אוניברסיטת בר-אילן
Bar-Ilan University

סמינריון במערכות מידע

MLimpute – שימוש באלגוריתם חדשני להשלמת נתונים הנעזר
במודלים מתחום למידת המכונה לשיפור תוצאות ניבוי מחלת לב

2020

מגיש:

יונתן הופמן - 304867930

מנחה:

פרופ' און שחורי

תקציר

אחת הסיבות העיקריות כיום לתמותה בטרם עת בעולם כיום היא מחלת לב. ניבוי מחלת לב מהווה את אחד האתגרים הקריטיים בתחום עיבוד הנתונים הקליני. למידת מכונה (Machine Learning - ML) היא אחד התחומים החמים והנפוצים כיום לניבוי מחלות לב וסיוע בלקיחת החלטות, תוך היעזרות במסדי נתונים של התעשייה הרפואית. אחד האתגרים עבור אלו המבקשים להסיק מסקנות ע״י שימוש במודלים מתחום ה – ML הוא שלרוב נדרש מסד נתונים שלם וללא חוסרים על מנת להריץ את המודלים המתאימים – דרישה שלרוב לא מתקיימת במסדי נתונים במציאות. בעבודה זו אנחנו מציעים אלגוריתם חדשני להשלמת נתונים חסרים במסד נתונים נתון – נקרא לו MLimpute. האלגוריתם משתמש במודלים מתחום למידת המכונה לטובת השלמת הנתונים. נשווה את שיטת MLimpute אל מול שיטות מסורתיות ומוכרות להשלמת נתונים חסרים כמו השלמת ממוצע יחידה (Single Mean Imputation), השלמת חציון (Median) יחידה והשלמת מצב (Mode) יחידה. נבדוק את דיוק והטיית השלמת הנתונים ע״י השיטות השונות ע״י בחינת מדדי accuracy ו RMSE אל מול מסד הנתונים המקורי, ונבצע השוואה של ביצועי מודלים שונים מתחום ה ML לניבוי מחלות לב, ע״י שימוש במסדי נתונים שהושלמו ע״י השיטות שהוזכרו. נראה כי אותם מודלים שהורצו עבור מסדי נתונים שהושלמו ע״י MLimpute הראו מדדי accuracy, roc-AUC ו F-score טובים יותר מהשיטות המסורתיות שהוזכרו. בנוסף, נציע אפשרות ליישם את MLimpute בשלב שלאחר פיצול הנתונים ל train ו validation ולפני הרצת מודל החיזוי, וזאת על מנת להקטין את הסיכוי ל overfitting.

תוכן עניינים

רשימת טבלאות

רשימת תרשימים

הקדמה

הערכת סיכונים וניבוי מחלות לב מהווה פקטור משמעותי בתחום הקליני של מחקר וטיפול במחלות לב. זיהוי וטיפול נכון במחלות לב הוא עקרון קריטי, שכן איחור בזיהוי או אי דיוק בטיפול יכולים לגרום למות מטופל. בעשור הקודם השתמשו בעיקר בשיטות מבוססות רגרסיה על מנת לחזות הימצאות של מחלות לב ולחשב סיכויי תמותה עבור חולים. אך לרוב שיטות אלה מוגבלות למספר מצומצם של משתנים מסבירים, מתקשות לפענח קשרים לא לינאריים ולהתגבר על קשרים שבין משתנים מרובים (Goldstein, Navar & Carter , 2016). בנוסף, עבור הרצה ופענוח של שיטות מסוג זה נדרשת התערבות אנושית משמעותית שעולה זמן וכסף (Steele, Denaxas, Shah, Hemingway & Luscombe, 2018). כיום, בעידן שבו גובר איסוף הנתונים על חולים וקיימת התעצמות של מערכות רפואיות אלקטרוניות (EHR) יש יותר ויותר שימוש בשיטות מתחום ה ML לטובת ניבוי והערכת סיכונים בתחום הרפואי בכלל, ובתחום מחלות הלב בפרט. בשנים האחרונות הולך וגובר השימוש במודלי קלסיפיקציה כמו רשת נוירונים (NN), עצי החלטה ורגרסיה (CART) ויערות רנדומיליים (RF). בנוסף נעשה שימוש של מודלים מתחום ה Cluster Analysis כמו K-Nearest Neighbours (KNN) (Mohan, Thirumalai, & Srivastava, 2019).

כאשר אנו עוסקים במסדי נתונים, ובמיוחד כאשר מדובר בנתוני עתק, נדרש לרוב לקחת בחשבון ולהתמודד עם מצב של חוסר בנתונים ובשלמות המידע. לדוגמה, חוסר בנתוני מעבדה מסוימים אצל מטופלים שנרצה לחזות לגביהם האם צפוי להם התקף לב או מוות פתאומי. רוב המודלים המקובלים בתחום ה ML דורשים מסד נתונים מלא (Complete Case), המכיל רשומות מלאות ללא חוסרים (Bertsimas, Pawlowski & Zhuo, 2017). במרוץ השנים פותחו מספר גישות לטיפול בבעיית הנתונים החסרים: גישה אחת שהייתה נפוצה מאוד בעבר היא שימוש ברשימות מלאות בלבד והתעלמות מרשומות המכילות חוסרים (Complete Cases). גישה מסורתית נוספת היא השלמה יחידה (single imputation) ע"י שימוש בנתון סטטיסטי עבור כל משתנה כמו ממוצע, חציון או מצב (mode) (Zhang, 2016). עם השנים פותחו גישות מתוחכמות יותר להשלמת נתונים חסרים מסוג multiple imputation, אשר מקבלות כקלט מסד נתונים עם חוסרים,

ונותנות כפלט סדרה של מסדי נתונים שהחוסרים בהם הושלמו, לרוב בעזרת דרישה מהמשתמש למידע על התפלגות המשתנים (Buuren & Groothuis-Oudshoorn, 2010). דוגמאות לגישה זו הן גישות Joint Modelling (JM) ו Fully Conditional Specification (FCS) (Van Buuren, 2007). בשנים האחרונות מתפתחת גישה חדשה - להשתמש באלגוריתמים מתחום ה ML על מנת להשלים נתונים חסרים. באמצעות מודלים כאלו ניתן להשלים את מסדי המידע ללא דרישה מהמשתמש למידע מקדים אודות התפלגות הנתונים. ניתן למצוא דוגמאות ליישום גישה זו ע"י שימוש ב CART (Burgette & Reiter, 2010) ו RF (Stekhoven & Bühlmann, 2012) לטובת המשימה. השיטה המוצעת בעבודה זו – Mlimpute תומכת בגישה זו של שימוש במודלים מתחום ה ML לבעיית השלמת הנתונים. MLimpute מהווה הכללה לפתרון של בעיית השלמת הנתונים ע"י שימוש במודלים מתחום ה ML. הרעיון הוא להשתמש באלגוריתם כ"קופסא שחורה" שמקבלת כפלט מסד נתונים עם חוסרים, ונותנת כפלט מסד נתונים ללא חוסרים ע"י שימוש במודלים גמישים מתחום ה ML. כאשר משתמשים בשיטות שונות להשלמת נתונים חסרים עלולה להיווצר תופעה של overfitting (Pawlowski & Zhuo, 2017). בעבודה זו נציע ליישם את השלמת הנתונים בשלב שלאחר חלוקת הנתונים ל train ו validation על מנת למתן את התופעה.

חשיבות המחקר

כשאנו עוסקים בניבוי בשיטות מתחום ה ML, המסתמך על מסד נתונים מסוים, חוסר בנתונים והשלמתם בצורה לא מיטבית תגרום להטיות ולאובדן מידע. צוינו מספר גישות להשלמת נתונים חסרים. גישה ראשונה שצוינה היא להשתמש אך ורק בחלק של מסד הנתונים המכיל רשומות ללא חוסרים. שיטה זו גורמת להטיה בנתונים ולאובדן רב של מידע (Masconi, Matsha, Erasmus, & Kengne, 2015). גישה מסורתית נוספת היא שימוש בשיטות מסוג single mean/median/mode imputation. שיטות אלה פשוטות, אך ממעיטות בערך השונות ומתעלמות ממערכות יחסים בין משתנים (Zhang, 2016). בנוסף, כשמדובר במידע רפואי, הממוצע, החציון והשכיח חוטאים לנתונים ולא ניתן להניח כי הם משקפים ערך רלוונטי שיכול להוות תחליף לערכים ריקים ברשומות רבות. קיימות גם שיטות מתקדמות ומתוחכמות להשלמת נתונים הנקראות השלמה מרובה (multiple imputation), לדוגמה אלגוריתם GAIN ו MICE (Yoon, Jordon & Van Der Schaar, 2018). שיטת אלה נחשבות למודרניות ואפקטיביות להשלמת נתונים, אך פחות נגישות ומובנות למשתמש הממוצע. בנוסף העלות החישובית של שיטות אלה גבוהה ויישומן על נתוני עתק עלול לקחת זמן רב לחישוב. קיימים גם ניסיונות לשלב אלגוריתמים מסוג למידת מכונה בתחום השלמת הנתונים, אך לא כל אותם ניסיונות מגיעים לתוצאות מספקות (Liu & Gopalakrishnan, 2017). ומכאן חשיבות המחקר שלנו – למצוא תחליף לשיטות הקיימות ע"י שימוש באלגוריתמים מוכרים מעולם למידת המכונה לטובת השלב המקדים של השלמת הנתונים, אשר יהווה פתרון שמצד אחד מביא לתוצאות טובות יותר משיטות ה single imputation המסורתיות, ומצד שני מהווה תחליף פשוט לשיטות מודרניות מורכבות יותר.

מטרות

MLimpute הינו הפתרון שאנו מציעים בעבודה זו לבעיית השלמת הנתונים החסרים למסדי נתונים. המטרה העיקרית שלנו היא לייצר פתרון אשר משלים את הנתונים החסרים בצורה

מושכלת ומביא לכך שמודלים לניבוי מחלות לב מתחום ה ML, המקבלים מסדי נתונים שהושלמו ע״י MLimpute, יהיו בעלי ביצועים טובים יותר מאותם מודלים המקבלים מסדי נתונים שהושלמו ע״י שיטות מסורתיות מסוג single imputation. הפתרון יהיה נגיש, מובן למשתמש ומתאים לבעיית השלמת הנתונים עבור נתוני עתק. מטרה נוספת היא שהפתרון שאנו מציעים יהיה מסוגל לבצע השלמה של מסד נתונים המכיל מידע מגוון ומערב סוגים שונים של משתנים – מספריים, קטגוריאליים ובינאריים. לבסוף, נציע אפשרות למימוש האלגוריתם על מסד הנתונים בשלב שלאחר הפיצול ל train ו validation, על מנת להקטין את הסבירות ל – overfitting.

