תמונה שמכילה סמל, לוגו, גרפיקה, גופן

התיאור נוצר באופן אוטומטי****

**תואר שני במנהל עסקים ומערכות מידע**

**70-949-01/15 כריית נתונים וגילוי ידע**

**מטלה מסכמת בנושא**

פרופיל חולה סוכרת



**מגישים**

**מנחה**

**תאריך**

כ״ח באב ה׳תשפ״ד

01.09.2024

1. **שאלת הידע - המחקר**
   1. בעיית הידע: פרופיל חולה סוכרת.
   2. שאלת המחקר: מהו המודל המאפשר ניבוי סוכרת בצורה המדויקת ביותר
   3. מטרת שאלת המחקר: זיהוי הפוטנציאל לחלות בסוכרת באמצעות למידת מכונה
   4. חשיבות ותרומות המחקר:

* לחולה הפוטנציאלי:
* הפחתת לחץ וחרדה: זיהוי מוקדם של מחלה יכול לספק לחולה הפוטנציאלי תחושת שליטה על מצב בריאותו. כאשר אדם יודע מהם הסיכונים ומה הוא יכול לעשות כדי למנוע את המחלה, הוא יכול להרגיש רגוע יותר ולחוש פחות לחץ וחרדה.
* מעקב אישי מותאם: ידיעת גורמי הסיכון והיסטוריה רפואית אישית יכולה להוביל לתוכנית מעקב אישית, הכוללת בדיקות רפואיות תקופתיות ומותאמות אישית, מה שיכול להוביל לזיהוי מוקדם יותר של שינויים במצב הבריאותי.
* חינוך והסברה: ידע על גורמי הסיכון למחלה מאפשר לחולה ללמוד ולחנך את עצמו בנושא הבריאות, מה שיכול לשפר את יכולתו לקבל החלטות מושכלות בנוגע לבריאותו.
* למערכת הבריאות:
* שיפור איכות הטיפול: זיהוי מוקדם של המחלה מאפשר למערכת הבריאות לספק טיפול מותאם אישית ומדויק יותר, מה שיכול לשפר את איכות הטיפול הכוללת ולתרום להצלחת התהליך הטיפולי.
* הפחתת עומס על הצוות הרפואי: חיזוי מוקדם וטיפול יעיל יכולים להקטין את הצורך בטיפולי חירום ובטיפולים מסובכים יותר, ובכך להקל על עומס העבודה של הצוות הרפואי ולאפשר להם להתמקד במטופלים נוספים.
* קידום מחקרים : נתונים על חיזוי וזיהוי מוקדם של המחלה יכול לסייע במחקרים, לספק הבנה מעמיקה יותר על המחלה ולתרום לפיתוח תרופות ושיטות טיפול חדשניות. זה יכול להוביל לפריצות דרך בתחום הרפואה והבריאות ולהגברת הידע הרפואי הכללי.
* להיבט החברתי:
* הפחתת עלויות אישיות: חולים המזהים את המחלה בשלב מוקדם יכולים להימנע מהוצאות גבוהות על טיפולים מסובכים ומורכבים בעתיד. הדבר יכול להקל על הנטל הכלכלי של החולה ושל משפחתו.
* שיפור איכות החיים של הקהילה: כאשר מערכת הבריאות מצליחה לזהות מחלות בשלב מוקדם ולטפל בהן בצורה יעילה, הדבר משפיע על הבריאות הכללית של האוכלוסייה ומשפר את איכות החיים של כלל הקהילה.
* מודעות ציבורית: חינוך הציבור והעלאת המודעות לגבי חשיבות הזיהוי המוקדם יכולים להוביל לשינויים חברתיים והתנהגותיים, כולל אימוץ אורח חיים בריא יותר, שיפור התזונה והגברת הפעילות הגופנית.

