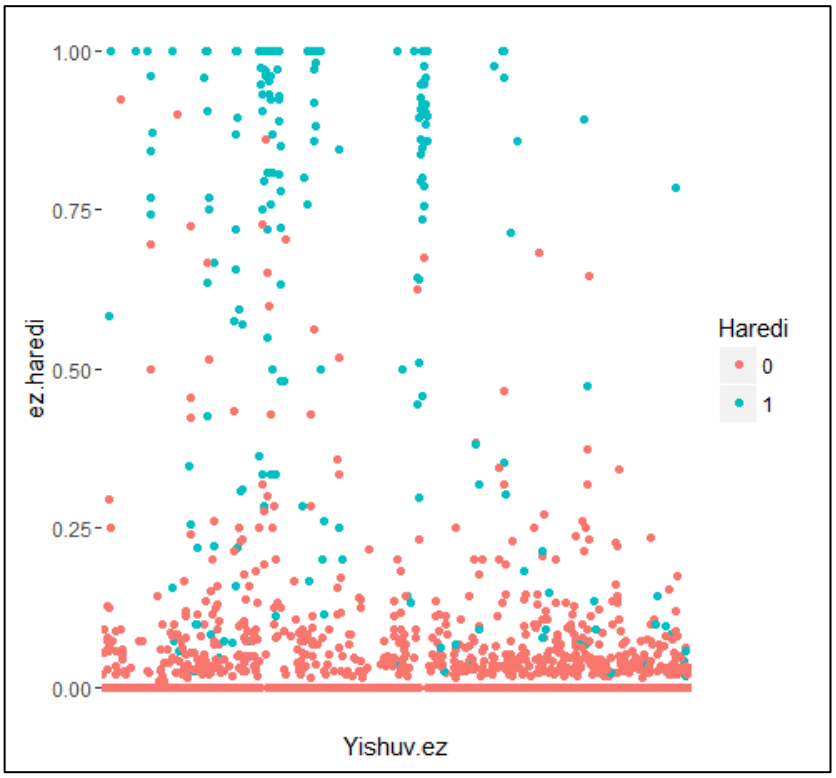
כיוון שמרבית המשתנים הם קטגוריאליים נעשה שימוש ב-Cramer's V על מנת לבחון את הקשרים בין המשתנים לבין עצמם וכן את הקשר שלהם למשתנה המטרה. אפשר לראות קשר חזק בין המשתנים אחוז חרדים (ez.haredi בתרשים), רמת דתיות על פי מוסדות חינוך (DatiutHinuch) ומספר אחים (Siblings) לבין משתנה המטרה חרדי (Haredi).

תרשים 3. מפת חום לפי Cramer's V

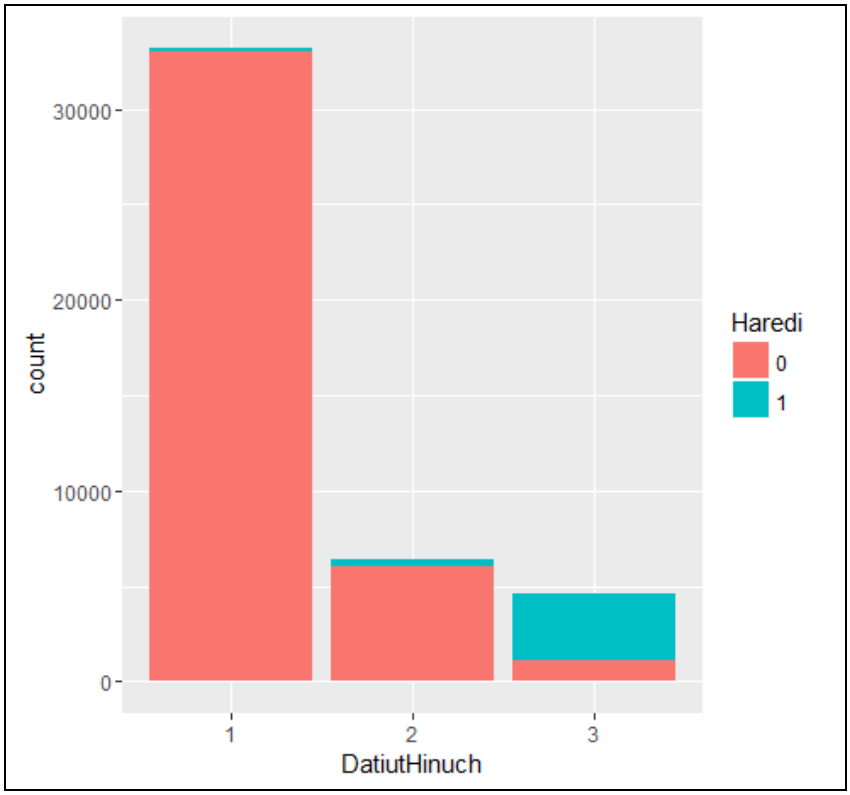


בנוסף בוצעו ויזואליזציות של המשתנים המסבירים הבולטים - אחוז חרדים, רמת דתיות על פי מוסדות חינוך ומספר אחים, מול משתנה המטרה (חרדי/לא חרדי). מהנתונים אפשר ללמוד כי יש קשר בין האוכלוסייה החרדית לבין מגורים באזורי מגורים בעלי רוב חרדי, למידה במוסדות חינוך בעלי פיקוח חרדי, ושככל שישנם יותר ילדים במשפחה חלקה היחסי של האוכלוסייה החרדית באזור הסטטיסטי/יישוב גדל.

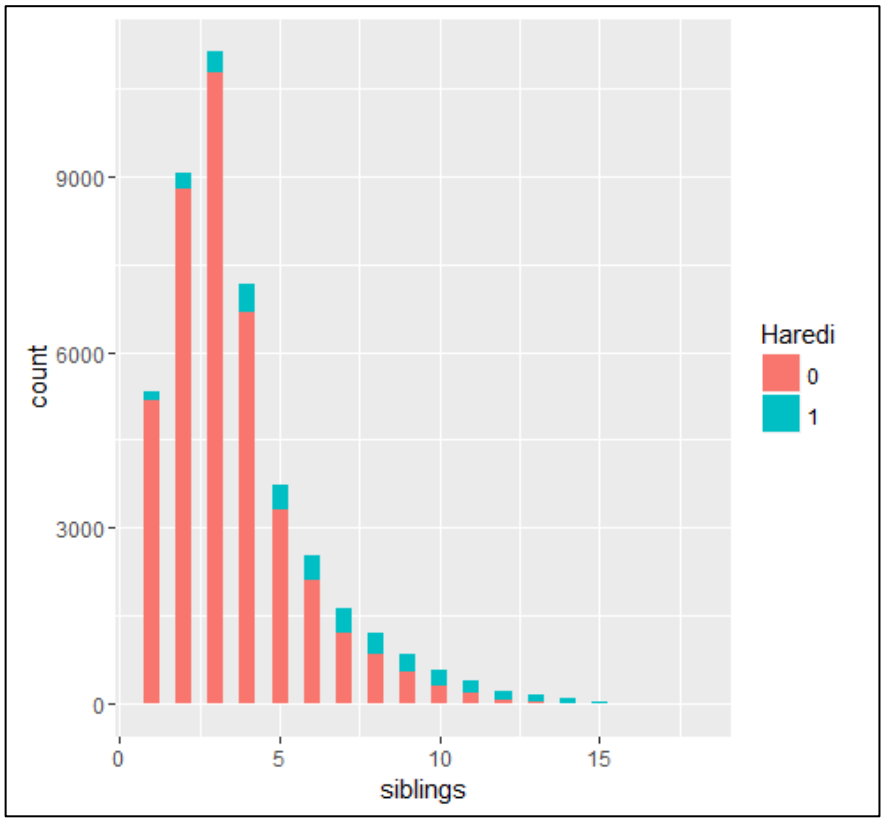
תרשים 4. השוואת המשתנים אחוז חרדים ויישוב מול משתנה חרדי