סקר ספרות

למידת המכונה הוא תחום שהסיקור שלו נמצא בתאוצה בעשור האחרון. מדי שנה מפורסמים אלגוריתמים חדשים, חדשניים, יעילים ואפקטיביים יותר מבעבר. במקטע זה נתמקד במאמרים העוסקים בתחום הטיפול המקדים במידע המגיע לאלגוריתם, ובדרכים שונות להתמודד עם בעיות ניבוי בתחום הרפואה.

במאמר שלו, Zhang (2016) עסק בדרכים להשלים נתונים חסרים במסדי מידע. המאמר מתמקד ביישום של single imputation ב R, ועוסק בשיטות פשוטות ומסורתיות להשלמת נתונים, כמו שימוש בממוצע, חציון, או ׳מצב׳ (עבור משתנה קטגוריאלי). דרך נוספת המוצגת במאמר היא השלמת נתונים באמצעות רגרסיה ליניארית עם הרעשה של הנתונים לטובת הורדת הוודאות שבהשלמת הנתונים. על מנת להציג את השיטות השונות יוצר הכותב מסד נתונים ע״י סימולציה, כאשר אין לנתונים במסד משמעות קלינית. בנוסף, אין תיעוד לרמת ההטיה והדיוק של השיטות המוצעות במאמר. המאמר שלנו בודק דרכים שונות המשתמשות בטכניקות מתחום למידת המכונה להשלמת הנתונים. בנוסף, נשתמש במסד נתונים קיים ולא מומצא. לבסוף, נעשה השוואה בין דרכים להשלמת מידע חסר המוצעות במאמר זה לאלגוריתם שנציע.

Amelia II הוצגה במאמרם של Honaker, J., King, G., & Blackwell (2011) כתוכנה פורצת דרך להשלמת נתונים. התוכנה משלימה נתונים בשיטת ההשלמה המרובה, קרי יצירת m מסדי נתונים, המכילים את כל הנתונים שלא היו חסרים במסד הנתונים המקורי ובנוסף m גרסאות שונות של השלמת נתונים עבור הנתונים שהיו חסרים. תוך שימוש באלגוריתם EMB, התוכנה משלימה את הנתונים שהיו חסרים בהתפלגות נורמלית המבטאת את חוסר הוודאות שבהשלמת הנתונים. האלגוריתם הוא מסוג EM (Expectation-Maximization). הנחת יסוד שמתקיימת היא שמסד הנתונים מורכב ממשתנים מרובים שמתפלגים נורמלית. זוהי דוגמה פרטית לתחום שנקרא Joint modeling (JM), אשר במהותו מניח התפלגות משותפת של המשתנים ומייצר פונקציית צפיפות פרמטרית למסד הנתונים. JM מהווה גם כן יישום של השלמה מרובה (multiple imputation). למודל מספר מעלות תיאורטיות שימושיות, אך לוקה בחסר בגמישות הדרושה לניתוח מסדי נתונים רבים ומגוונים (Van Buuren, 2007). המודל המוצע במחקר שלנו, המשתמש באלגוריתמים מתחום למידת המכונה, לא דורש הנחת נורמליות או התפלגות אחרת של המשתנים. בנוסף נשתמש באלגוריתמים גמישים, המסוגלים להשתמש בסוגים שונים של משתנים, וללא צורך בהשלמה מרובה על מנת להגיע למודל ניבוי (בתחום מחלות הלב) עם תוצאות טובות.

גישה נוספת להשלמת נתונים בשיטת multiple imputation היא Fully Conditional Specification (FCS). Van Buuren (2007) משווה במאמרו בין שיטת JM ל שיטת FCS. שיטת JM מגובה בהסברים תיאורטיים מוצקים, אך עלולה לסבול מחוסר גמישות כאשר נדרש ליישם את המודל על משתנים רבים מן "החיים האמיתיים", ועל כן ליצור הטיה. שיטת FCS נחשבת פרמטרית למחצה, ומגלה גמישות רבה יותר. גישה זו מטפלת במודל רב-משתני ע"י סידרה של מודלים מותנים – מודל לכל משתנה שיש להשלימו. שיטת FCS גמישה יותר וקלה למימוש, אך קיים קושי בלבסס אותה תיאורטית. במחקר שנעשה במאמר זה נאספו נתונים על 3801 בנות בגיל ההתבגרות, כאשר היו נתונים חסרים במספר משתנים, כגון ווסת, שיער ערווה והתפתחות של סרטן. במחקר זה שיטת JM יצרה הטיה, כאשר המודל של שיטת FCS לא סבל מתופעה זו.
המסקנה הייתה שמודל FCS יכול להתאים טוב יותר ממודל JM, כאשר לא ניתן לבסס הנחה של התפלגות משותפת לכל המשתנים. נטען כי מודל FCS מתאים יותר ממודל JM, כאשר מדובר בהשלמת נתונים מגוונים מסוגים שונים. אך Lee & Carlin (2010) הראו במאמרם כי לא תמיד כך המצב. במחקר שלהם החוקרים יצרו מסד נתונים המדמה מחקר אמיתי עם נתונים חסרים. מבין הנתונים החסרים היו גם משתנים בינאריים ואורדינליים, שלא ניתן לטעון לגביהם התפלגות נורמלית משותפת. החוקרים השתמשו במודלים מסוג JM, FCS ו Complete Case (מחיקת רשומות עם נתונים חסרים). הם הראו כי שיטות JM ו FCS הגיעו לתוצאות דומות, כאשר מחיקת רשומות ריקות הייתה השיטה הנחותה ביותר. קיים דמיון בין המחקר שלנו לשיטת FCS – בשני המקרים מתייחסים לכל משתנה בנפרד, ללא צורך בהנחה כוללת על התפלגות כל המשתנים יחד. המחקר שלנו יעסוק בהשוואה בין שיטת השלמת נתונים עם אלגוריתמים המשלבים למידת מכונה לשיטות מסוג single imputation. לא נתעמק בשיטות ה multiple imputation, אך ניתן לומר כי שיטות רבות מתחום למידת המכונה גמישות יותר, ודורשות פחות הנחות לגבי התפלגות הנתונים. לדוגמה, ברשת נוירונים לא נבצע הנחות על התפלגות הנתונים, אלא ניתן לאלגוריתם "להבין לבד" מה היא התלות שבין סט הנתונים המסבירים למשתנה המוסבר וע"י כך להציע ערך להשלמת החסר. בנוסף לא ניצור סט נתונים מלאכותי, מחשש שמסקנות על סט שכזה חוטאות למציאות.

אלגוריתם Multivariate Imputation by Chained Equation (MICE) הינו יישום תוכנתי של שיטת FCS. Buuren & Groothuis-Oudshoorn (2010) הציגו במאמרם את יישום התוכנה בספריית mice ב R. האלגוריתם מקבל כפלט מסד נתונים ומייצר מערך של m מסדי נתונים שהושלמו ע"י פונקציית השלמת נתונים גנרית. בנוסף ניתן לייצר מערך של ציוני תרומה מדעית לכל מערך שהושלם ו pooling סופי. על המשתמש מוטלות מספר משימות על מנת לצלוח את משימת השלמת הנתונים. נדרש לציין את תבנית המידע החסר – MAR (המידע חסר אקראית עם תלות במשתנים הבלתי תלויים) או MCAR (המידע חסר באקראית לחלוטין), באילו תכונות להשתמש לטובת ניבוי והשלמת הנתונים לשאר התכונות, האם להשלים משתנים התלויים כפונקציה של משתנים אחרים, סדר התכונות למילוי, מס' האיטרציות ואת ערכו של m – מס' מסדי הנתונים בפלט. נוסף על כך יש לבחור את השיטה הגנרית לניבוי. מבין השיטות: Predictive mean matching (PMM), Bayesian linear regression, Linear regression, Mean imputation ועוד. הספרייה מספקת אפשרויות רבות לניתוח הנתונים שהתקבלו מהאלגוריתם. במחקר שלנו השאיפה היא לייצר אלגוריתם עם מספר עקרונות דומים, אך בדגש על פשטות וידידותיות למשתמש. האלגוריתם שלנו יעבוד גם כן בצורה איטרטיבית ויפעיל מודל ניבוי גנרי על

כל תכונה בנפרד, אך באופן פשוט יותר – פחות דרישות מהמשתמש, ותוצר אחד פשוט בצורת מסד נתונים שלם וללא חוסרים (single imputation). בנוסף, המודלים לניבוי הגנריים שנשתמש בהם יהיו מתחום למידת המכונה (לדוג' רשתות נוירונים, RF).