1. **מיפוי: נתונים ,מידע וידע**

|  |  |  |
| --- | --- | --- |
| במערכת הממוכנת | בעולם האנושי- ארגוני | מאפיין/תהליך |
| נתוני גלם ממכשירי ניטור סוכר בדם, יומני רמות סוכר בדם, תכונות מקריאות סוכר (זמן המדידה, רמת הסוכר, מגמות לאורך זמן). | קריאות סוכר המצביעות על רמות לא תקינות, דוחות אירועים רפואיים הקשורים למצבי חירום סוכרתיים, רשומות מטופלים שסומנו עם אבחנת סוכרת. | נתונים |
| דפוסים ומגמות מנותחים ברמות סוכר בדם, נתונים מצטברים על מקרים ידועים של היפרגליקמיה או היפוגליקמיה, תובנות ממעקב סוכר רציף. | דפוסים ומאפיינים מנותחים של קריאות סוכר המצביעות על סיכון לסוכרת, תובנות על מאפיינים משותפים בקרב מטופלים עם סוכרת. | מידע |
| ההבנה של מודלים מבוססי למידה ממוחשבת על גורמים המצביעים על סיכון לסוכרת, תובנות המופקות מניתוח מערכי נתונים גדולים של מדידות סוכר. | חינוך לגבי גורמי סיכון ותסמינים של סוכרת, מודעות למחקרים רפואיים מתקדמים ואפשרויות טיפול בסוכרת. | ידע |
| שימוש באלגוריתמים של למידה ממוחשבת לעיבוד נתוני סוכר גולמיים, הפקת תכונות וזיהוי דפוסים המצביעים על סיכון לסוכרת, והפיכת הנתונים למידע וידע מעשי. | אנשי מקצוע רפואיים מפרשים נתוני סוכר גולמיים לזיהוי דפוסים ומאפיינים המצביעים על סיכון לסוכרת, והפקת תובנות משמעותיות. | הפיכת נתונים למידע וידע |
| שילוב תובנות ממחקרים רפואיים, ניתוח מומחים ומודלים מבוססי למידה ממוחשבת לנתונים מובנים לשיפור מודלי חיזוי הסיכון לסוכרת. | המרת דוחות רפואיים, ממצאי מחקר וידע מומחים לנתונים מובנים הניתנים לשימוש לאימון מודלים של למידה ממוחשבת לחיזוי סיכון לסוכרת. | הפיכת מידע וידע לנתונים |
| תיעוד הידע הסמוי הגלום במודלים של למידה ממוחשבת למפרטים, פרמטרים וקווים מנחים רפואיים לצורך שיתוף והבנה גלויים. | תיעוד הידע של אנשי מקצוע רפואיים למדריכים, קווים מנחים או חומרי הדרכה לשיתוף מפורש של הידע שלהם על פירוש קריאות סוכר והערכת סיכון לסוכרת. | הפיכת ידע סמוי לגלוי |
| הפעלת מודלים של למידה ממוחשבת לחיזוי בזמן אמת של סיכון לסוכרת, מתן אפשרות למודל ליישם ידע גלוי באופן סמוי ללא התערבות אנושית ישירה. | אנשי מקצוע מנוסים בתחום הבריאות מיישמים ידע מתועד בתרחישי טיפול במטופלים בזמן אמת, תוך שילוב הבנות ותובנות מעמיקות שלא נלכדו באופן גלוי בקווים מנחים כתובים. | הפיכת ידע גלוי לסמוי |
| מערכות אוטומטיות המסתמכות על הידע הנלמד לקבלת החלטות מהירות בזמן אמת, כגון התרעה למטופלים על אירועים פוטנציאליים של היפרגליקמיה או היפוגליקמיה בהתבסס על קריאות סוכר. | שימוש בידע על קריאות סוכר וגורמי סיכון לסוכרת לקבלת החלטות מושכלות לגבי תוכניות טיפול במטופלים, התערבויות באורח חיים ואמצעי מניעה. | תרומת הידע לתמיכה בהחלטות |
| שיפור איטרטיבי של מודלים מבוססי למידה ממוחשבת על סמך ידע חדש הנרכש מניתוח מתמשך של מגמות סוכר, ותרומה לגישות חדשניות לחיזוי וניהול סיכון סוכרת. | יישום ידע מתקדם על גורמי סיכון לסוכרת לחדשנות בפרוטוקולים רפואיים וטכנולוגיות חדשות לגילוי וניהול יעילים יותר של סוכרת. | תרומת הידע לחידוש ולחדשנות |
| שיפור מתמיד של המודלים באמצעות למידה מנתונים חדשים, שיתוף תובנות רפואיות בפלטפורמות בריאות והפצת עדכוני מודלים לשיפור גילוי וניהול מוקדם של סוכרת. | למידה מתמדת באמצעות תוכניות חינוך רפואי, שיתוף תובנות בקהילות בריאות והפצת שיטות עבודה מומלצות לשיפור היכולת הקולקטיבית של אנשי מקצוע בתחום הבריאות בניהול סוכרת. | למידה, שיתוף והפצת ידע |