תרשים 5. השוואת המשתנה רמת דתיות על פי מוסדות חינוך מול משתנה חרדי



תרשים 6. השוואת המשתנה מספר אחים מול משתנה חרדי

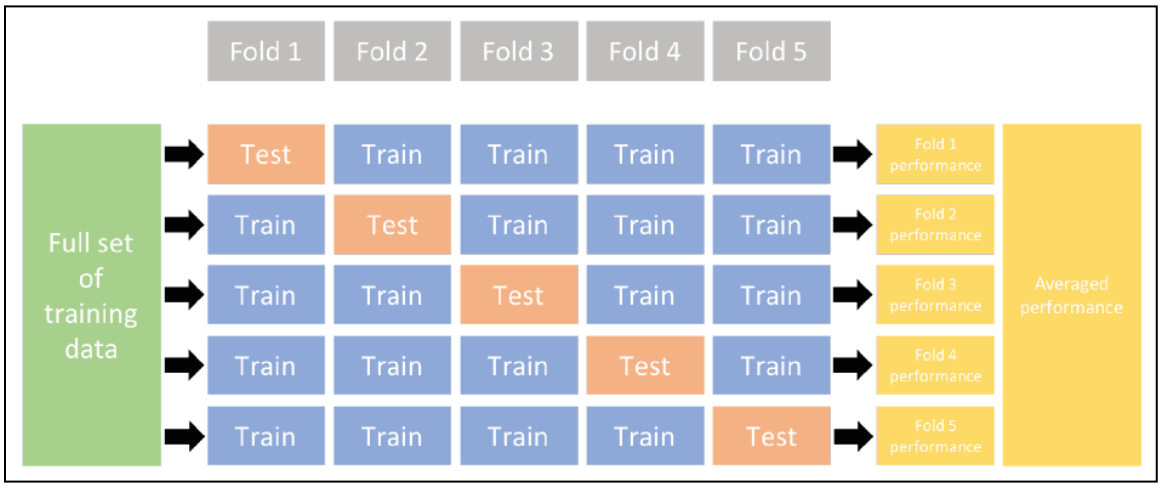


3.4. בחירת ואימון האלגוריתם

אחת המטרות העיקריות בתהליך למידת המכונה היא למצוא אלגוריתם המנבא בצורה המדויקת ביותר ערכים עתידיים בהתבסס על קבוצה של תכונות. במילים אחרות, נרצה אלגוריתם שלא רק מתאים היטב לנתוני העבר, אלא חשוב מכך, כזה שמנבא תוצאה עתידית במדוייק. זה נקרא הכלליות (Generalizability) של האלגוריתם. כדי להבין באיזו מידה האלגוריתם מצליח לנבא נכון נתונים חדשים נהוג לפצל את הנתונים למדגם אימון (Training set) ולמדגם אימות (Validation set). מדגם האימון משמש לאימון האלגוריתם, כיוון של היפר-פרמטרים, השוואה בין מודלים שונים ועוד. לאחר בחירת מודל סופי מדגם האימות משמש לבחינת טיב ודיוק המודל על מידע שלא היה ידוע בעת בנייתו. יחס הפיצול שנבחר בעבודה זו הוא 80% אימון 20% אימות, בהתאם לגודל קובץ הנתונים המלא. כמו כן, כיוון שקיים חוסר איזון מובנה בקובץ הנתונים (כ-90% אינם חרדים ו-10% חרדים) קיים חשש שפיצול אקראי של הנתונים ייפגע בפרופורציית החרדים שקיימת באוכלוסייה. לכן פיצול הנתונים נעשה בשיטה של דגימה מרובדת (Stratified Sampling) כך שלאחר הפיצול היחס בין חרדי ללא-חרדי יישמר בשני המדגמים.

כדי לא 'לזהם' את התהליך לא נעשה שימוש במדגם האימות להערכת ביצועי האלגוריתם בשלב האימון. על מנת להעריך את ביצועי האלגוריתם נהוג להשתמש בשיטות דגימה חוזרת (Resampling) המאפשרות להתאים אלגוריתם שוב ושוב לחלקים ממדגם האימון ולהעריך את ביצועיו על החלקים הנותרים. בעבודה זו נעשה שימוש בשיטת k-Fold Cross Validation אשר מחלקת את מדגם האימון ל-k קבוצות שוות בגודלן, מאמנת אלגוריתם על בסיס k-1 קבוצות ואז משתמשת בקבוצה הנותרת כדי לבדוק את ביצועיו. תהליך זה חוזר על עצמו k פעמים כאשר בכל פעם קבוצה אחרת נחשבת לקבוצת הבדיקה (Test set). בסוף התהליך מחושב ממוצע התוצאות אשר מהווה קירוב למידת הדיוק שאפשר לצפות מהרצת האלגוריתם על נתונים חדשים. בדרך כלל נהוג לתת ערך של k=5 או k=10 אך אין כלל פורמלי לגבי גודלו של הפרמטר k, עדיף להתאים את k לגודל המדגם כדי שיהיו מספיק רשומות בכל קבוצה. בעבודה זו נעשה שימוש ב- k=10. שיטה זו טובה גם כדי להקטין מצב של התאמת יתר לנתונים (Overfitting), בו מודל משיג אחוזי דיוק גבוהים מול נתוני מדגם האימון אך משיג אחוזי דיוק נמוכים משמעותית מול מדגם האימות כתוצאה מהתאמת המודל בצורה מוגזמת לניואנסים הקיימים במדגם האימון.

תרשים 7. תהליך העבודה של שיטת k-Fold Cross Validation



הגישה הרווחת להערכת ביצועי אלגוריתם היא לעשות שימוש במטריקות (מדדים) שהן למעשה פונקציות הפסד המודדות עד כמה הערכים החזויים קרובים לערכים בפועל. ישנן מטריקות רבות אשר שמות דגש על היבטים שונים ושמתאימות לסוגי בעיות שונות, למשל לבעיות רגרסיה ישנן מטריקות שונות מבעיות סיווג. בעבודה זו נעשה שימוש במטריקה ROC-AUC.

ROC-AUC

בעת יישום מודלים של סיווג, לעתים קרובות נעשה שימוש במטריצת טעות (Confusion Matrix) כדי להעריך מדדי ביצוע. מטריצת טעות היא מטריצה שמשווה רמות קטגוריות (או אירועים) בפועל לרמות הקטגוריות החזויות.

כאשר בוצע חיזוי של האירוע הרצוי, נתייחס לכך כאל חיובי אמיתי (True Positive), כאשר בוצע חיזוי נכון של האירוע הלא-רצוי נתייחס לכך כאל שלילי אמיתי (True Negative). לעומת זאת, אם בוצע חיזוי של אירוע שלא קרה נתייחס לכך כאל חיובי כוזב (False Positive). לחלופין, אם לא בוצע חיזוי של אירוע והוא התרחש נתייחס לכך כאל שלילי כוזב (False Negative).

המדד הבסיסי ביותר שעושה שימוש במטריצת הטעות הוא מדד דיוק – Accuracy שבוחן בסה"כ באיזו תדירות המודל מבצע חיזוי נכון, כלומר: $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$, המטרה היא למקסם את הערך כדי להגיע לדיוק הגבוה ביותר.