דוגמה פשוטה יותר לספציפיקציה מותנית (Conditional Specification) היא Least-Squares specification (LS). שיטה זו מייצרת השלמת נתונים יחידה (single imputation) ע"י סט של רגרסיות פר משתנה. Bø, Dysvik & Jonassen (2004) השוו במחקרם את שיטת LS אל מול שיטות מסוג EM ו KNN. המטרה הייתה להשלים נתונים במיקרו-מערכים של נתונים בנושא מחקר גנים. לצורך המחקר הם יצרו שלושה מסדי נתונים ללא חוסרים והוציאו רנדומלית נתונים. כדי להגיע למסדי נתונים ללא חוסרים הם השתמשו במסד נתונים מתחום חקר הגנים והסרטן הנקרא NCI60, וגזרו מתוכו רשומות ללא חוסרים כלל. אל מול KNN שיטת LS הגיעה לתוצאות מדויקות יותר. בנוסף מצאו כי תוצאות LS היו מדויקות לפחות כמו שיטת EM. על מנת ליישם את שיטת LS החוקרים ביצעו רגרסיות על שני סוגים של סטים של נתונים – gene ו array. בכל רגרסיה בוצע bootstrapping, כאשר העיקרון המנחה לכל רגרסיה הוא הקטנת סכום ריבועי הטעויות (least squares). נבדקה קוראלציה ובסופו של דבר התוצאה הסופית להשלמת הנתונים הייתה ממוצע משוקלל של שתי הרגרסיות. החוקרים בדקו RMSE בין התוצאות החזויות למקוריות על מנת להעריך את ביצועי המודל. במחקר שלנו נשתמש בצורה דומה בשיטת bootstrapping על מנת להגיע לערך ממוצע משוקלל בשיטת single imputation. אחת הדרכים להעריך את ביצועי השלמת הנתונים תהיה גם כן ע"י מדידת RMSE לעומת הנתונים המקוריים במסד הנתונים, אך גם נרצה להעריך את שיטת השלמת הנתונים ע"י בחינת ביצועי מודל לניבוי מחלות לב. לגבי שיטת איסוף הנתונים – אנו סבורים כי לקחת נגזרת של מסד נתונים המכילה רק רשומות ללא חוסרים כפי שבוצעה במחקר המדובר אינה מיטבית וחוטאת למטרת המחקר. בכלל, המוטיבציה לביצוע imputation לנתונים נובעת מכך שלקיחת נגזרת של מסד נתונים ללא חוסרים לצורך מחקר (נקראת גם full case) גורמת להטיות בנתונים ומגיעה לתוצאות נחותות לעומת גישות ה imputation השונות (Zhang, 2016; Buuren & Groothuis-Oudshoorn, 2010).

שיטות נוספות אשר משתמשות ברגרסיה תחת conditional specification בדומה ל LS הן Predictive-Mean Matching (pmm) ו Support Vector Regression (SVR). ההבדל בין שיטת pmm לשיטת LS הוא שבשיטת pmm הערך שמתקבל להשלמת הנתונים נלקח באקראית מתוך קבוצת ערכים אשר ערכיה קרובים לערך הרגרסיה (Buuren & Groothuis-Oudshoorn, 2010). לעומת זאת ב SVR הערכים נבחרים ע"י הרגרסיה, רק שכאן אופי הרגרסיה שונה מ LS (Wang, Li, Jiang & Feng, 2006).

כאשר הקשרים בין המשתנים אינם לינאריים, שיטות המבוססות על רגרסיה לינארית מראות ביצועים נחותים. Burgette & Reiter (2010) הציעו להשתמש בעצי רגרסיה וקלסיפיקציה בשיטת CART. המחקר שלהם מציע לבצע השלמת מידע מסוג multiple imputation באלגוריתם איטרטיבי הדומה לשיטת MICE, רק שהמודל להשלמת הנתונים יבוצע על ידע עצי קלסיפיקציה או רגרסיה. החוקרים יצרו מסד נתונים מלאכותי עם התפלגויות מוגדרות לכל משתנה והראו כי יישום של MICE עם מודל CART הגיע לתוצאות טובות יותר מ MICE מסורתי. אחד היתרונות במודל CART על פני שיטות קודמות שצוינו נובע מכך שאין צורך להניח התפלגות למשתנים השונים. צוין במאמר כי החסרונות העיקריים של CART נובעים מעלות חישובית גבוהה, בעיקר

בנוכחות משתנים קטגוריאליים עם מס׳ קטגוריות רב ויצור של עצים מסורבלים לפיענוח. בנוסף אחת המסקנות הנובעות מהמאמר כי מומלץ לבחון שיטות נוספות, אקזוטיות יותר כמו רשתות נוירונים, יערות רנדומליים ועוד. Stekhoven & Bühlmann (2012) הראו במאמר שלהם הרחבה ל CART ע״י שימוש ביערות רנדומליים. המחקר שלנו למעשה עונה על המסקנה מהמחקר על CART וממשיכה את כיוון המחקר לאלגוריתמים נוספים מתחום למידת המכונה. המחקר שלנו יממש שיטת השלמת נתונים איטרטיבית, בדומה לשיטת MICE תוך שימוש במודלים שונים מתחום למידת המכונה כדי להשלים את הנתונים החסרים במסד הנתונים. אנחנו נציע מימוש של single imputation תחת ההנחה כי שיטה זו יותר ידידותית ופשוטה למשתמש.

שיטה א-פרמטרית נוספת ומאוד פופלרית להשלמת נתונים היא K-Nearest Neighbors imputation (KNN). Troyanskaya et al. (2001) הראו במחקרם כי KNN יכולה להיות יותר מדוייקת ורגישה כשמדובר בהשלמת מסד נתונים בתחום המחקר הגנטי. השיטה משלימה כל משבצת ריקה כממוצע של המימד הרלוונטי של K השכנים הקרובים שנצפו להם ערכים באותו מימד. לשיטה זו מספר וריאציות, ובאחת מההרחבות השתמשו ב Sequential KNN. לפי שיטה זו מתחילים להשלים את הנתונים במשתנה בעל מספר הערכים החסרים הנמוך ביותר, וממשיכים להשלים את המימדים הבאים על סמך הדאטא שהושלם עד כה (Kim, Kim & Yi, 2004). Iterative KNN משתמש בתהליך איטרטיבי כדי לדייק ולבחור את K השכנים הקרובים על סמך הערכים שהתקבלו מאיטרציה קודמת (Brás, & Menezes, 2007). במאמר שלנו נציע בצורה דומה אלגוריתם סידרתי ואיטרטיבי המשלים מידע חסר תכונה אחר תכונה. גם הפתרון שלנו מהווה פתרון א-פרמטרי ומייצר השלמת נתונים מסוג single imputation. המחקר שלנו לא עוסק במסד נתונים מתחום הגנטיקה ולא מיישם את שיטת KNN, אלא משתמש במודלים לניבוי מתחום למידת המכונה כדי להשלים נתונים מתחום מחלות הלב.

Yoon, Jordon & Van Der Schaar (2018) הציגו אלגוריתם מבטיח להשלמת נתונים. אלגוריתם GAIN הינו אלגוריתם מודרני ומתקדם, ובהשוואה לאלגוריתמים חדישים מגיע לתוצאות טובות יותר. במאמר מוצגות תוצאות תחת הנחת MCAR לנתונים חסרים, קרי הנתונים חסרים בצורה אקראית לחלוטין, ללא תלות בסוג המשתנים. ההשוואה בין האלגוריתמים מתבצעת על 5 מסדי נתונים ממאגרי המידע של UCI ללמידת מכונה (Lichman, 2013), בעלי מאות אחדות של נתונים כ״א. ההשוואה בין האלגוריתמים התבצעה על פני מספר מישורים. התבצעה בדיקה איכותנית וכמותית על ביצועי האלגוריתם בהשלמת מסדי הנתונים, תחת תנאים שונים (לדוגמה שינוי ביחס הנתונים החסרים מסך כל הנתונים). המדד להצלחה היה RMSE. בנוסף, נבדקו תוצאות אלגוריתם של רגרסיה לוגיסטית לאחר השלמת נתונים בדרכים שונות. כאן להערכת המודל מדדו AUROC. הבעיה באלגוריתם GAIN היא שכרגע לא קיים מימוש פשוט בתוכנות מוכרות כמו R ו PYTHON. בנוסף, לא הייתה התייחסות במאמר לגבי העלות החישובית של האלגוריתם, אשר מורכב ממספר רב של צעדים. הפתרון האלטרנטיבי המוצע במאמר שלנו מתבסס על אלגוריתמים ידועים ומוכרים מתחום למידת המכונה, כמו רשת נוירונים או עצי החלטה, ומשתמשת במימושים קיימים ב R על מנת להשלים מידע חסר. באופן דומה יתבצע על מאגרים ממאגר UCI, אך ההשוואה תהיה לדרכים מסורתיות להשלמת נתונים כמו השלמה ע״י ממוצע או חציון.

הצעה מעניינת לאלגוריתמים מתחום למידת המכונה ניתן למצוא במאמר של Bertsimas, Pawlowski & Zhuo (2017). במאמרם, הציעו סדרת אלגוריתמים שנקראה באופן כללי

opt.impute. המחקר נעשה על 84 מאגרי מידע מ UCI Machine Learning Repository ועל דרגות שונות של מידע חסר – 10%, 30% ו 50%. האלגוריתם שמוצע משתמש במודלים מתחום למידת המכונה – KNN, SVM regression ו Classification tree. קיימת אפשרות לשימוש בהשלמה יחידה (single) או מרובה (multiple) כתלות בגודל מסד הנתונים. החוקרים ניסחו את בעיית השלמת הנתונים כבעיית אופטימיזציה. החוקרים השוו את ביצועי opt.impute למספר אלגוריתמים – השלמת ממוצע, KNN, iterative KNN, Bayesian PCA, ו pmm. המודלים נבדקו במישור דיוק השלמת הנתונים ובתוצאות הרצת מודל לחיזוי על מסדי הנתונים שהושלמו. התוצאות הראו כי בכ 80% מהמקרים התוצאות של אלגוריתם opt.impute היו עדיפות. במחקר שלנו נציע פתרון חלופי ל opt.impute כדרך לשימוש במודלים לחיזוי מתחום למידת המכונה לטובת בעיית השלמת הנתונים. המודל שלנו יהיה בשאיפה פשוט יותר למימוש. בנוסף, נתמקד בתחום ניבוי מחלות הלב. מימד נוסף שהמודל שלנו יעסוק בו הוא האפשרות לביצוע השלמת נתונים לאחר שלב פיצול מסד הנתונים ל train ו validation. השיטה שלנו תתמוך בפתרון מסוג השלמה יחידה ולא בהשלמה מרובה, אם כי לא נשלול את האפשרות להרחבה בעתיד.