1. **סקר ספרות**
   1. סקירת ספרות אודות עולם התוכן הנידון:

* שם המאמר:

Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR)

* סיכום המאמר:

המאמר מבחין בין משתנים קליניים, ביולוגיים וגנטיים כמנבאים לסוכרת. במסגרת המחקר, שמטרתו הייתה לפתח מדד סיכון קליני פשוט לסוכרת ולזהות מנבאים לסוכרת עתידית, נבחנו המשתנים הבאים כמנבאים אפשריים למחלה: היקף מותניים, יתר לחץ דם, עישון, היסטוריה משפחתית של סוכרת, גלוקוז בצום, BMI, , טריגליצרידים ופולימורפיזמים גנטיים. המחקר כלל גברים ונשים בגילאים 30-65 שנים, שנבדקו במשך תשע שנים, כאשר אבחנת הסוכרת הוגדרה לפי טיפול או רמות גלוקוז בצום. דיוק החיזוי, נמדד ע"י השטח מתחת לעקומות מקובלות (AROCs) והיה 0.713 אצל גברים ו0.827 אצל נשים, ועבר את מדד הסיכון ה-FINDRIEC לסוכרת.

ממצאי המחקר מלמדים שהמנבאים שונים בין המינים, וכי המשתנים הביולוגיים הטובים ביותר לניבוי סוכרת הם השמנת יתר (היקף מותניים+ BMI) ורמת הגלוקוז בדם במצב צום. המחקר מדגיש את החשיבות של שילוב נתונים קליניים וביולוגיים לחיזוי הסיכון לסוכרת, בעוד שנתונים גנטיים שנבדקו בעלי ערך חיזוי נוסף מוגבל כאשר כבר נלקחים בחשבון גורמים קליניים וביולוגים, ולכן אינם משפיעים באופן משמעותי על היכולת לנבא סוכרת.

* ציטוט (APA):

Balkau, B., Lange, C., Fezeu, L., Tichet, J., de Lauzon-Guillain, B., Czernichow, S., ... & Eschwège, E. (2008). Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes care, 31(10), 2056-2061.

* 1. סקירת ספרות אודות כריית נתונים בעולם התוכן הנידון
* שם המאמר:

Comparison of three data mining models for predicting diabetes or prediabetes by risk factors

* סיכום המאמר:

המחקר בחן את השימוש בטכניקות כריית נתונים לצורך פיתוח מודלים חיזויים לזיהוי סוכרת או טרום-סוכרת. המחקר התבסס על נתונים שנאספו משתי קבוצות מדגם: 735 אנשים עם סוכרת או טרום סוכרת ו-752 אנשים ללא מחלה. שתי הקבוצות מילאו שאלון אחיד בנוגע לאורח חייהן ונתונים אלה שימשו כקלט עבור שלושת המודלים. במסגרת המחקר, פותחו שלושה מודלים חיזויים אשר הוערכו במונחים של דיוק (Accuracy), רגישות (Sensitivity) וספציפיות (Specificity). שאלת המחקר המרכזית הייתה: איזה מודל כריית נתונים - רגרסיה לוגיסטית, רשתות עצביות מלאכותיות (ANN) או עץ החלטות - הוא המדויק ביותר בניבוי סוכרת או טרום-סוכרת בהתבסס על 12 גורמי סיכון נפוצים, כגון גיל, היסטוריה משפחתית, מצב משפחתי, רמת השכלה, לחץ בעבודה, משך שינה, פעילות גופנית, העדפה לאוכל מלוח, מגדר, אכילת דגים, שתיית קפה ומדד מסת גוף (BMI).תוצאות המחקר הראו כי מודל עץ ההחלטה הציג את דיוק הסיווג הגבוה ביותר (77.87%), בעוד שרגרסיה לוגיסטית הגיעה לדיוק של 76.13% ורשתות עצביות מלאכותיות (ANN) הגיעו לדיוק של 73.23%. בנוסף, מודל עץ ההחלטה הציג את הרגישות הגבוהה ביותר (80.68%) וכן את הספציפיות הגבוהה ביותר (75.13%) בהשוואה לשני המודלים האחרים. עוד נמצא כי מטריצת הבלבול (Confusion Matrix) של מודל עץ ההחלטה היא כדלקמן: חיובי אמיתי (TP) = 593, חיובי שגוי (FP) = 142, שלילי שגוי (FN) = 187, ושלילי אמיתי (TN) = 565.