תרשים 8. מטריצת הטעות

	Actual events	Actual non-events
Predicted events	True Positive (TP)	False Positive (FP)
Predicted non-events	False Negative (FN)	True Negative (TN)

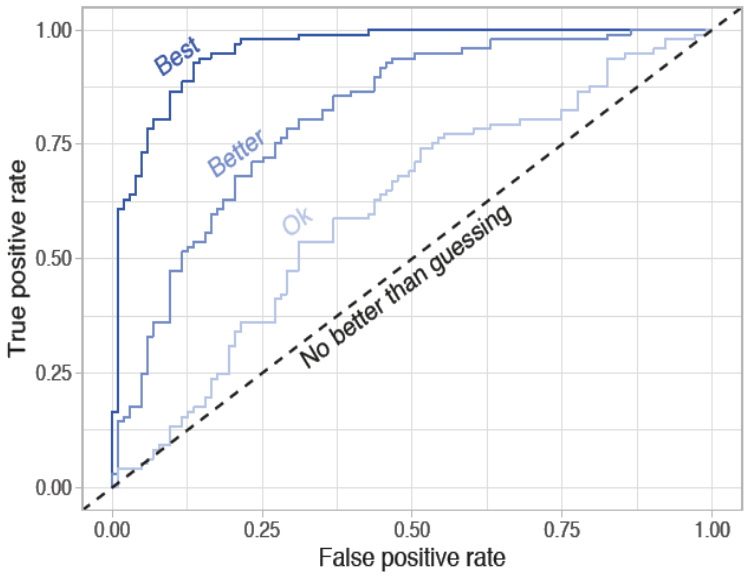
מטריקה נוספת היא שיעור החיוביים האמיתיים (TPR – True Positive Rate) שמודדת את שיעור הסיווגים החיוביים מתוך כלל הפריטים החיוביים: $\text{TPR} = \frac{TP}{TP+FN}$.

כמו כן ישנה המטריקה שיעור החיוביים הכוזבים (FPR – False Positive Rate) שמודדת את שיעור הסיווגים החיוביים מתוך כלל הפריטים השליליים: $\text{FPR} = \frac{FP}{TN+FP}$.

המטריקה ROC-AUC שפירוש שמה הוא Receiver Operating Characteristics- Area Under The Curve, יוצרת עקומה אשר מטרתה היא לאזן בין FPR לאורך ציר ה-x לבין TPR לאורך ציר ה-y, על פני ספי החלטה שונים. ככל שהשטח מתחת לעקומה מתקרב לאחד כך האלגוריתם מבצע פחות חיוביים כוזבים ושליליים כוזבים ולמעשה מצליח לחזות נכון את שתי הרמות של משתנה המטרה. בעבודה זו נעשה שימוש במטריקה זו כיוון שהיא מאפשרת לאמן אלגוריתם מדוייק יותר במקרים בהם

משתנה המטרה אינו מאוזן כמו במקרה של משתנה חרדי. נקודה זו חשובה במיוחד במקרה של האוכלוסייה החרדית כיוון שהיא מהווה נתח קטן יחסית מכלל האוכלוסייה ולכן חשוב לבדוק עד כמה המודל מזהה נכון חרדים אך לא פחות חשוב כמה הוא מזהה נכון לא-חרדים.

תרשים 9. עקומת ROC



על מנת לבחור את האלגוריתם הטוב ביותר עבור הנתונים נעשה שימוש במטריקת ROC-AUC. נבדקו מודלים שונים אשר מתאימים לבעיות מסוג סיווג, להלן המודלים שנבחנו עם ערך ה--ROC AUCשהתקבל עבורם:

- 0.975 – Logistic Regression
- 0.964 – Random Forest
- 0.955 – Support Vector Machines
- 0.985 – XGBoost

המודל שהשיג את התוצאות הטובות ביותר הוא XGBoost עם מדד ROC-AUC של 0.985. המשתנים המסבירים שנבחרו במודל הסופי הם: אחוז חרדים, רמת דתיות על פי מוסדות חינוך, מספר אחים, ארץ לידה, ארץ לידת אב ומין.

מודל XGBoost (eXtreme Gradient Boosting)[6]

XGBoost הוא אלגוריתם למידת מכונה מבוסס עץ החלטות (Decision Tree). האלגוריתם פותח ב-2016 כפרוייקט מחקר באוניברסיטת וושינגטון בארצות הברית. מאז השקתו זכה המודל לפופולריות רבה בעיקר בזכות ביצועי חיזוי עדיפים על פני מודלים אחרים.

בכדי להבין את האלגוריתם עלינו להסביר מהם המונחים Gradient Descent ו-Gradient Boosting:
Gradient Descent – פונקציית הפסד מודדת עד כמה הערכים החזויים קרובים לערכים בפועל. באופן אידיאלי, אנו רוצים הבדל קטן ככל האפשר בין הערכים החזויים לערכים בפועל. לפיכך, אנו רוצים למזער את פונקציית ההפסד. המשקולות המשוייכות למודל מאומן, גורמות לו לחזות ערכים הקרובים לערכים בפועל. לפיכך, ככל שהמשקולות המשויכות למודל טובות יותר, כך הערכים החזויים מדוייקים יותר וערך פונקציית ההפסד קטן יותר. ככל שמדגם האימון יכיל יותר רשומות כך המשקולות יתעדכנו ויהיו מדוייקות יותר. Gradient Descent הוא אלגוריתם אופטימיזציה איטרטיבי. זוהי שיטה למזער פונקציה עם מספר משתנים. לפיכך, ניתן להשתמש ב-Gradient Descent כדי למזער את פונקציית ההפסד. תחילה האלגוריתם מריץ את המודל עם משקולות ראשוניות, ואז מבקש למזער את פונקציית ההפסד על ידי עדכון המשקולות על פני מספר איטרציות.

Gradient Boosting – נבנה אנסמבל של מסווגים (Classifiers) חלשים, כאשר לרשומות שסווגו לא נכון ניתן משקל גדול יותר ("מוגבר") כדי לחזות אותם נכון במודלים מאוחרים יותר. מסווגים חלשים אלה משולבים מאוחר יותר כדי לייצר מסווג חזק אחד. ישנם אלגוריתמי Boosting רבים כגון AdaBoost, Gradient Boosting ו-XGBoost. שני האחרונים הם מודלים מבוססי עצים. Gradient Boosting מביא את העיקרון של Gradient Descent ו-Boosting לתחום הלמידה המונחית. מודלים משופרים אלה הם עצים שנבנו ברצף, ובהם אנו לוקחים את הסכום המשוקלל של מספר מסווגים.

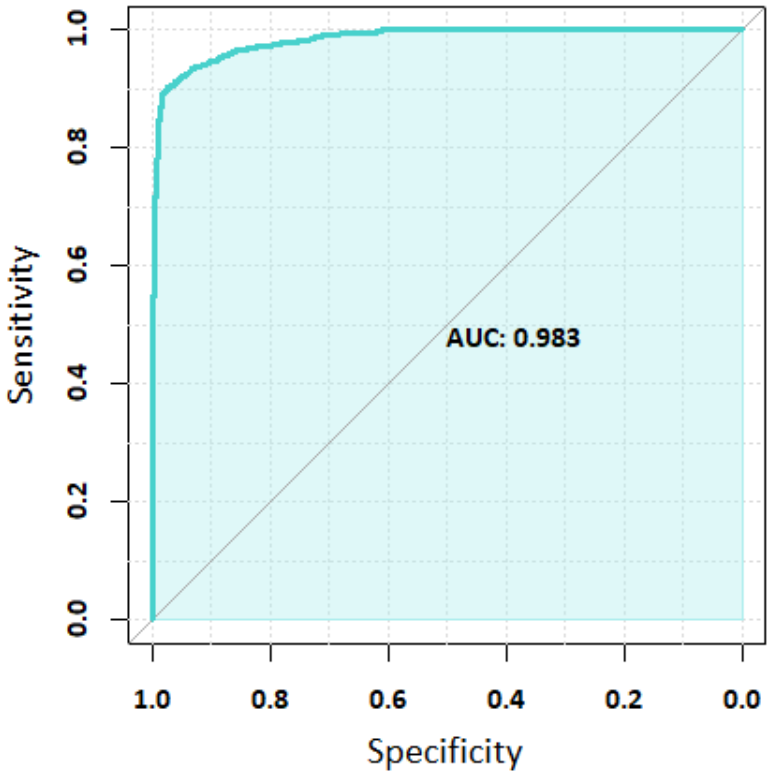
XGBoost נבנה כדי לשפר את מגבלת משאבי החישוב לעצים מוגברים, הוא יישום של אלגוריתמי Boosting עם שיפורים משמעותיים. אלגוריתמי Boosting בונים עצים ברצף, XGBoost בונה עצים במקביל וזה הופך אותו לאלגוריתם מהיר יותר. כמו כן, האלגוריתם מוסיף אובייקט רגולציה לפונקציית ההפסד. אובייקט הרגולציה "מעניש" את המודל כדי למנוע ממנו להפוך למורכב מדי ובכך מקטין את הסיכוי להתאמת יתר של המודל לנתוני האימון (Overfitting). פונקציית ההפסד המעודכנת נראית כך: $J(\theta) = L(\theta) + \Omega(\theta)$, כאשר L היא פונקציית ההפסד המבוססת על מדגם האימון ו-Ω הוא אובייקט הרגולציה.