ראינו עיסוק בניבוי רפואי ע״י למידת למידה במאמר של Goldstein, Navar & Carter (2016). המאמר עוסק בהבדלים בין שיטות שונות לניבוי סכנה לתמותה אצל חולים שסבלו מהתקף לב. המאמר עוסק בהשוואה בין שיטות רגרסיה ולמידת מכונה על מאגר מידע רפואי EHR מהמרכז הרפואי של אוניברסיטת Duke מצפון קרוליינה, ארה״ב. כדי להתמודד עם ערכים חסרים בתוצאות מעבדה השתמשו בשיטת השלמת הממוצע בערכים החסרים. הם מצאו כי לאלגוריתמים מתחום למידת המכונה יש יתרונות בניבוי לעומת רגרסיות פשוטות. גם במאמר שלנו יש עיסוק בניבוי של מחלות לב ע״י רגרסיות ואלגוריתמים של למידת מכונה, אבל בניגוד להשלמת הנתונים החסרים ע״י השלמת הממוצע לכל ערך, המיקוד של המאמר עוסק בשימוש בלמידת מכונה להשלמת הנתונים החסרים, ובדיקת ההשפעה על אלגוריתם הניבוי.

Mohan, Thirumalai, & Srivastava (2019) הציעו אלגוריתם משלהם מתחום למידת המכונה לניבוי מחלות לב. המאמר סוקר שיטות שונות לניבוי מתחום למידת המכונה. צוין כי אלגוריתם רשתות נוירונים מהווה מודל לניבוי מחלות לב עם תוצאות טובות באופן יחסי. המאמר מציג אלגוריתם היברידי המשלב יער רנדומלי עם מודל לינארי (HRFLM). החוקרים השתמשו במסד הנתונים Cleveland UCI repository data set, ובשלב ה preprocess הורדו מופעים בודדים עם נתונים חסרים. בוצע feature selection על מנת לטייב את תוצאות המודל. החוקרים הראו שהמודל שלהם מגיע לתוצאות טובות לעומת מודלים אחרים. במחקר שלנו נשתמש גם כן במסד הנתונים Cleveland UCI repository, אך המיקוד במחקר שונה. אנו נתמקד בשלב ה preprocessing ונבחן שיטות שונות להשלמת נתונים חסרים. לא נתמקד במציאת האלגוריתם הטוב ביותר לניבוי מחלת לב, אלא נשתמש באותם מודלים ככלי להשוואה בין השיטות השונות להשלמת הנתונים. בנוסף, לא נבצע feature selection ונשתמש בכל 13 המשתנים על מנת שתהיה אחידות בין ההרצות.

המאמר של Steele et al. (2018) משווה בין מודלים מתחום למידת המכונה לבין מודלים מסורתיים בתחום הניבוי לסיכויי תמותה לחולים במחלת לב כלילית. המחקר מבוצע על מידע רפואי אלקטרוני הכולל 80000 מטופלים. תחילה, מבוצעת השוואה בין מודל COX ל random survival forests עם וללא השלמת נתונים חסרים, על 27 משתנים נבחרים. בהמשך, השתמשו

במודל COX, random forests ו elastic net regression על מאגר מידע מורחב עם 586 משתנים ללא עיבוד מוקדם של מומחים. הערכת המודלים בוצעה ע״י C-index ו calibration score הנובע מהדיוק של המודלים. נמצא כי מודלים מונחי מידע בעלי ביצועים טובים יותר לעומת המודלים המסורתיים כשלא בוצע עיבוד מקדים של מאגר המידע. בנוסף, מודלים מתחום מכונת הלמידה מאפשרים הסקה חדשנית לגבי טיפול המשך, הסקה לגבי משמעות הנתונים החסרים ומציאת קשרים לא ליניאריים. כאשר בוצעה השלמת מידע היא בוצעה באמצעות אלגוריתם MICE בתוכנת R. לגבי משתנים קטגוריאליים חסרים – בוצעה הוספה של משתנה דמה המעיד על הימצאות/חיסרון המשתנה. נמצא כי לאי המצאות של משתנים יכולה להיות משמעות על רמת הסיכון לתמותה, ככל הנראה מהסיבה להמצאות/חוסר המידע. בעצם המידע החסר במסד הנתונים הוא מסוג MAR (חסר בצורה מקרית), קרי החוסר בנתונים תלוי בנתונים הנמדדים. להשלמת המידע החסר הייתה משמעות נמוכה בהיבט ביצועי המודלים, יתכן ונובע מכך שרק ל 6 משתנים היו ערכים חסרים, חלקם לא נחשבים כמשמעותיים. כותבי המאמר האירו שמצד אחד השלמת מידע על כל מסד נתונים (בטרם פיצול לסט אימון ואימות) מגבירה את ההטיה של הנתונים, ומצד שני לא קיים ישום של השלמת מידע בנפרד לסט האימון והאימות. המחקר שלנו יעסוק בשימוש באלגוריתמים מתחום למידת המכונה לניבוי מחלת לב, ולא יבדוק שיטות מסורתיות. במאמר שלנו ננסה לתת פתרון לבעיית השלמת הנתונים בצורה נפרדת לסט האימון והאימות. בנוסף, נשתמש במגוון אלגוריתמים שלא קיבלו התייחסות כמו רשת נוירונים.

Masconi, Matsha, Erasmus, & Kengne (2015) עסקו גם כן בהשוואה בין שיטות שונות של השלמת נתונים והשפעתן על ביצועי מודל חיזוי בתחום רפואי – אבחון סוכרת. המיקוד היה סביב שיטות השלמה יחידה (single) ומרובה (multiple). בוצעה השוואה בין 5 מודלים שונים לטיפול בנתונים חסרים – מחיקת רשומות בעלות מידע חסר, השלמת ממוצע יחידה, השלמת ממוצע מותנית, רגרסיה סטוכסטית, ושיטת MICE. המחקר בוצע על נתוני אמת שנאספו בדרום אפריקה על כ 1300 תושבי, כאשר לכ 30% מהרשומות היה מידע חסר. נבדקה ההשפעה של השיטות השונות על 5 מודלי ניבוי לסוכרת – המודל של קיימברידג׳, המודל הכווייתי, המודל העומאני, מודל הניבוי מרוטרדם 1 והמודל הפיני המפושט לניבוי סוכרת. נבדקו מדדי הערכה שונים, כמו calibration, brier score, Yates score ו C-statistic. מסקנות המחקר היו שהתוצאות הגרועות ביותר נמצאו עבור השיטה של מחיקת רשומות עם חוסרים. בנוסף, בעוד ששיטות ה multiple imputation (כמו MICE) נחשבות למתקדמות ומדויקות יותר, לא נמצאו הבדלים מהותיים בינן לבין השיטות הפשוטות כמו השלמת ממוצע פשוטה. המחקר שלנו עוסק בהשוואה בין שיטות פשוטות כמו השלמת ממוצע, לבין שימוש באלגוריתמים מוכרים מתחום למידת המכונה. בנוסף, נשתמש באלגוריתמים מסוג דומה עבור מודל הניבוי של המשתנה המוסבר במחקר. התחום הרפואי שנעסוק בו הוא חיזוי מחלת לב.

Liu & Gopalakrishnan (2017) בחנו שיטות שונות להשלמת נתונים, על מנת לנצל בצורה טובה יותר מאגרי מידע לניבוי בתחום הרדיולוגיה של כלי הדם. הם ביצעו השוואה בין 4 שיטות להשלמת נתונים: השלמת ממוצע יחיד, עץ החלטה, KNN ו self-organized maps. לאחר מכן השתמשו במסד הנתונים להרצת אלגוריתם BRL לניבוי מחלת לב. את הנתונים קיבלו מארכיב הנתונים הרפואיים של אוניברסיטת פיטסבורג. הם השלימו את הנתונים תוך הנחת MNAR (החוסר בנתונים תלוי במשתנה התלוי). הם מצאו כי ככל ששיעור הנתונים החסרים גדול יותר, כך

ביצועי המודלים הפשוטים הקיימים להשלמת מידע פחות מדויקים ורובסטיים. במאמר שלנו נעסוק גם כן בשימוש באלגוריתמים מתחום למידת המכונה לטובת השלמת הנתונים. השיטה שאנו מציעים עובדת תחת הנחת MCAR (הנתונים חסרים בצורה אקראית לחלוטין) ולכן לא דורשים ידע מקצועי ספציפי בניתוח מקדים של התכונות השונות על מנת לטייב את תוצאות האלגוריתם. בנוסף, נשתמש באלגוריתמים שונים מתחום למידת המכונה על מנת להשלים את הנתונים ואופן יישומם יהיה שונה.

הגדרת שאלת המחקר

המחקר שלנו מבקש לערוך השוואה בין ביצועי אלגוריתם MLimpute לבין השיטות המסורתיות מסוג single imputataion. ראינו כי לאופן השלמת הנתונים יכולה להיות השפעה על איכות מסד הנתונים שמתקבל ועל רמת ההטיה של הנתונים. בראש ובראשונה נרצה לדעת האם שימוש ב MLimpute יכול לגרום לשיפור ביצועים של מודלי חיזוי מחלות לב. בנוסף, נרצה לבחון האם שחזור מסד נתונים ע״י MLimpute מדויק יותר משחזור אותו מסד הנתונים ע״י שיטות השלמת נתונים מסורתית מסוג single imputation. לבסוף, נרצה לבחון היתכנות עבור ביצוע השלמת נתונים ע״י אלגוריתם MLimpute בשלב שלאחר הפיצול של הנתונים ל train ו validation:

1. השלמת נתונים חסרים: האם עיבוד מקדים בעזרת למידת מכונה, Mlimpute, יכול לשפר אלגוריתמים לחיזוי מחלת לב, תוך השוואה לשיטות השלמת נתונים חסרים מסורתית מסוג single imputation?
2. האם שחזור מסד נתונים ע״י MLimpute יותר מדויק משחזור ע״י שיטות השלמת נתונים חסרים מסורתית מסוג single imputation?
3. האם ניתן ליישם את שיטת MLimpute עבור השלמת נתונים לאחר פיצול הנתונים ל train ו validation, טרם הרצת מודל מתחום ה ML לחיזוי מחלת לב?