* ציטוט (APA):

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

* 1. סקירת ספרות וטווח הערכים שידועים בספרות לאורם תיבחן שאלת המחקר:
* שם האמר:

Comparative analysis of predictive machine learning algorithms for diabetes mellitus

* סיכום המאמר:

מאמר זה בחן אלגוריתמים שונים בלמידת מכונה (ML) לזיהוי גורמי סיכון מרכזיים להתפתחות סוכרת. במסגרת המחקר, החוקרים השוו את ביצועיהם של מספר אלגוריתמים מוכרים בתוכנת Weka, ביניהם SVM, Naive Bayes, KNN, יער אקראי, רגרסיה לוגיסטית ועץ החלטות. הנתונים למחקר נאספו משני מקורות באתר Kaggle.com: מאגר הנתונים של Pima Indian Diabetes ומאגר הנתונים של בית החולים פרנקפורט בגרמניה. ממצאי המחקר מצביעים על הבדלים משמעותיים ביעילות האלגוריתמים בהתאם למאגר הנתונים. עבור מאגר הנתונים ההודי, אלגוריתם SVM השיג את הדיוק הגבוה ביותר עם 74.3%, והציג תוצאות של Precision, Recall ו-ROC area בשיעור של 74%, וערכים של MCC ו-Kappa בשיעור של 48%. לעומת זאת, במאגר הנתונים הגרמני, אלגוריתמי KNN ו- RF הצטיינו במיוחד, והשיגו דיוק מרשים של 98.7%. עבור אלגוריתמים אלו נמדדו Precision ו-Recall של 98%, ROC area של 97% עבור KNN ו-99% עבור RF, וערכים של MCC בשיעור של 97%, ו-Kappa של 97% עבור KNN ו-98% עבור RF.

* ציטוט (APA):

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. Bulletin of Electrical Engineering and Informatics, 12(3), 1728-1737.

1. **שיטת המחקר**
   1. תיאור הנתונים, מקורם ושימושים קודמים:

* מערך הנתונים:

מערך נתוני חיזוי לסוכרת מכילה 100,000 רשומות עם נתונים רפואיים ודמוגרפיים של מטופלים, יחד עם סטטוס הסוכרת שלהם (חיובי או שלילי). הנתונים כוללים מאפיינים כגון גיל, מגדר, מדד מסת גוף ,יתר לחץ דם, מחלות לב, היסטוריית עישון, רמת HbA1c ורמת גלוקוז בדם.

* שדה מטרה:

חולה/ לא חולה= משתנה מספרי (0/1).

חולה= נמצא בסיכון לפתח את מחלת הסוכרת.

לא חולה= אינו נמצא בסיכון לפתח את מחלת הסוכרת.

* מקור הנתונים:

אתר Kaggle

* קישור למערך הנתונים:

<https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>

* שימושים קודמים:

לפי הספרות המשתנים גיל, מגדר וBMI קיימים ברוב המחקרים. ישנם משתנים נוספים החופפים בין המחקר שלנו ומחקרים אחרים בספרות (נספח 4.1.1, 4.1.2, 4.1.3 ), כמו יתר לחץ דם, גלוקוז בדם ורמת סוכר.