[6] Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen & R. Rastogi (eds.), *KDD*

3.5. הערכת האלגוריתם

לאחר בחירת האלגוריתם שהשיג את התוצאות הטובות ביותר על מדגם האימון יש לבדוק עד כמה האלגוריתם מצליח לבצע חיזוי נכון על נתוני מדגם האימות. אימות זה הוא הדרך בה נוכל להעריך מה צפויה להיות מידת הדיוק שלו בעולם האמיתי. כפי שאפשר לראות בתרשים 10 הרצת המודל על מדגם האימות החזירה מטריקת ROC-AUC עם ערך של 0.983 בדומה למה שהתקבל בעת הערכת ביצועי המודל על נתוני מדגם האימון. משמעות התוצאה הזאת היא שהאלגוריתם מבצע הכללה טובה לנתונים חדשים ולא סובל מהתאמת יתר. כמו כן, בחינת פירוט מטריצת הטעות מעלה שהמודל מנבא נכון 98.3% ממי שסווגו כלא-חרדים ו-88.0% ממי שסווגו כחרדים ובסה"כ מזהה נכון 97.4% מהרשומות.

תרשים 10. עקומת ROC שהתקבלה על בסיס מדגם האימות



תוצאות המודל על מדגם האימון מתייחסות לחלוקת ההסתברויות של חרדי/לא-חרדי על פי סף של 0.31. בדרך כלל כאשר יש צורך למפות הסתברויות לקטגוריות של משתנה בינארי נהוג להשתמש בסף של 0.5, כלומר כל הערכים שמעל 0.5 ימופו לקטגוריה אחת וכל הערכים שמתחת לחצי ימופו לקטגוריה השנייה. אך במקרים של חוסר איזון בין הקטגוריות, כמו המקרה בעבודה זו, סף ברירת המחדל הזה עלול להביא לביצועים ירודים של המודל. במקרים אלו שינוי הסף על בסיס מדד כלשהו יכול לשפר את ביצועי המודל. ישנן שיטות שונות העוזרות לבחור את הסף המתאים למודל, בעבודה זו נעשה שימוש

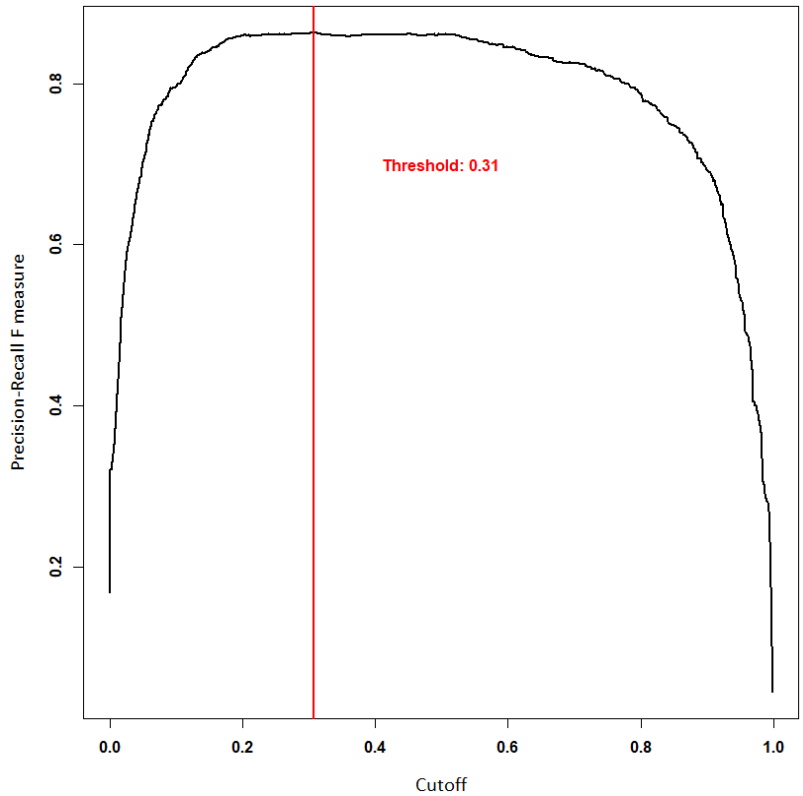
במדד F1- Score, המכונה גם F1- Measure, אשר נמצא יעיל במקרים של משתנה בינארי לא מאוזן כיוון שהוא נותן יותר דגש למצבים בהם לחיובים כוזבים ולשליליים כוזבים יש משקל שונה.

F1- Score מוגדר כממוצע ההרמוני של שני מדדים:

- Precision – מודד את שיעור הפריטים החיוביים מתוך כלל הפריטים שסווגו לקבוצה "חיובי" ומוגדר על ידי $\text{Precision} = \frac{TP}{TP+FP}$.
- Recall – שם אחר לשיעור החיוביים האמיתיים (TPR) שכאמור מודד את שיעור הסיווגים החיוביים מתוך כלל הפריטים החיוביים.

ומכאן ש-F1- Score מוגדר כ- $F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$ וערכיו נעים בין אפס לאחד. חישוב המדד עבור כל סף אפשרי יוצר עקומה שנקודת המקסימום שלה מסמלת את ערך הסף האופטימלי. במקרה זה התקבל ערך F1- Score מקסימלי של 0.86 שהיתרגם כאמור לערך סף של 0.31 ובו נעשה שימוש כדי לסווג את המשתנה חרדי לשתי הקטגוריות חרדי ולא-חרדי.

תרשים 11. עקומת F1-Score



3.6. פירוש האלגוריתם

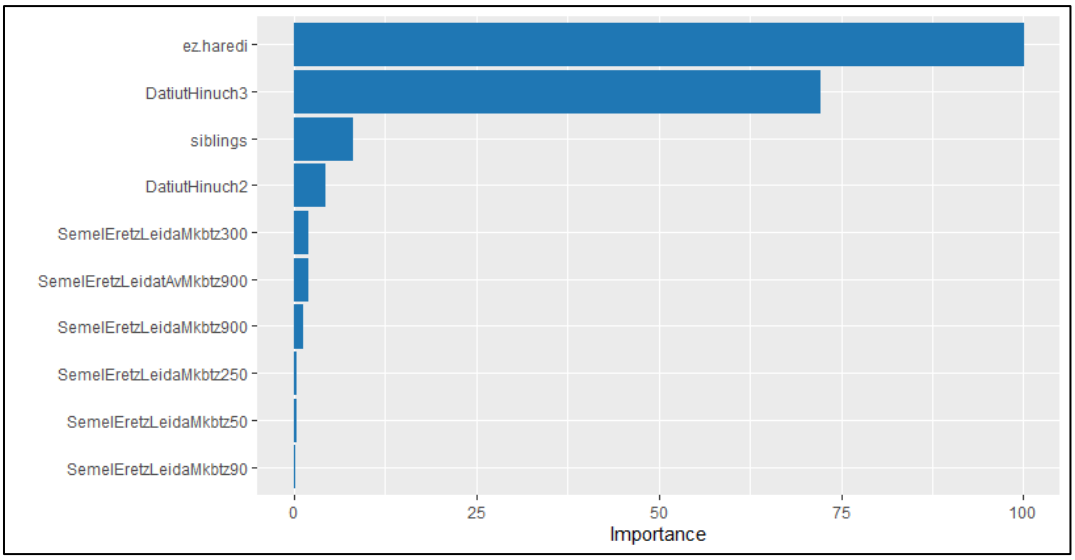
מודלים מתקדמים של למידת מכונה, דוגמת XGBoost, נחשבים לעיתים קרובות כ"קופסאות שחורות" בשל פעולתם הפנימית המורכבת. עם זאת, בגלל מורכבותם הם בדרך כלל מדויקים יותר לחיזוי תופעות לא לינאריות או נדירות. לרוע המזל דיוק טוב יותר בא לעיתים קרובות על חשבון פרשנות (Interpretability), ופרשנות היא חיונית עבור תיעוד המודל, פיקוח על דרך פעולתו וכן עבור קבלת אמונם של מקבלי ההחלטות. בשנים האחרונות פותחו מספר שיטות המסייעות בפירוש מודלים של למידת מכונה וניתן להשתמש בהן על מנת לחלץ תובנות חשובות על ביצועי המודל.