שיטת מחקר

במחקר שלנו נציג את אלגוריתם MLimpute, הנותן פתרון לבעיית השלמת הנתונים ע״י שימוש באלגוריתמים מתחום למידת המכונה. נציע שני מימושים לאלגוריתם – מימוש אחד ע״י שימוש ב – NN, ומימוש נוסף ע״י שימוש ב – RF. נקרא להם MLimpute.nn ו MLimpute.rf בהתאמה. על מנת לבחון את תפקוד האלגוריתמים המוצעים נשתמש במאגר נתונים Cleveland Heart Disease המפורסם ב UCI Machine Learning Repository. מסד הנתונים מכיל 13 משתנים בלתי תלויים, מסוגים שונים (מספרי, קטגוריאלי ובינארי), המכיל מידע רפואי ומשתנה אחד תלוי – האם קיימת מחלת לב או לא. במסד הנתונים קיימות כ – 300 רשומות מלאות (Lichman, 2013). הנתונים יהיו חסרים תחת הנחת MCAR, קרי הנתונים יהיו חסרים בצורה אקראית לחלוטין, כאשר בהרצות השונות נבחר את שיעור הנתונים החסרים מתוך המאגר המידע השלם. נבחן את האלגוריתמים המוצעים אל מול שיטות השלמה יחידה (mean, median, mode), כאשר ההרצות יבדלו במספר מישורים:

1. אחוז הנתונים החסרים – נבצע השוואה של ביצועי האלגוריתמים תחת דרגות שונות של החסרת נתונים: 10/20/30 אחוזים מן הנתונים יוחסרו. זאת על מנת לראות ולהציג האם

קיימת מגמה הקשורה לביצועי המודלים להשלמת הנתונים ולניבוי מחלת לב ולשיעור הנתונים החסרים. נרצה לראות האם יש מודלים שביצועיהם נפגעים יותר ככל ששיעור הנתונים החסרים גדול יותר.

2. מסדי נתונים שונים בכל דרגה של נתונים חסרים – בכל דרגה של נתונים חסרים ניצור בצורה אקראית 40 גרסאות שונות של מסדי נתונים. נבדוק האם כאשר לוקחים מדגם רחב של מסדי נתונים ניתן לראות הבדלים משמעותיים בין המודלים ובכך להגדיל את אמינות התוצאות.
3. האלגוריתם הנבחר להשלמת הנתונים – עבור כל מסד נתונים חסר נבצע השלמת נתונים ע״י האלגוריתמים: MLimpute.nn, MLimpute.rf, Single mean imputation ו single median imputation, כאשר עבור 2 השיטות האחרונות נשתמש ב single mode imputation עבור משתנים קטגוריאליים. למעשה בכך אנו יוצרים השוואה (benchmarking) בין MLimpute לבין השיטות המסורתיות מסוג single imputation.
4. המודל מתחום למידת המכונה הנבחר לניבוי מחלת הלב – עבור כל מסד נתונים שהושלם ע״י אחד מהאלגוריתמים נשתמש במודלים הבאים לניבוי מחלת לב – RF, NN, ורגרסיה לוגיסטית (LR). המודל הנבחר לניבוי מחלת לב אינו במוקד המחקר שלנו, אולם נרצה לראות האם ניתן לצפות בהבדלים מהותיים בביצועי המודלים לניבוי, כאשר אותם מודלים מקבלים מסדי נתונים דומים, שהנתונים החסרים הושלמו בשיטות שונות ובכך ניתן להשוות ביניהן.
5. חלוקות שונות ל train ו validation – כל מודל לניבוי נריץ 40 פעם, כאשר כל הרצה תיבדל ע״י חלוקה שונה ל train ו validation. כך עבור כל מודל לניבוי המקבל מסד נתונים שהושלם בשיטה מסוימת נבטיח שהממצאים שקיבלנו עבור ביצועי המודל לא תלויים בחלוקה ספציפית ל train ו validation אלא מהווים ממוצע של מדגם של 40 חלוקות שונות.
6. עיתוי השלמת הנתונים – נבצע בדיקת היתכנות עבור ביצוע השלמת נתונים לאחר פיצול ל train ו validation ע״י הרצה של MLimpute.rf עבור מסדי הנתונים החסרים בדרגה של 30% ובדיקה של ביצועי מודלי הניבוי למחלת לב עבור אותם מסדי נתונים שהושלמו.

את הערכת ביצועי האלגוריתמים להשלמת נתונים חסרים נבצע בשני מישורים:

1. איכות השלמת הנתונים – גודל הטעות של השלמת הנתונים אל מול המאגר המקורי במונחי RMSE/accuracy. נרצה לראות האם קיים הבדל בין השיטות השונות להשלמת הנתונים ברמת ההטיה שלהן. עבור משתנים מספריים שהושלמו נבדוק את מדד ה RMSE ועבור משתנים קטגוריאליים accuracy.
2. ביצועי המודל לניבוי מחלת לב – accuracy, roc-AUC, f-score. נאסוף עבור כל שיעור נתונים חסרים, שיטת השלמת נתונים ומודל לניבוי מחלת לב את ממוצע המדדים שצוינו.

את המודלים השונים מימשנו באמצעות תוכנת R.

אלגוריתם MLimpute:

1. דרג את התכונות לפי שיעור הנתונים החסרים בכל תכונה.

2. השלם זמנית את כל התכונות לפי שיטת השלמת ממוצע יחידה.
3. כל עוד לא הושלמו באופן סופי כל התכונות:
   a. בחר את התכונה שטרם הושלמה סופית עם שיעור החוסרים הכי נמוך, נסמנה a.
   b. בחר את a כמשתנה מוסבר עבור מודל ML נבחר, שאר התכונות יהוו המשתנים המסבירים עבור המודל.
   c. הרץ את מודל החיזוי והשלם את הנתונים החסרים ב a על סמך תוצאות המודל (החלף את הנתונים שהושלמו באופן זמני ע״י שיטת הממוצע בנתונים שהתקבלו במודל החיזוי).
   d. סמן את a כתכונה שהושלמה באופן סופי.

תוצרים מהמחקר

ביצענו את ההרצות של המודלים השונים על מסד הנתונים Heart Disease data set מתוך UCI Machine Learning Repository. עבור כל שיעור נתונים חסרים (10%/20%/30%) יצרנו באקראי 40 מסדי נתונים חסרים. עבור כל מסד נתונים חסר ביצענו השלמת נתונים ע״י השיטות השונות, ביצענו 40 חלוקות שונות ל train ו validation והרצנו את המודלים השונים לניבוי מחלת לב. בנוסף, עבור מסדי הנתונים החסרים בשיעור 30% שיצרנו ביצענו פיצול ל train ו validation בטרם הושלמו הנתונים, ביצענו השלמת נתונים ע״י MLimpute.rf עבור מסדי הנתונים המפוצלים, והרצנו מודלים לניבוי מחלת לב. נראה בתוצאות שאספנו כי מבחינת מדדי ההטיה RMSE ו accuracy כאשר משווים בין מסדי הנתונים שהושלמו בשיטות השונות אל מול מסד הנתונים המקורי השיטה MLimpute.rf הגיעה לתוצאות הטובות ביותר. עבור השלמת הנתונים הקטגוריאליים MLimpute.nn הייתה עדיפה על השיטות המסורתיות של ה single mean, אולם במשתנים המספריים לא היה הבדל משמעותי. כאשר משווים את ביצועי המודלים השונים לניבוי מחלת לב במונחים של accuracy, roc-AUC ו f-score ניתן לראות הבדלים מהותיים לטובת MLimpute אל מול שיטות המסורתיות של ה single imputation. בנוסף, ראינו בהדגמה של ביצוע השלמת הנתונים לאחר הפיצול ל train ו validation שאין פגיעה מהותית בביצועי המודל לניבוי מחלות לב בהשוואה להשלמת הנתונים טרם שלב הפיצול.

הטיה בהשלמת הנתונים בשיטות השונות

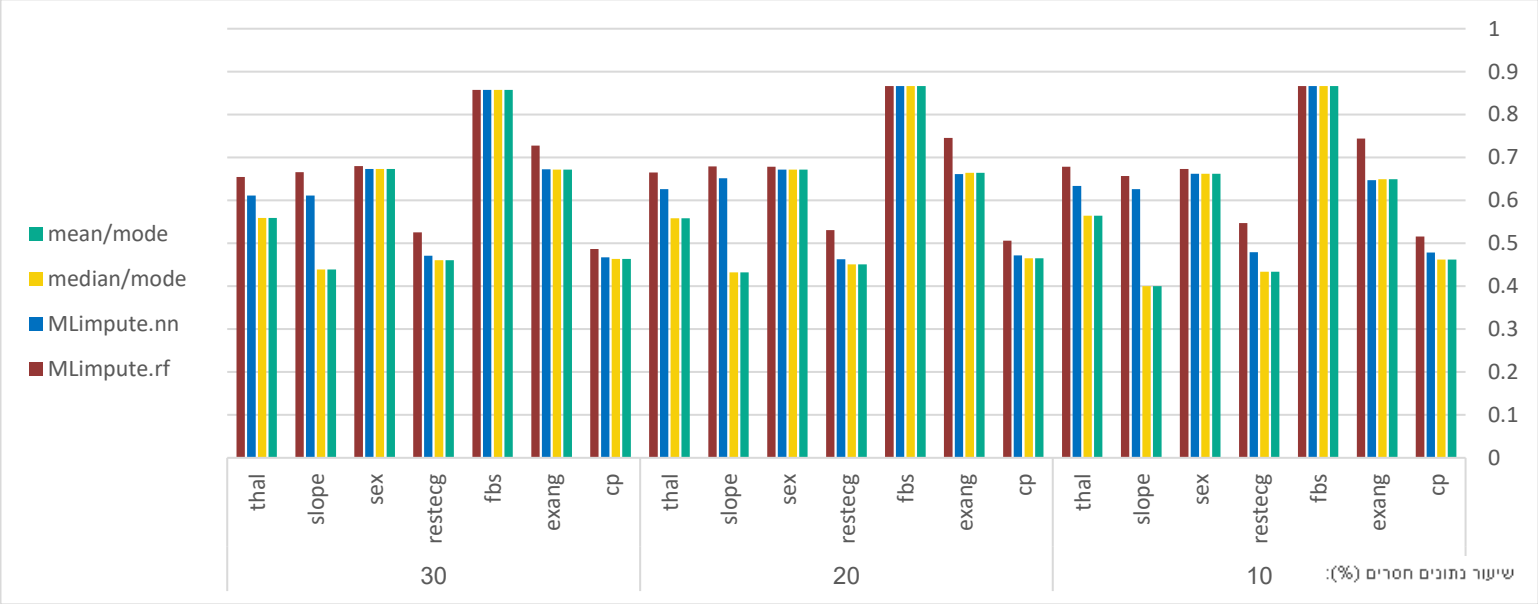
בטבלה 1 ניתן לראות את מידת ההטיה הממוצעת שנמדדה עבור כל שיעור נתונים חסרים, לכל שיטת השלמת נתונים ולכל תכונה בנפרד. התכונות הקטגוריאליות (או בינאריות) הינן sex, cp, fbs, restecg, exang, slope ו thal. עבור השלמת תכונות אלה מדדנו את ההטיה במונחי accuracy. התכונות המספריות הינן age, trestbps, chol, thalach, oldpeak ו ca. עבור השלמת תכונות אלה מדדנו את ההטיה במונחי RMSE. בתרשים 1 ניתן לראות כי עבור משתנים קטגוריאליים, שיטות MLimpute היו עדיפות על השיטות המסורתיות מסוג single mean בדיוק בהשלמת הנתונים, כאשר MLimpute.rf הגיעה לרמת הדיוק הגבוהה ביותר. גם עבור המשתנים המספריים MLimpute.rf הגיעה לרמת דיוק הגבוהה ביותר, אולם היו תכונות ש MLimpute.nn השלימה בצורה פחות מדויקת מהשיטות המסורתיות, זאת ניתן לראות בתרשים 2.