* 1. פירוט המשתנים ואופן הטיפול במשתנים חסרים או רשומות עם נתונים חסרים:
* פירוט המשתנים:
* מין: מין מתייחס למין הביולוגי של הפרט, אשר יכול להשפיע על הסבירות לפתח סוכרת. קיימות שלוש קטגוריות: זכר, נקבה ואחר.
* גיל: גיל הוא גורם חשוב שכן סוכרת מאובחנת יותר בקרב מבוגרים. טווח הגילאים במאגר הנתונים שלנו נע בין 0 ל-80.
* יתר לחץ דם: הוא מצב רפואי שבו לחץ הדם בעורקים גבוה באופן מתמשך. ישנם שני ערכים: 0 מציין שאין יתר לחץ דם ו-1 מציין שיש יתר לחץ דם.
* מחלת לב: מחלת לב היא מצב רפואי נוסף שמקושר לסיכון מוגבר לפיתוח סוכרת. ישנם שני ערכים: 0 מציין שאין מחלת לב ו-1 מציין שיש מחלת לב.
* היסטוריית עישון: היסטוריית עישון נחשבת גם כגורם סיכון לסוכרת ויכולה להחמיר את הסיבוכים הקשורים לסוכרת. במאגר הנתונים שלנו ישנן חמש קטגוריות: לא נוכחי, בעבר, אין מידע, נוכחי, לעולם לא, ותמיד.
* BMI : מדד מסת גוף הוא מדד לשומן הגוף המבוסס על משקל וגובה. ערכי BMI גבוהים יותר מקושרים לסיכון מוגבר לסוכרת. טווח ה-BMI במאגר הנתונים הוא מ-10.16 עד 71.55. BMI פחות מ-18.5 נחשב כתת משקל, 18.5-24.9 כנורמלי, 25-29.9 כעודף משקל ו-30 ומעלה כהשמנה.
* רמת HbA1c: רמת HbA1c (המוגלובין (A1c היא מדד לרמת הסוכר הממוצעת בדם במהלך 2-3 חודשים האחרונים. רמות גבוהות יותר מצביעות על סיכון גבוה יותר לפיתוח סוכרת. ברוב המקרים רמת HbA1c מעל 6.5% מצביעה על סוכרת.
* גלוקוז בדם: רמת גלוקוז בדם מתייחסת לכמות הגלוקוז בזרם הדם בזמן נתון. רמות גבוהות של גלוקוז בדם הן אינדיקטור מרכזי לסוכרת.
* סוכרת: סוכרת היא המשתנה המנובא, עם ערכים של 1 המציינים נוכחות של סוכרת ו-0 המציינים היעדר סוכרת.
* משתנים חסרים:

חלק מהמשתנים קיימים בספרות אך לא קיימים בdataset של המחקר (נספח 4.2). על חלקם הגדול ניתן לגשר וחלקם אינם בעלי השפעה מהותית. היסטוריה משפחתית של סוכרת נמצא בספרות כמשתנה בעל השפעה גבוהה, יש המעידים כי מדובר על השפעה רק בקרב הנשים והאחרים אינם שוללים בוודאות את

הטענה.

* רשומות חסרות:

לא נמצאו בdataset שלנו רשומות חסרות. היה והיו קיימות כאלו יש למחוק אותם כך שלא יהיו חלק מקבוצת המחקר. מדובר ב100,000 מקרים, גם אם יהיו רשומות חסרות שנבחר שלא להתייחס אליהם עדיין נשאר לנו מדגם מספיק גדול בשביל להסיק מסקנות.

* 1. פירוט המניפולציות שהופעלו בייצוג הנתונים:

תחילה היה עלינו לבדוק שאין קשר או תלות בין המשתנים, בכדי לוודא שאין בעיה במשתני המחקר.

לפי נספח 4.3 ניתן להוכיח כי למשתנים יש אך ורק קשר או תלות עם עצמם.