ניתן לחלק את הגישות לפירוש המודל לשתי קבוצות עיקריות: פרשנות גלובלית ופרשנות מקומית.

פרשנות גלובלית עוסקת בהבנת האופן בו המודל מבצע חיזויים בהתבסס על ראייה כוללת של תכונותיו וכיצד הן משפיעות על מבנה המודל הבסיסי. פרשנות גלובלית מסייעת בהבנת הקשר בין משתנה המטרה לבין המשתנים המסבירים, לרוב נעשה שימוש בגישה זו כדי להציג את המשתנים המסבירים המשפיעים ביותר על המודל.

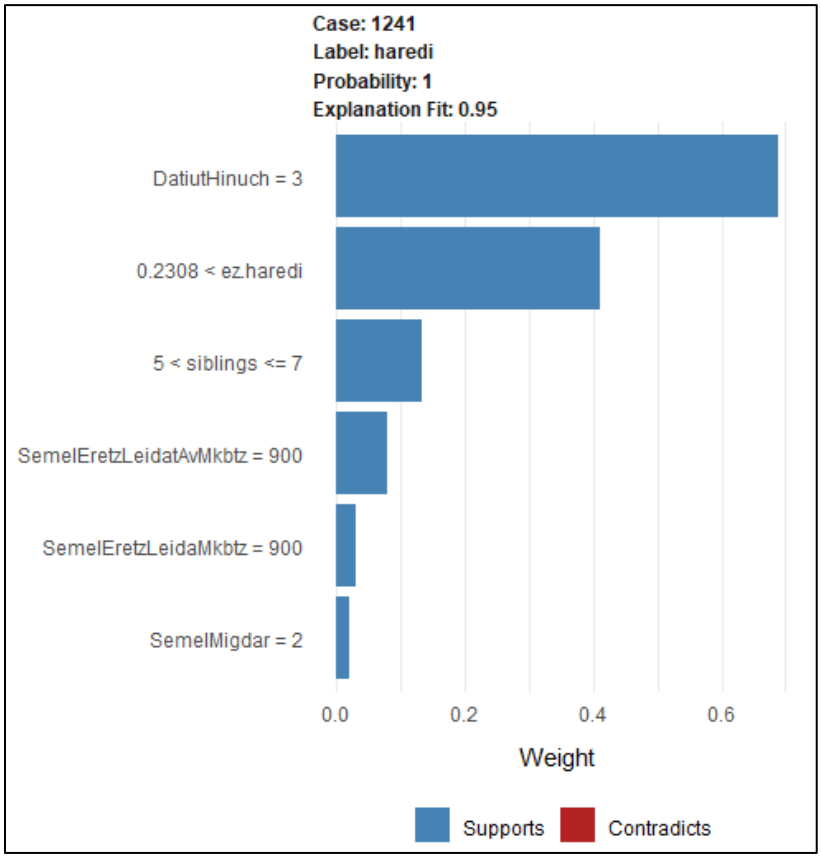
אחת השיטות הנפוצות לעשות זאת היא שיטת חשיבות המשתנים (Feature Importance) בה מחושבת חשיבותו של משתנה על ידי חישוב העלייה בטעות החיזוי של המודל לאחר החלפה אקראית של ערכי המשתנה, באופן שמנתק את הקשר בינו לבין משתנה המטרה. משתנה הוא "חשוב" אם ערבוב אקראי של ערכיו מגדיל את טעות המודל, מכיוון שבמקרה זה המודל הסתמך על המשתנה הזה לצורך חיזוי. משתנה הוא "לא חשוב" אם ערבוב הערכים משאיר את טעות המודל ללא שינוי. בתרשים 12 אפשר לראות שהמשתנים החשובים ביותר שהתקבלו הם אחוז חרדים, רמת דתיות על פי מוסדות חינוך ומספר אחים (השיטה, כמו המודל עצמו, מתייחסת לכל קטגוריה כמשתנה בינארי בפני עצמו).

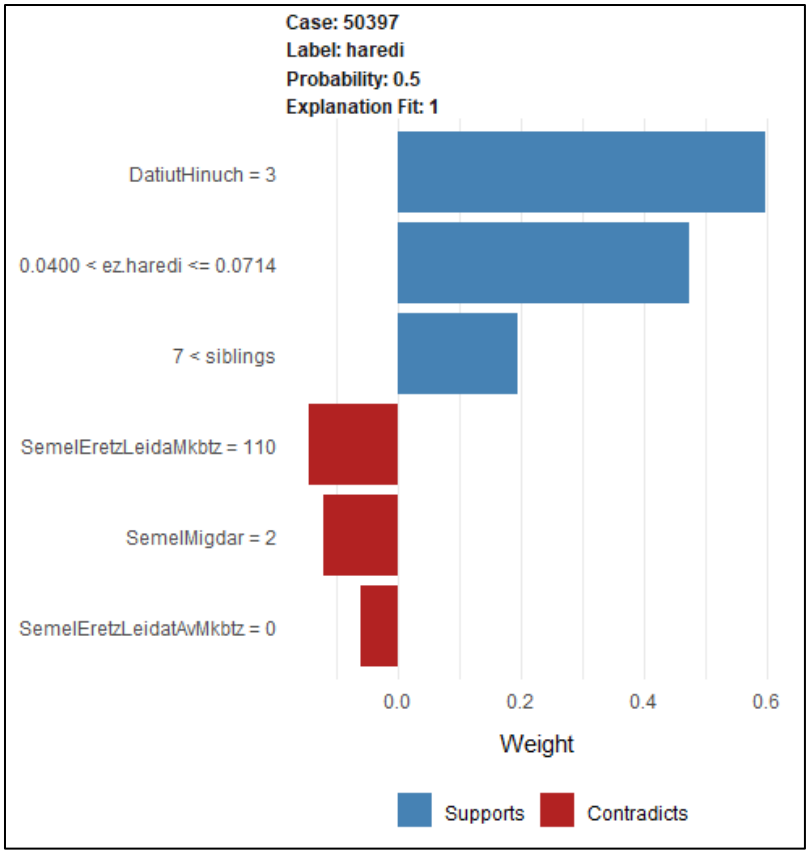
תרשים 12. פרשנות גלובלית – חשיבות המשתנים