טבלה 1 – מדדי הטיה ממוצעים לכל תכונה עבור שיטות השלמת הנתונים השונות ביחס למסד הנתונים המקורי, מחולק לפי שיעור נתונים חסרים במסדי הנתונים. עבור משתנים מספריים המדד המוצג הוא RMSE ועבור משתנים קטגוריאליים ובינאריים המדד המוצג הוא accuracy.

RMSE						accuracy							מדד	
trestbps	thalach	oldpeak	chol	ca	age	thal	slope	sex	restecg	fbs	exang	cp	תכונה / מודל השלמה	שיעור חוסרים (%)
18.081	20.121	0.976	56.818	0.961	7.872	0.633	0.626	0.662	0.479	0.866	0.647	0.479	MLimpute.nn	10
17.032	18.403	0.959	50.809	0.947	7.637	0.678	0.657	0.673	0.547	0.866	0.744	0.516	MLimpute.rf	
17.524	23.026	1.136	51.582	1.018	8.963	0.564	0.400	0.662	0.434	0.866	0.649	0.462	mean/mode	
17.526	23.245	1.153	51.942	1.185	9.014	0.564	0.400	0.662	0.434	0.866	0.649	0.462	median/mode	
18.319	20.836	0.999	54.353	0.948	8.040	0.626	0.652	0.671	0.463	0.866	0.661	0.472	MLimpute.nn	20
17.556	18.934	0.975	50.322	0.942	7.930	0.665	0.679	0.678	0.531	0.866	0.746	0.506	MLimpute.rf	
17.984	22.801	1.142	50.703	1.000	9.002	0.558	0.433	0.672	0.451	0.866	0.664	0.465	mean/mode	
18.014	23.016	1.165	51.166	1.171	9.054	0.558	0.433	0.672	0.451	0.866	0.664	0.465	median/mode	
18.569	21.076	1.008	54.466	0.938	8.286	0.612	0.611	0.673	0.471	0.857	0.673	0.467	MLimpute.nn	30
17.467	19.471	1.002	51.205	0.940	8.057	0.654	0.666	0.680	0.525	0.857	0.728	0.487	MLimpute.rf	
17.623	22.715	1.143	51.269	0.994	8.935	0.559	0.439	0.673	0.461	0.857	0.672	0.463	mean/mode	
17.639	22.988	1.168	51.579	1.162	9.007	0.559	0.439	0.673	0.461	0.857	0.672	0.463	median/mode	

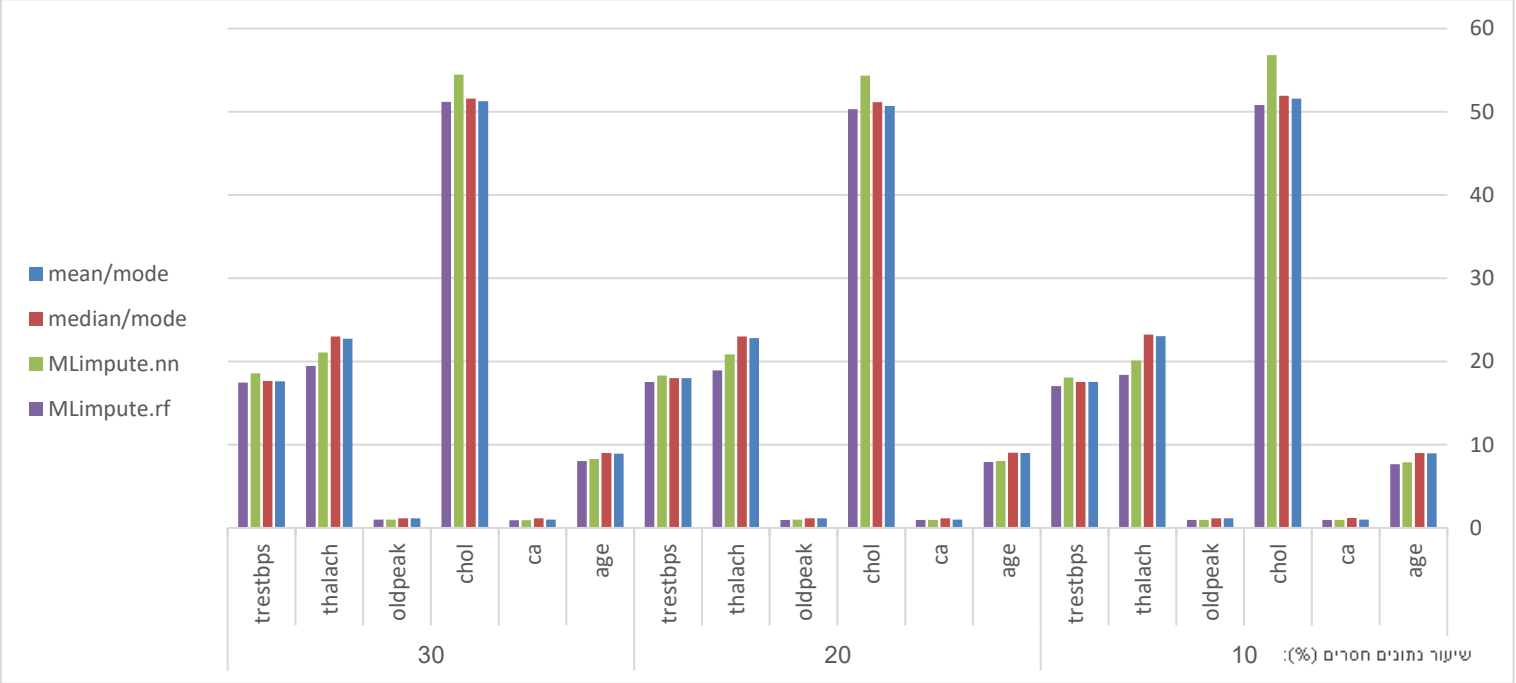
הערה : מדדי ההטיה מהווים ממוצע של מדגם של 40 הרצות שונות.

תרשים 1 – מדד accuracy ממוצע עבור שיטות השלמת הנתונים השונות לפי שיעור נתונים חסרים במסדי הנתונים עבור משתנים קטגוריאליים ובינאריים.



הערה : מדדי ההטיה מהווים ממוצע של מדגם של 40 הרצות שונות.

תרשים 2 – מדד RMSE ממוצע עבור שיטות השלמת הנתונים השונות לפי שיעור נתונים חסרים במסדי הנתונים עבור משתנים מספריים.



הערה : מדדי ההטיה מהווים ממוצע של מדגם של 40 הרצות שונות.

ביצועי המודלים לניבוי מחלת לב

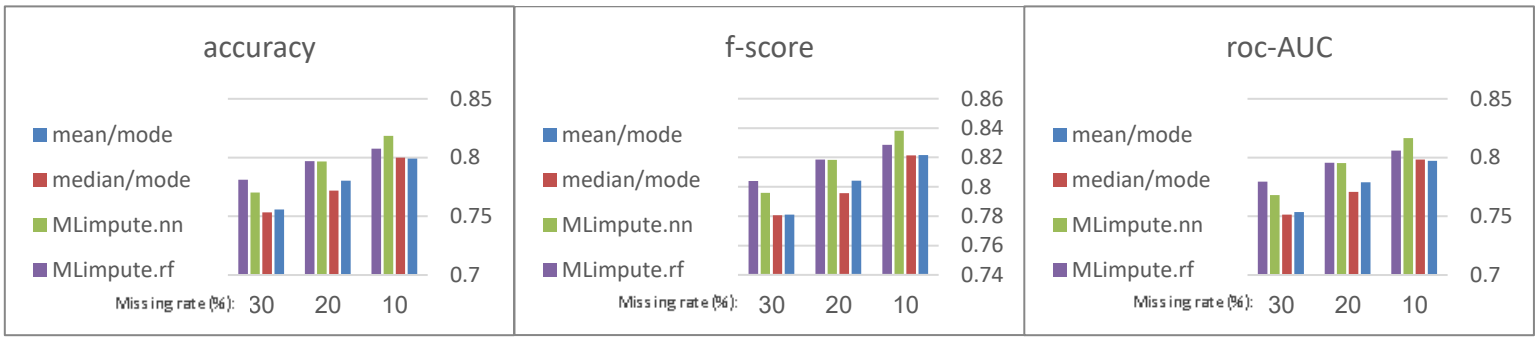
בחרנו במודלים הבאים לניבוי מחלת לב: NN , RF ו LR. ביצועי המודלים וטיוב שלהם אינם בפוקוס של מחקר זה. למעשה השתמשנו במודלים השונים לניבוי מחלת לב וביצועיהם ככלי השוואתי, על מנת לבחון את כדאיות השימוש בשיטות שונות להשלמת נתונים. רצינו לראות האם שימוש בשיטת MLimpute תביא לשיפור ביצועי המודלים השונים בהשוואה לשימוש בהשלמת נתונים בשיטות מסורתיות מסוג single imputation. בטבלה 2 ניתן לראות את מדדי הביצועים הממוצעים של המודלים שנבחרו לניבוי מחלת לב בחלוקה לפי שיעור הנתונים החסרים ומודל השלמת הנתונים שהשלים את מסד הנתונים שנמסר למודלי הניבוי. המדדים שמדדנו הם accuracy, roc-AUC ו f-score. עבור כל שיעור נתונים חסרים התקבלו 40 גרסאות שונות שנוצרו באקראי ממסד הנתונים המקורי, ועבור כל מסד נתונים שהתקבל לאחר השלמת נתונים בוצעו 40 חלוקות שונות ל train ו validation. בתרשים 3 אספנו עבור כל מדד ביצוע את הממוצע המשותף ל 3 מודלי הניבוי השונים. בתרשים ניתן לראות כי השלמת הנתונים ע"י MLimpute הביאה לשיפור ניכר בתוצאות מודלי הניבוי ובביצועיהם, זאת לעומת השלמה בשיטות המסורתיות מסוג single imputation. מגמה נוספת שניתן לראות היא שככל ששיעור הנתונים החסרים גבוה יותר, כך מודלי הניבוי השונים הגיעו למדדי ביצועים ממוצעים נמוכים יותר, זאת על אף שלא ראינו מגמה דומה במדדי ההטיה של השלמת הנתונים החסרים בהשוואה למסד הנתונים המקורי, כפי שניתן לראות בתרשימים 1 ו 2. בנוסף, ראינו כי ככל ששיעור הנתונים החסרים גבוה יותר, כך מדדי הביצוע של מודל MLimpute.rf נפגעו פחות בהשוואה לשיטות האחרות.