* שינוי סקאלה:

בכדי לבחון מספר מגוון של אלגוריתמים ולהתאים את שיטות כריית הנתונים לספרות, היה עלינו לקבוע את שדה המטרה כמספרי ונומינלי. בdataset שדה המטרה הוא מספרי כאשר 0 = אינו חולה ו1= חולה. בשביל להריץ אלגוריתמים כמו J48 היינו צריכים להמיר את השדה לנומינלי ככה שY= חולה וN- אינו חולה.

* נירמול הנתונים:

המשתנים במחקר שלנו עם טווחים שונים, כך לדוגמא גיל הוא בסקאלה של 0-80 ואילו יתר לחץ דם הוא בסקאלה של 0 או 1, ולכן אין אנו יכולים להשוות בין המשנים כי הם נמצאים על אותה הסקאלה בין 0 ל 1. חשוב לזכור שהנרמול מאפשר השוואה בין המשתנים ובכך מציף תוצאות חריגות, אך אין למהר לנטרל במחשבה שמדובר בטעות, כי קיימת האפשרות שזו התחלה של מגמה.

* תקנון הנתונים:

בתקנון אין בעיה עם הטווחים, ולכן שיטה זו עדיפה במצב בו יש חריגים. במחקר שלנו הרצנו את האלגוריתמים שלוש פעמים, על dataset רגיל, מנורמל ומתוקנן. ניתן לראות שבחלק מהמקרים היה שינוי בתוצאות בעקבות מניפולציות שבוצעו בנתונים ואילו בחלק התוצאה הייתה זהה לחלוטין.

* 1. האלגוריתמים שהופעלו והפרמטרים הייחודים:

בוצעו מספר הרצות שונות בתוכנת Weka, כאשר כל הרצה התייחסה לשילוב שונה של dataset ואלגוריתם.