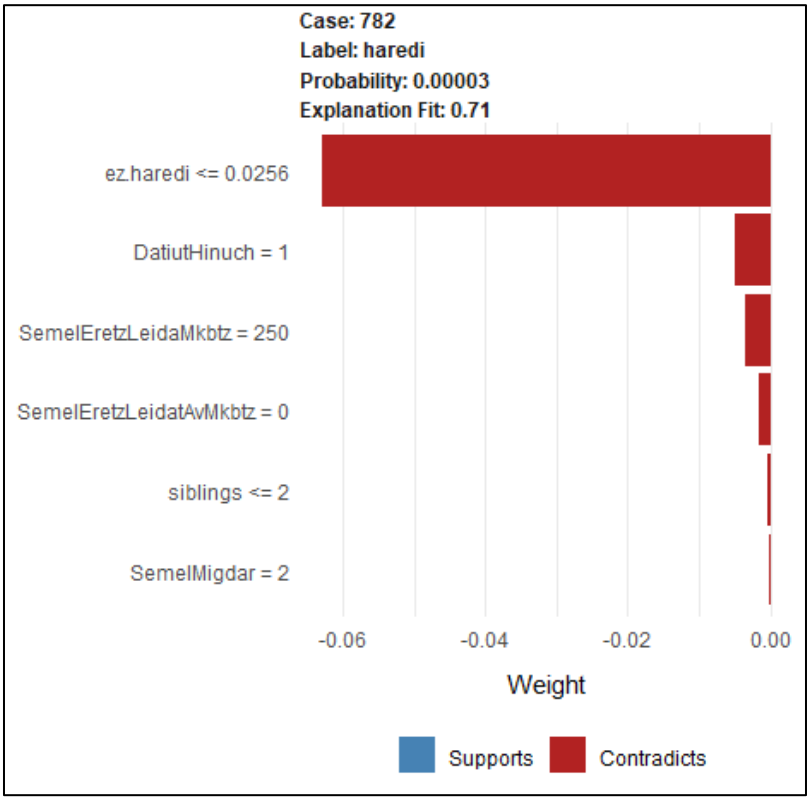


בשונה מפרשנות גלובלית, **פרשנות מקומית** עוזרת להבין כיצד המשתנים המסבירים משפיעים על החיזוי לרשומה נתונה וכך ניתן לראות טוב יותר את יחסי הגומלין בין המשתנים באופן שלא אפשרי בראייה כוללת. בדרך זו ניתן לראות מדוע המודל ביצע חיזוי לרשומה מסויימת על בסיס מאפייני המשתנים המסבירים של אותה רשומה. ישנן מספר שיטות לביצוע פרשנות מקומית, בעבודה זו נעשה שימוש בשיטת LIME (הערת שוליים - Local Interpretable Model-agnostic Explanations) שעוזרת להסביר חיזוי ספציפי. מאחורי פעולתה של LIME מסתתרת ההנחה שכל מודל מורכב הוא לינארי בקנה מידה מקומי (כלומר באזור קטן סביב רשומה מעניינת) וכי ניתן להתאים מודל פשוט סביב רשומה אחת שתחקה כיצד המודל הגלובלי מתנהג באותו אזור. תרשימים 15-13 מציגים שלוש רשומות שנבחרו לפי רמת ההסתברות לחרדיות שהמודל העניק להן – הסתברות גבוהה (1), הסתברות בינונית (0.5) והסתברות נמוכה (0). אפשר לראות שאחוז ההומוגניות באזור המגורים, המוסד החינוכי, מספר האחים ומשתנים נוספים משפיעים מאוד על האופן בו קובע המודל את ההסתברות.

תרשים 15-13. פרשנות מקומית – פירוט ההסתברות של רשומות נבחרות




Case: 50397
Label: haredi
Probability: 0.5
Explanation Fit: 1
DatiutHinuch = 3
0.0400 < ez.haredi <= 0.0714
7 < siblings
SemelEretzLeidaMkbtz = 110
SemelMigdar = 2
SemelEretzLeidatAvMkbtz = 0
0.0
0.2
0.4
0.6
Weight
Supports
Contradicts


Case: 782
Label: haredi
Probability: 0.00003
Explanation Fit: 0.71
ez.haredi <= 0.0256
DatiutHinuch = 1
SemelEretzLeidaMkbtz = 250
SemelEretzLeidatAvMkbtz = 0
siblings <= 2
SemelMigdar = 2
-0.06
-0.04
-0.02
0.00
Weight
Supports
Contradicts

4. תוצאות

על מנת להפעיל את האלגוריתם המאומן על כלל האוכלוסייה היהודית בישראל היה צורך לבנות את המשתנים המסבירים עבור כלל האוכלוסייה. משתנה אחוז חרדים נוצר על בסיס החישוב שבוצע בנתוני מדגם האימון, אזורים שלא נדגמו בסקר החברתי קיבלו ערך של אפס. משתנה רמת דתיות על פי מוסדות חינוך נלקח כפי שהוא מנתוני אומדן רמת דתיות על פי זיקה למוסדות חינוך ובוצעה בו השלמת ערכים חסרים בשיטת Hot Deck כפי שנעשה במהלך בניית המודל. כמו כן היות והאומדן מבוסס על מרשם אפריל בכל שנה, נוצר מחסור בגיל אפס כיוון שחלק ניכר מהתינוקות נולדו לאחר חודש אפריל. נתונים אלו הושלמו על ידי מתן רמת הדתיות של האם לכל תינוק. משתנה אחים נבנה באותה צורה כמו במהלך בניית המודל ובוצעה בו השלמת ערכים חסרים בשיטת Hot Deck. המשתנים ארץ לידה, ארץ לידת אב ומין היו זמינים ומלאים בקובצי אומדני האוכלוסייה ברמת הפרט לשנת 2017.

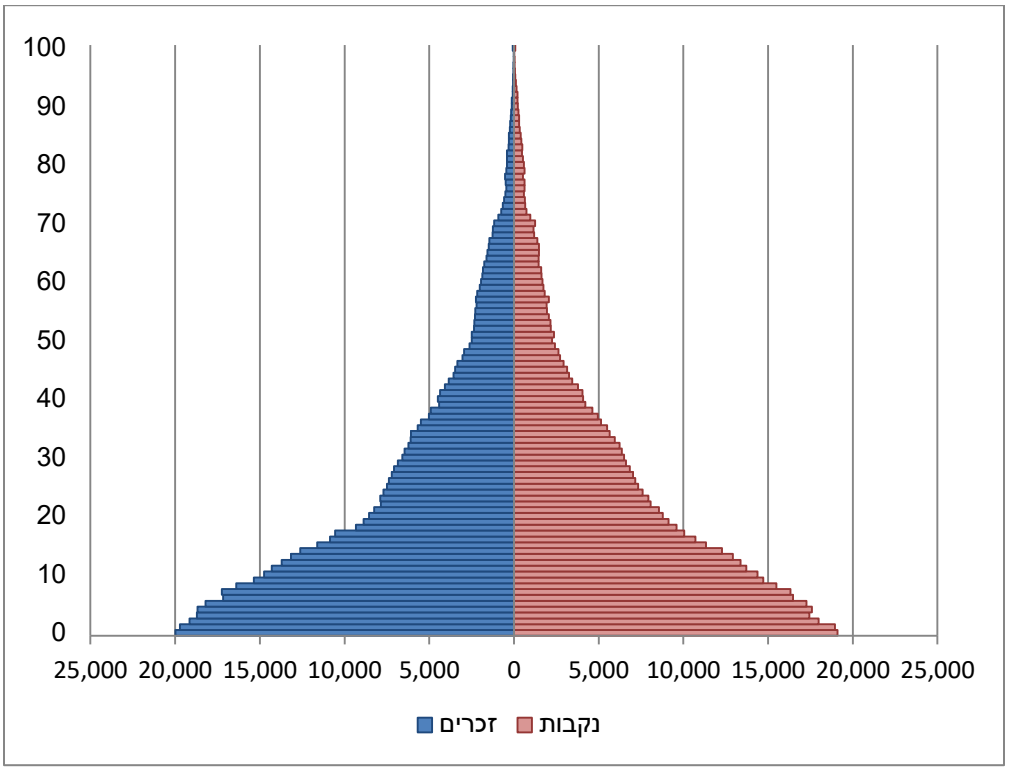
הרצת המודל על כלל האוכלוסייה היהודית מצאה כי נכון לסוף 2017 האוכלוסייה החרדית מנתה 1,018,672 תושבים. על פי המודל האוכלוסייה החרדית מהווה 15.5% מכלל האוכלוסייה היהודית ו-11.5% מקרב כלל אוכלוסיית ישראל. כמו כן, הנתונים מראים כי האוכלוסייה החרדית היא קבוצת אוכלוסייה צעירה מאוד כאשר כמחצית מהאוכלוסייה החרדית מורכבת מילדים מתחת לגיל 15 וכי 4% בלבד הינם מבוגרים מעל גיל 65.