טבלה 2 – מדדי ביצועים ממוצעים עבור מודלי הניבוי למחלות לב, מחולק לפי מודל ניבוי, מודל השלמת נתונים חסרים ושיעור נתונים חסרים. מדדי הביצועים שנמדדו הם roc-AUC ,accuracy ו f-score.

MLimpute.rf			MLimpute.nn			median/mode			mean/mode			מודל השלמת נתונים / מדד / מודל ניבוי	שיעור חוסרים (%)
roc-AUC	f-score	accuracy	roc-AUC	f-score	accuracy	roc-AUC	f-score	accuracy	roc-AUC	f-score	accuracy		
0.809	0.837	0.813	0.813	0.839	0.817	0.801	0.830	0.805	0.803	0.832	0.807	lr	10
0.787	0.814	0.790	0.810	0.835	0.813	0.788	0.814	0.790	0.785	0.813	0.787	nn	
0.821	0.836	0.820	0.826	0.840	0.825	0.806	0.820	0.805	0.804	0.820	0.803	rf	
0.797	0.824	0.800	0.794	0.822	0.797	0.776	0.807	0.779	0.785	0.816	0.789	lr	20
0.787	0.812	0.788	0.787	0.813	0.789	0.760	0.789	0.762	0.774	0.803	0.776	nn	
0.803	0.819	0.803	0.804	0.820	0.803	0.777	0.790	0.775	0.778	0.793	0.776	rf	
0.778	0.809	0.782	0.766	0.801	0.771	0.759	0.797	0.764	0.758	0.794	0.763	lr	30
0.771	0.799	0.773	0.756	0.789	0.759	0.732	0.768	0.734	0.731	0.760	0.733	nn	
0.789	0.804	0.788	0.782	0.797	0.780	0.763	0.777	0.761	0.771	0.789	0.771	rf	

הערה : מדדי הביצועים מהווים ממוצע של מדגם של 40*40 הרצות שונות.

תרשים 3 – מדדי ביצועים ממוצעים עבור מודלי הניבוי למחלות לב, מחולק לפי מודל השלמת נתונים חסרים ושיעור נתונים חסרים. מדדי הביצועים שנמדדו הם roc-AUC ,accuracy ו f-score.



הערה : מדדי הביצועים מהווים ממוצע של מדגם של 40*40 הרצות שונות.

השלמת נתונים לאחר פיצול ל train ו validation

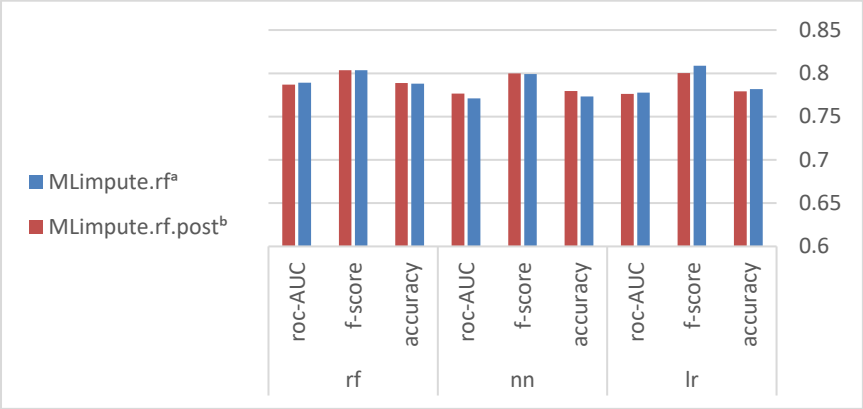
בדיקה נוספת שביצענו היא האם ניתן לבצע את השלמת הנתונים החסרים ע"י MLimpute לאחר הפיצול ל train ו validation, זאת על מנת לבדוק שיטה שעשויה להקטין את הסיכון ב overfitting. את הבדיקה ביצענו עבור MLimpute ושימוש ב RF, קראנו לשיטה זו MLimpute.rf.post. הרצנו את השיטה על 40 מסדי הנתונים שיצרנו בשיעור נתונים חסרים של 30% והשווינו את ביצועי המודלים שבחרנו לחיזוי מחלת לב אל מול MLimpute.rf. רצינו לראות האם קיים שוני מהותי בביצועים של המודלים במונחי accuracy, roc-AUC ו f-score, את הנתונים שאספנו ניתן לראות בטבלה 3. בתרשים 4 רואים כי אין שוני מהותי במדדי הביצועים של המודלים השונים כשמשווים בין MLimpute.rf ו MLimpute.rf.post.

טבלה 3 – מדדי ביצועים ממוצעים עבור מודלי הניבוי למחלות לב, מחולק לפי מודל ניבוי, ומודל השלמת הנתונים החסרים: MLimpute.rf ו MLimpute.rf.post. מדדי הביצועים שנמדדו הם accuracy, roc-AUC ו f-score.

מודל השלמת נתונים	MLimpute.rf[b]			MLimpute.rf.post[a]		
מדד / מודל ניבוי	accuracy	f-score	roc-AUC	accuracy	f-score	roc-AUC
lr	0.782	0.809	0.778	0.779	0.800	0.776
nn	0.773	0.799	0.771	0.779	0.800	0.776
rf	0.788	0.804	0.789	0.789	0.804	0.787

[a]מדדי הביצועים מהווים ממוצע של מדגם של 40 הרצות שונות.
[b] מדדי הביצועים מהווים ממוצע של מדגם של 40*40 הרצות שונות.

תרשים 4 – מדדי ביצועים ממוצעים עבור מודלי הניבוי למחלות לב, מחולק לפי מודל ניבוי, ומודל השלמת הנתונים החסרים: MLimpute.rf ו MLimpute.rf.post. מדדי הביצועים שנמדדו הם accuracy, roc-AUC ו f-score. בוצע על מסדי נתונים עם שיעור נתונים חסרים של 30%.



[a]מדדי הביצועים מהווים ממוצע של מדגם של 40*40 הרצות שונות.
[b] מדדי הביצועים מהווים ממוצע של מדגם של 40 הרצות שונות.

דיון

תרומת המחקר שלנו מתבטאת במספר אופנים. ראשית, MLimpute מתפקדת כ"מכונה" המקבלת כפלט מסד נתונים עם חוסרים, ונותנת כפלט מסד נתונים ללא חוסרים, המוכן לשימוש לטובת מודלים שונים לניבוי. ראינו כי בעוד ששיטות מסורתיות מסוג single imputation ממלאות את אותה התכלית, מסדי הנתונים שהתקבלו ע"י MLimpute הביאו לשיפור בביצועים של מודלי ניבוי למחלות לב. ל MLimpute מספר יתרונות על פני מודלים מודרניים אחרים להשלמת נתונים. ראשית, MLimpute מתפקדת כשיטה מודרנית מסוג single imputation. שיטות מודרניות אחרות מסוג multiple imputation כמו MICE ו GAIN,לוקחות מסד נתונים עם חוסרים ומחזירות כפלט סידרה של מסדי נתונים שלמים. על כן, עבור פלט המתקבל משיטה מסוג multiple imputation המשתמש נדרש למטלות המשך כמו ניתוח ובחירת נתונים לאחר הרצת המודל. MLimpute מחזיר מסד נתונים יחיד. היתרון כאן הוא בפשטות, שכן עבור המשתמש הממוצע, המבקש למלא חוסרים במסד הנתונים שברשותו, יותר אינטואיטיבי ופשוט לעבוד עם מסד נתונים אחד ושלם לעומת סדרה של מסדי נתונים, הטומנת בתוכה מטלות נוספות עבור המשתמש. יתרון נוסף של MLimpute על פני מודלים מודרניים אחרים נובע מהגמישות של

המודל. המודל יכול להתמודד עם נתונים חסרים מכל סוג שהוא (מספריים, קטגוריאליים, בינאריים וכו'), וללא דרישות כלל מהמשתמש לספק מידע לגבי אופי מסד הנתונים ואופי התפלגות המשתנים. למעשה, MLimpute מהווה מעין "קופסה שחורה" הממלאה את תכליתה ללא עזרה או דרישות נוספות (מקדימות או לאחר הרצה) מהמשתמש.