* האלגוריתמים שהופעלו:
* J-48
* NN
* Logistic Regression
* SMO (SVM)
* Linear Regression
* Rep Tree
* Random Forest
* IBK (KNN)
* הפרמטרים הייחודים שהופעלו:
* Diabetes-N: שדה המטרה הוא נומינלי (Y/N).
* Diabetes -N without BMI: הסרת המשתנה BMI ושדה המטרה הוא נומינלי (Y/N) .
* Diabetes -N without HbA1c\_level: הסרת המשתנה HbA1c\_level ושדה המטרה הוא נומינלי (Y/N).
* Diabetes -N without blood\_glucose\_level: הסרת המשתנה blood\_glucose\_level ושדה המטרה הוא נומינלי (Y/N).
* Diabetes -N without HbA1c\_level and blood\_glucose\_level: הסרת המשתנה HbA1c\_level ו- blood\_glucose\_level כאשר שדה המטרה הוא נומינלי (Y/N).
* Diabetes –O: שדה המטרה הוא מספרי (0/1), כאשר N=0 וY=1.
* Diabetes –O (Normalized 0-1): dataset מנורמל עם ערכים בין 0-1, ושדה המטרה הוא מספרי (0/1).
* Diabetes –O (Standardized): dataset מתוקנן, ושדה המטרה הוא מספרי (0/1).
* Diabetes –O (0,1000): שדה המטרה הוא מספרי (0/1000), כאשר N=0 וY=1000.
  1. אופן חישוב מדדי הערכה ופונקציית ההפסד:
* מדדי הערכה: (Confusion Matrix)
* Positive: הערכה שאדם חולה בסוכרת
* Negative: הערכה שאדם בריא (לא חולה בסוכרת)
* True Positive: אדם בריא שמערכת העריכה שהוא בריא
* True Negative: אדם חולה שמערכת העריכה שהוא חולה
* False Positive: אדם בריא שמערכת העריכה שהוא חולה
* False Negative: אדם חולה שמערכת העריכה שהוא בריא
* Accuracy: הדיוק מציין את אחוז המקרים בהם המערכת העריכה נכון אם מטופל חולה סוכרת או לא.
* Recall: רגישות גבוהה מצביעה על כך שהמערכת מצליחה לזהות בצורה טובה את מרבית החולים האמיתיים בסוכרת.
* Precision: דיוק חיובי גבוה מציין שהמערכת כמעט ולא מתריעה לשווא
* Kappa: מקדם קאפה נותן הערכה עד כמה המערכת מסכימה עם האבחנה האמיתית של המומחים בתחום, תוך התחשבות באפשרות של הסכמה מקרית. קאפה גבוה מציין כי המערכת אמינה ומבצעת חיזוי מדויק שאינו מבוסס על מקריות.
* טווח הערכים הרצוי:
* Accuracy: מכיוון שבמאמר 2 הערך היה 77.87% ובמאמר 3 הערך היה 74% וגם 98.7% - בחרנו במקרה שלנו ב 83.5%.
* Recall: מכיוון שבמאמר 2 הערך היה 80.68% ובמאמר 3 הערך היה 74% וגם 98% - בחרנו במקרה שלנו ב 84.2%.
* Precision: מכיוון שבמאמר 2 הערך היה 80.75% ובמאמר 3 הערך היה 74% וגם 98.7% - בחרנו במקרה שלנו ב 84.5%.
* Kappa: מכיוון שבמאמר 2 הערך היה 55% ובמאמר 3 הערך היה 48% וגם 97% וגם 98% - בחרנו במקרה שלנו ב 74.5%.
* פונקציית ההפסד:
* "מחיר" ה - FP סיווג שגוי חיובי מתאר מצב שבו האלגוריתם מנבא שלמטופל יש סוכרת, כאשר בפועל אין לו סוכרת. המטופל יעבור בדיקות נוספות מיותרות, שיכולות להיות פולשניות ולא נעימות. מהמטופל עלול להתחיל טיפול מיותר בסוכרת, כולל תרופות שעלולות לגרום לתופעות לוואי לא רצויות. המטופל יחווה לחץ וחרדה בלתי מוצדקים בגלל החשש מסוכרת.
* "מחיר" ה - FN סיווג שגוי שלילי מתאר מצב שבו האלגוריתם מנבא שלמטופל אין סוכרת, כאשר בפועל יש לו סוכרת. המטופל לא יקבל טיפול מתאים בזמן, מה שעלול להוביל לסיבוכים חמורים כמו פגיעות בכליות, בעיות לבביות, פגיעות ברגליים וכו'. איכות החיים של המטופל תיפגע באופן משמעותי עקב התקדמות המחלה ללא טיפול. מערכת הבריאות תיאלץ להתמודד עם טיפולים מורכבים ויקרים יותר בשלבים מאוחרים יותר של המחלה.
  1. אופן ביצוע ניתוח הרגישות לבדיקת יציבות המודל:

לצורך הערכת איכות המודל ביצענו ניתוח רגישות. חילקנו את מערך הנתונים לtrain ו Test כך שלקחו כ-66% ממערך הנתונים לtrain ו34% לTest. הפעלנו את המערכת על מקטע של נתונים וכך קיבלנו מודל. לאחר מכן, לקחנו את המודל והרצנו אותו בתוכנת הweka באלגוריתם J-48 מסוג Training 66%, על סדרת נתונים שהמודל לא ראה (test), וביצענו השוואה בין התוצאות לתוצאה המקורית, כך בעצם יצרנו הערכה של איכות המודל.

בחנו מהם המשתנים בעלי ההשפעה הגבוהה ביותר באמצעות תוכנת הweka, ועליהם הפעלנו את הרעש. כל עוצמת רעש נבדקה 3 פעמים, כל פעם עם מספרים שונים ובסוף נלקח הממוצע בין 3 ההרצות. ניתוח הרגישות בוצע באמצעות הפעלת רעש על 34,000 רשומות וטווח ההרעשה נע בסטיית תקן בהתפלגות נורמלית של 0-0.2. מטרת הפעלת הרעש הייתה לשנות את הערכים בנתוני המבחן, ובדיקת השינוי בתוצאות שהתקבלו. (נספח 4.6).

1. **תוצאות**
   1. תוצאות

כדי להציג את התהליך באופן מקביל לסיפור, ניתן לחלק את התיאור לשלושה שלבים עוקבים, שכל אחד מהם מייצג שלב מתקדם יותר בהבנת המודל.