האוכלוסייה החרדית בישראל לפי גיל ומין – סוף 2017

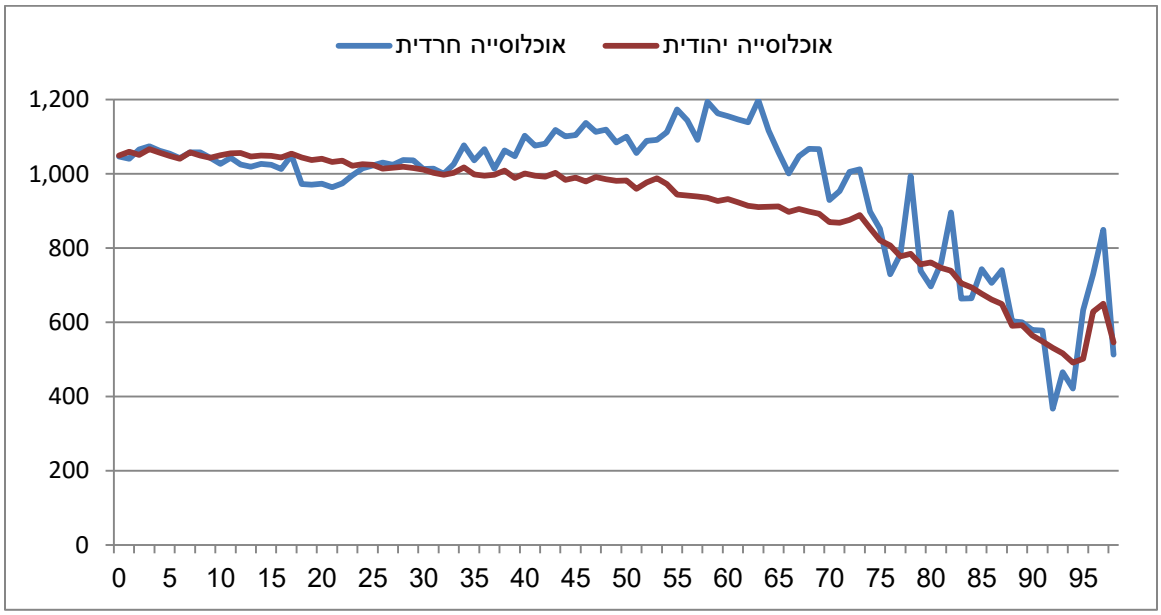
גיל	מספרים מוחלטים			אחוזים		
	סה"כ	זכרים	נקבות	סה"כ	זכרים	נקבות
סה"כ	1,018,672	518,631	500,041	100%	100%	100%
0-14	487,226	249,214	238,012	48%	48%	48%
15-64	494,931	252,243	242,688	49%	49%	49%
65+	36,515	17,175	19,340	4%	3%	4%

מבנה הגילים המלא כפי שנראה מתרשים פירמידת הגילים מתאים לאוכלוסיות צעירות המיוצגות על ידי פירמידות עם בסיס רחב של ילדים צעירים ופסגה צרה של אנשים מבוגרים, בעוד שאוכלוסיות מבוגרות מאופיינות במספר אחיד יותר של אנשים בקטגוריות הגיל. מבנה זה נראה הולם עבור אוכלוסייה כמו האוכלוסייה החרדית המאופיינת במשפחות מרובות ילדים. בחינת יחס המינים שהתקבל עבור האוכלוסייה החרדית מעידה על מחסור קל בנשים ביחס לכלל האוכלוסייה היהודית.

תרשים 16. פירמידת גילים של האוכלוסייה החרדית – סוף 2017



תרשים 17. יחס המינים לפי גיל, זכרים לאלף נקבות – סוף 2017



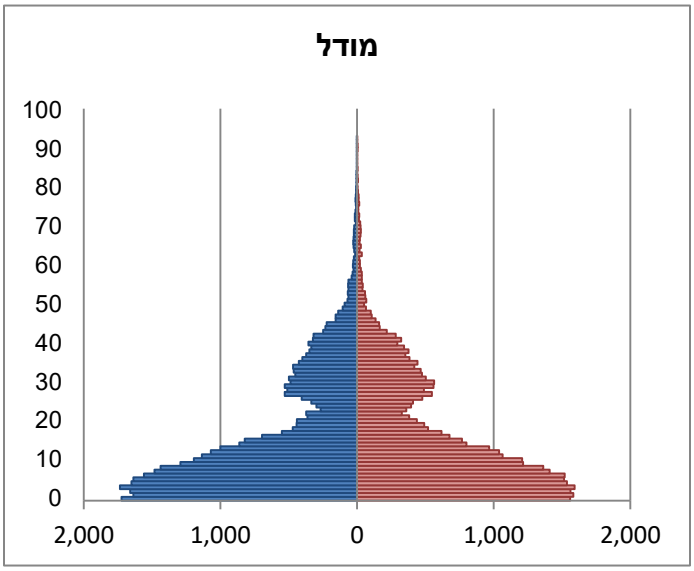
תיקוף התוצאות נעשה בשתי שיטות שונות, האחת השוואת תוצאות המודל עם נתוני סקר כוח אדם לשנת 2017, והשנייה השוואת תוצאות המודל ליישובים שידוע שיש בהם רוב חרדי מובהק. על מנת לתקף את התוצאות מול נתוני סקר כוח אדם התקבלו נתוני פרט לשנת 2017 של נדגמים שענו על שאלת רמת הדתיות במשק הבית. סה"כ התקבלו 80,188 רשומות מתוכן נותרו 61,676 רשומות המתייחסות ליהודים. 6,308 רשומות סומנו כבעלות אורח חיים חרדי (10.2%). רמות הדתיות של סקר כוח אדם קובצו לשתי רמות – חרדי ולא-חרדי. לאחר מכן הנתונים קושרו לפי מספר תעודת זהות לתוצאות המודל עבור כלל האוכלוסייה היהודית כך שלכל פרט שנדגם בסקר היה ערך רמת דתיות אמיתי וערך רמת דתיות חזוי על ידי המודל. השוואת רמת הדתיות של סקר כוח אדם לזו שהתקבלה מהמודל החזירה את התוצאות הבאות:

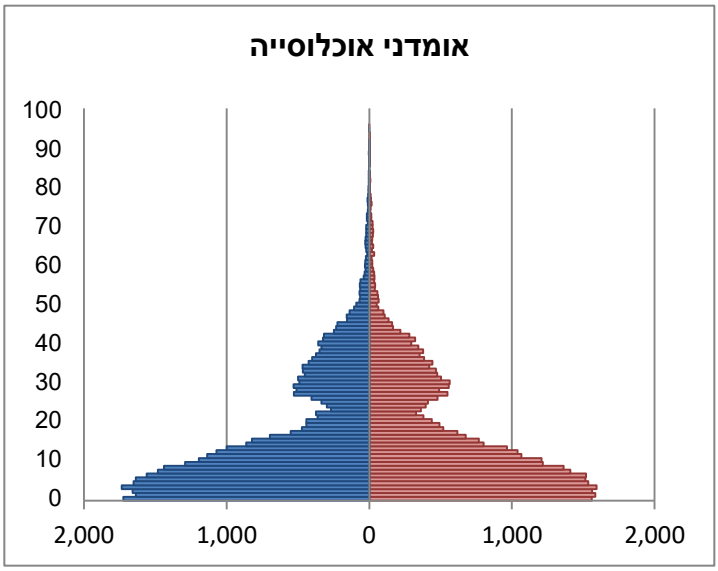
סכ"א / מודל	**לא-חרדי**	**חרדי**
לא-חרדי	54,171	869
חרדי	1,224	5,412

משמעות התוצאות היא שדיוק המודל הכולל עומד על 96.6%, ובפירוט לפי רמות דתיות המודל מזהה נכון לא-חרדים בשיעור של 97.8% ומזהה נכון חרדים בשיעור של 86.2%. תוצאות אלה מעט נמוכות יותר מהתוצאות שהתקבלו בהשוואה לנתוני מדגם האימות המבוסס על הסקר החברתי בוא נעשה שימוש לתיקוף הבסיסי של המודל, בו כאמור התקבל דיוק של 98.3% ו-88.0%, בהתאמה. בדיקת הרשומות שזוהו בטעות כחרדים מעלה כי 65% מהן הגדירו את רמת הדתיות במשק הבית כדתית או דתית מאוד.