תרומה אפשרית נוספת הנובעת ממימוש MLimpute היא פתיחת צוהר לשינוי גישה בתחום ה ML. בשנים האחרונות אנו עדים להתפתחות מואצת של תחומים כמו בינה מלאכותית ולמידת מכונה. כשאנחנו חוזים בדברים המדהימים שמודלים מתחום ה ML יכולים לעשות, הדרישה מהם להתמודד עם נתונים חסרים נראית טריוויאלית ומתבקשת. הדרישה כיום של מודלים לקבל אך ורק מסדי נתונים מלאים וללא חוסרים אינה תואמת את הקידמה והיכולות של אותם מודלים. לשיטתנו, מודלים מתחום ה ML לא צריכים לדרוש מהמשתמש לקבל מסד נתונים ללא חוסרים. במידה ומודל מסוים מקבל מסד נתונים עם חוסרים, על אותו מודל לנסות לטפל בעצמו ולהשלים את אותם חוסרים, ע"י שימוש ב MLimpute. למעשה אנחנו מעבירים את הדרישה להשלמת הנתונים מהמשתמש לאלגוריתם עצמו. דרישה זו מתחדדת עבור מודלים גמישים כמו NN ו RF היכולים להתמודד עם סוגים שונים של משתנים.

להמשך מחקר ישנם מספר כיוונים שיש לבחון. במחקר שלנו הרצנו את המודל על מסד נתונים מלא, שהוחסרו ממנו נתונים בצורה אקראית לחלוטין על פני כל התכונות. באופן טבעי החוסר בנתונים התפלג בצורה דומה ומונוטונית בין התכונות. יש לבחון את המודל על צורות והתפלגויות שונות של חוסר בנתונים ועל מסדי נתונים מהעולם האמיתי המגיעים עם חוסרים בצורה טבעית. בנוסף, יש לבחון את ביצועי המודל עבור מסדי נתוני עתק. יתכן ויש יתרון בשימוש במסדי נתונים גדולים, שכן יש משמעות לכמות הרשומות שמודל מתחום ה ML מקבל לאימון ולמידה. מקום נוסף לחקר MLimpute הוא שימוש במודל עבור מסדי נתונים מסוגים שונים, מתחום הרפואה ומתחומים אחרים כמו פיננסים, דמוגרפיה ועוד. יש מקום לבצע השוואה בין ביצועי MLimpte אל מול שיטות מודרניות אחרות כמו MICE, GAIN ו opt.impute, הן בהיבטי דיוק ושיפור תוצאות מודלי ניבוי והן בהיבט זמני הריצה.

יש מקום למחקר המשך ע"י ניסיון לשפר את אלגוריתם MLimpute. יש לבחון מימושים נוספים של MLimpute ע"י מודלים גמישים נוספים מתחום ה ML. בנוסף, יש לבחון האם יש מקום לשיפור ואופטימיזציה בבחירת הפרמטרים במודלים שנבחרו ע"י ביצוע בחירת פרמטרים עבור כל תכונה בנפרד. יש לבחון מספר אפשרויות לאופטימיזציה של האלגוריתם עצמו. ייתכן וניתן לשפר את תוצאות המודל כתוצאה משינוי סדר בחירת התכונות להשלמה, ע"י הוספת משקולות ציון של מתאם עבור כל תכונה עם שאר התכונות או רמת החשיבות של כל תכונה עבור מודל הניבוי. אפשרות נוספת לאופטימיזציה של המודל היא הרצה איטרטיבית של המודל מספר פעמים עד הגעה להתכנסות (convergance). במצב הקיים המודל משתמש תחילה במסד נתונים שהושלם בשיטת single mean imputation ומשלים פעם אחת את התכונות אחת אחר השניה. יתכן וחזרה איטרטיבית של סדרת פעולות השלמת הנתונים, כאשר בכל איטרציה הנתונים שהושלמו "מתקדמים" לתוצאה יותר טובה תביא לשיפור ביצועי המודל.

ישנן אפשרויות נוספות להכנסת שינויים במודל ולקיחתו למקומות שונים במימוש שלו. אפשרות אחת היא מימוש של מודל היברידי על בסיס MLimpute. ראינו כי ניתן לממש את MLimpute

ע״י שימוש במודלים שונים מתחום ה ML. ההצעה שלנו היא לייצר מודל היברידי המשתמש בסדרת מימושים שונים של MLimpute ומשלים את הנתונים ע״י שימוש בערך אמצעי או שכיח מתוך כלל הערכים שהתקבלו ע״י סדרת המודלים. בצורה דומה ניתן לממש את MLimpute שיתפקד כשיטה מסוג multiple imputation ומציע סדרה של מסדי נתונים שלמים עבור מסד נתונים חסר. במימוש הנוכחי שלו MLimpute משלים את הנתונים תחת הנחת MCAR. ניתן לממש מודל התומך הנחות MAR ו MNAR ע״י הוספת משתנה בינארי לכל תכונה – חסר/לא חסר.

שאלה שנשארת פתוחה בשלב זה היא רמת ה overfitting שהמודל מייצר. ייתכן ומימושים שונים מגיעים לרמות שונות של overfitting. בנוסף, ראינו כי ניתן לממש את MLimpute לשלב שאחרי הפיצול ל train ו validation מבלי לפגוע משמעותית בביצועי המודל. בצורה כזו הסבירות ל overfitting קטנה, שכן הנתונים שבמסד ה validation לא הושפעו מהשלמת הנתונים שבוצעו במסד ה train ולהפך. בכל מקרה לא ברור בשלב זה האם רמת ה overfitting גבוהה או נמוכה באופן יחסי, והאם ביצוע השלמת הנתונים לאחר הפיצול ל train ו validation מקטין בצורה משמעותית את הסיכוי ל overfitting.

סיכום ומסקנות

בעיית השלמת הנתונים החסרים עבור מסדי נתונים מהחיים האמיתיים היא בעיה נפוצה ומוכרת לרוב העוסקים במלאכת ניבוי על סמך נתונים. בעבודה זו אנו מציעים משפחה של פתרונות לבעיה הנקראת MLimpute. ראינו כי כשמדובר בשחזור של מסדי נתונים מלאים שהוחסרו מהם נתונים בצורה אקראית, מימושים שונים של MLimpute יכולים להתעלות על מודלים להשלמת נתונים חסרים מסורתיים מסוג single imputation בדיוק וברמת ההטיה. בנוסף ראינו כי מודלים שונים לניבוי מחלת לב אשר עשו שימוש במסדי נתונים שהכילו נתונים חסרים והושלמו ע״י MLimpute הראו ביצועים טובים יותר לעומת אותם מודלים לניבוי אשר קיבלו את אותם מסדי נתונים שהכילו נתונים חסרים והושלמו ע״י שימוש בשיטות מסורתיות מסוג single imputation. העלינו הצעה למימוש לא קונבנציונלי של המודל, ע״י השלמת מסדי הנתונים בשלב שלאחר הפיצול ל train ו validation, וראינו שמימוש זה לא פוגע משמעותית בביצועים, בעוד שישנה אפשרות שמימוש מסוג זה עשוי להקטין את הסיכוי לתופעת ה overfitting.

מעבר לרמת הביצועים של MLimpute, היתרונות הטמונים בו נובעים בעיקר מפשטות השימוש בו, הגמישות שמובנית בו ובמודלים מתחום ML בהם הוא משתמש והוורסטיליות שלו. ראשית, כשמשתמשים במודלי ניבוי גמישים כמו NN ו RF, המסוגלים להתמודד עם סוגים שונים של משתנים (קטגוריאליים, מספריים, בינאריים וכו׳), אותה גמישות מוקרנת ל MLimpute. למעשה MLimpute גמיש כמו המודל ML אשר הוא עושה בו שימוש. שנית, לעומת שיטות מודרניות רבות אחרות, MLimpute לא דורש מידע מקדים מהמשתמש לגבי אופי הנתונים וצורת התפלגות של המשתנים השונים. בנוסף, בעצם היותו מימוש מודרני של single imputation, המודל מספק למשתמש מוצר מוגמר בדמות מסד נתונים שלם המוכן לשימוש, בניגוד לשיטות מודרניות מסוג multiple imputation המספקות למשתמש סט של מסדי נתונים שלמים המצריך המשך ניתוח ובחירת משתנים לטובת שימוש פרקטי בנתונים.

ביבליוגרפיה

Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2017). From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research*, *18*(1), 7133-7171. https://dl.acm.org/doi/abs/10.5555/3122009.3242053

Bø, T. H., Dysvik, B., & Jonassen, I. (2004). LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic acids research*, *32*(3), e34-e34. https://academic.oup.com/nar/article-abstract/32/3/e34/2904603

Brás, L. P., & Menezes, J. C. (2007). Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular engineering*, *24*(2), 273-282. https://www.sciencedirect.com/science/article/pii/S1389034407000354

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, *172*(9), 1070-1076. https://academic.oup.com/aje/article-abstract/172/9/1070/148540

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68. https://dspace.library.uu.nl/handle/1874/44635

Goldstein, B. A., Navar, A. M., & Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal*, *38*(23), 1805-1814. https://academic.oup.com/eurheartj/article-abstract/38/23/1805/3056931

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software*, *45*(7), 1-47. http://artax.karlin.mff.cuni.cz/~hans/src/doc/r-cran-amelia/amelia.pdf

Kim, K. Y., Kim, B. J., & Yi, G. S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, *5*(1), 160. https://link.springer.com/article/10.1186/1471-2105-5-160

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology*, *171*(5), 624-632. https://academic.oup.com/aje/article-abstract/171/5/624/137388

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Liu, Y., & Gopalakrishnan, V. (2017). An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data*, *2*(1), 8. https://www.mdpi.com/2306-5729/2/1/8

Masconi, K. L., Matsha, T. E., Erasmus, R. T., & Kengne, A. P. (2015). Effects of different missing data imputation techniques on the performance of undiagnosed diabetes risk prediction models in a mixed-ancestry population of South Africa. *PloS one*, *10*(9), e0139210. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139210

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554. https://ieeexplore.ieee.org/abstract/document/8740989/

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One*, *13*(8), e0202344. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202344

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118. https://academic.oup.com/bioinformatics/article-abstract/28/1/112/219101

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525. https://academic.oup.com/bioinformatics/article-abstract/17/6/520/272365

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, *16*(3), 219-242. https://journals.sagepub.com/doi/abs/10.1177/0962280206074463

Wang, X., Li, A., Jiang, Z., & Feng, H. (2006). Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC bioinformatics*, *7*(1), 32. https://link.springer.com/article/10.1186/1471-2105-7-32

Yoon, J., Jordon, J., & Van Der Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. *arXiv preprint arXiv: 1806.02920*. https://arxiv.org/abs/1806.02920

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1).
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716933/