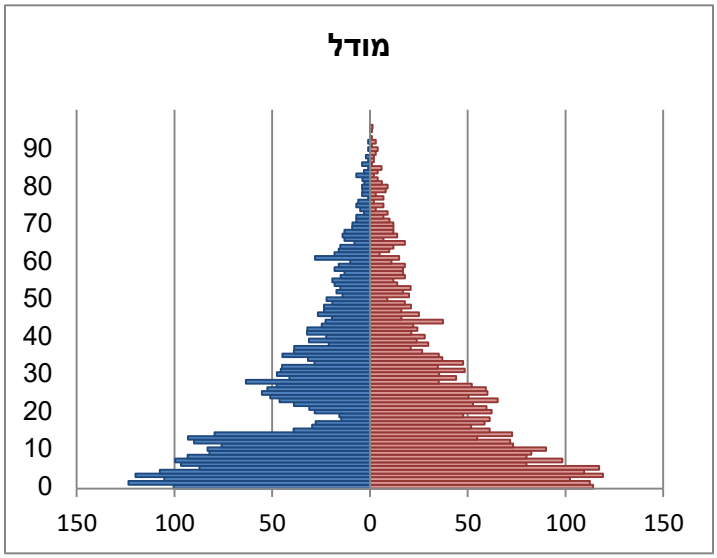
שיטת התיקוף השנייה בה נעשה שימוש היא השוואת תוצאות המודל עם אומדני האוכלוסייה הרשמיים ביישובים שידוע שיש בהן רוב חרדי מובהק ולכן נצפה שהמודל יזהה שם קרוב למאה אחוז חרדים. היישובים שנבחרו הן מודיעין עילית וכפר חב"ד, שניהם יישובים המוכרים כיישובים חרדיים. על פי אומדני האוכלוסייה הרשמיים חיו בשנת 2017 במודיעין עילית 70,081 תושבים. תוצאות המודל סיפקו אומדן אוכלוסייה של 70,071 תושבים חרדים. על פי אומדני האוכלוסייה הרשמיים חיו בשנת 2017 בכפר חב"ד 6,214 תושבים. תוצאות המודל סיפקו אומדן אוכלוסייה של 6,206 תושבים חרדים. כמו כן מתרשימים 14 ו-15 ניתן לראות את ההשוואה בין האומדנים לפי גיל ומין.

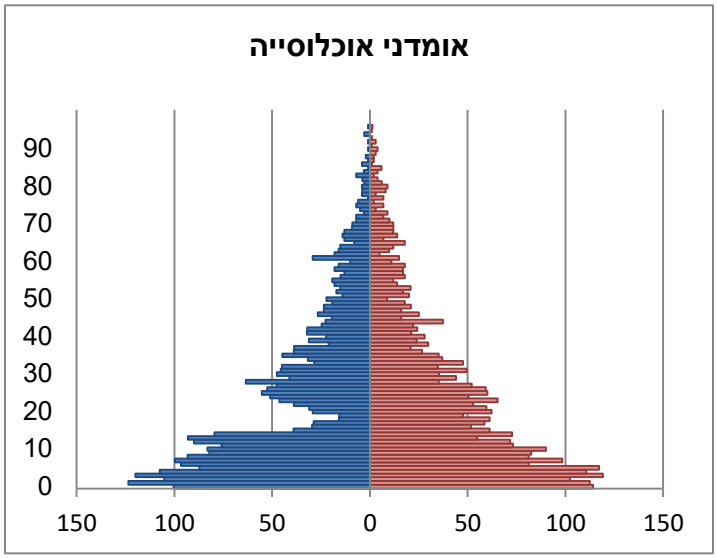
תרשים 18. השוואת פירמידות גילים עבור מודיעין עילית





תרשים 19. השוואת פירמידות גילים עבור כפר חב"ד





5. סיכום

עבודה זו הציגה מודל חדש לאמידת גודלה והרכבה של האוכלוסייה החרדית בישראל באמצעות שיטות מתקדמות של למידת מכונה. באמצעות נתונים מתוייגים מהסקר החברתי נבנה אלגוריתם אשר לימד את עצמו לזהות האם פרט הוא חרדי על בסיס מאפיינים שונים (בראשם אחוז החרדים באזור הסטטיסטי, רמת דתיות על פי מוסדות חינוך ומספר אחים). המודל בצורתו הנוכחית משיג רמת דיוק כללית של כ-97% ולאחר שהופעל על כלל האוכלוסייה היהודית מצא כי נכון לסוף 2017 האוכלוסייה החרדית מנתה 1,018,672 תושבים, שהם 15.5% מכלל האוכלוסייה היהודית ו-11.5% מקרב כלל אוכלוסיית ישראל.

המודל עונה על צורך הולך וגובר לידע מפורט על האוכלוסייה החרדית בכלל ובימי הקורונה בפרט. משרדי ממשלה וארגונים שונים זקוקים לנתונים אמינים, מדוייקים ומעודכנים עד כמה שניתן על מנת לקבוע מדיניות ולספק שירותים לאוכלוסייה החרדית. היכולת לעבוד עם מאגר נתונים מלא שמכיל את כלל האוכלוסייה החרדית בישראל ברמת הפרט יכולה לתרום רבות לצרכי מחקר ותכנון. המודל מאפשר הפקת נתונים בפירוט רב לפי משתנים דמוגרפיים כמו גיל, מין, אזור סטטיסטי ועוד, וגם מוסיף אפשרות לקישור הנתונים למקורות מידע מנהליים נוספים בתוך הלמ"ס ומחוצה לה על מנת להפיק תובנות בנושאים דמוגרפיים, חברתיים-כלכליים ועוד. עם כניסת המודל לשימוש תיבנה סדרה רב שנתית של אומדני אוכלוסייה לאוכלוסייה החרדית שתאפשר בעתיד מחקרים שיוכל לזהות מגמות ושינויים מהיבטים דמוגרפיים, כלכליים, מרחביים ועוד.

העבודה על המודל היא מתמשכת ויש להמשיך לבדוק ולתקף אותו כל שנה, כולל עדכון בסיס הנתונים כדי להמשיך ולזהות מגמות בחברה החרדית. כיוון שהטכנולוגיה מתפתחת ניתן להמשיך לשפר ולעדכן את המודל ככל שעוד נתונים ומקורות מידע יהפכו לזמינים עם הזמן.

6. **ביבליוגרפיה**

פרידמן י', שאול-מנע, נ', פוגל, נ', רומנוב, ד', עמדי, ד', פרידמן, מ', סחייק, ר', שיפריס, ג', ופורטנוי, ח', (2011). *שיטות מדידה ואמידת גודלה של האוכלוסייה החרדית בישראל*. סדרת ניירות טכניים מס' 25. ירושלים: הלשכה המרכזית לסטטיסטיקה.

פורטנוי, ח'. (2007). אפיון רמת הדתיות באוכלוסייה היהודית על פי זיקה למוסדות חינוך. סדרת ניירות טכניים מס' 19. ירושלים: הלשכה המרכזית לסטטיסטיקה.

Hayashi, Chikio (1998). "What is Data Science? Fundamental Concepts and a Heuristic Example". Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan.

Cranmer, S.J. and Gill, J.M.. (2013) “We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data.” *British Journal of Political Science* 43:2

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen & R. Rastogi (eds.), *KDD*