תיאור זה משקף את המסע שאנו עוברים מהתוצאה ההתחלתית ועד לתוצאה הסופית, המדויקת והמושכלת, המבוססת על הבנה מעמיקה של כל המרכיבים במודל.

בהמשך המסמך נציג את התוצאות עבור כל שלב. השלב האחרון (שלב השלישי)

* השלב הראשון: התוצאה הראשונית

בהתחלה, אנו עדים לתוצאה הראשונית, שהיא התוצאה המתקבלת כאשר המודל פועל ללא שום התערבות או התאמה. זוהי התוצאה המייצגת את המצב כפי שהוא, מבלי שנלקחו בחשבון שיקולים נוספים כמו קבוצות מיעוט או מתווכים.

* טבלת התוצאות של השלב הראשון:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Dataset** | **Algorithm** | **F-score** | **RMSE** | **Kappa** |
| 1 | Diabetes-N | J-48 (Cross Validation) | 0.969\* | --- | 0.7866 |
| 2 | Diabetes-N | J-48 (Training 66%) | 0.967\* | --- | 0.7826 |
| 3 | Diabetes-N | J-48 (Training 66%)- Preserve order for % split | 0.971\* | --- | 0.7904 |
| 4 | Diabetes-N | NN (Cross Validation ) | 0.969\* | --- | 0.7858 |
| 5 | Diabetes-N | Logistic regression (Cross Validation) | 0.957\* | --- | 0.7082 |
| 6 | Diabetes -N  Without BMI | J-48 (Cross Validation) | 0.969\* | --- | 0.7871 |
| 7 | Diabetes -N  Without HbA1c\_level | J-48 (Cross Validation) | 0.935\* | --- | 0.5465 |
| 8 | Diabetes -N  Without blood\_glucose\_level | J-48 (Cross Validation) | 0.946\* | --- | 0.6145 |
| 9 | Diabetes -N  Without HbA1c\_level and blood\_glucose\_level | J-48 (Cross Validation) | 0.880\* | --- | 0.055 |
| 10 | Diabetes –O | Linear Regression (Cross Validation) |  | 0.2251 |  |
| 11 | Diabetes –O | NN (Cross Validation ) |  | 0.1884 |  |
| 12 | Diabetes –O | Rep-Tree (Cross Validation ) | 0.9026\*\* | 0.1558 | 0.7862\*\* |
| 13 | Diabetes –O  (Normalized 0-1) | Linear Regression (Cross Validation) | --- | 0.2251 | --- |
| 14 | Diabetes –O  (Normalized 0-1) | NN (Cross Validation ) | --- | 0.1847 | --- |
| 15 | Diabetes –O  (Normalized 0-1) | Rep-Tree (Cross Validation ) | --- | 0.1558 | --- |
| 16 | Diabetes –O )Standardized( | Linear Regression (Cross Validation) | --- | 0.2251 | --- |
| 17 | Diabetes –O  )Standardized( | NN (Cross Validation ) | --- | 0.1864 | --- |
| 18 | Diabetes –O  )Standardized( | Rep-Tree (Cross Validation ) | 0.9028\*\* | 0.1555 | 0.7865\*\* |
| 19 | Diabetes –O (0,1000) | Rep-Tree (Cross Validation ) | --- | 225.0537 | --- |
| 20 | Diabetes –O | Random Forest CV\*\*\* | --- | 0.1596 | --- |
| 21 | Diabetes –O | IBK CV (KNN)\*\*\* | --- | 0.2232 | --- |
| 22 | Diabetes –O  (Normalized 0-1) | Random Forest CV\*\*\* | --- | 0.1595 | --- |
| 23 | Diabetes –O  (Normalized 0-1) | IBK CV (KNN)\*\*\* | --- | 0.2235 | --- |
| 24 | Diabetes –O  )Standardized( | Random Forest CV\*\*\* | --- | 0.1593 | --- |
| 25 | Diabetes –O  )Standardized( | IBK CV (KNN)\*\*\* | --- | 0.2235 | --- |
| 26 | Diabetes-N | SMO CV (SVM)\*\*\* | 0.957\* | --- | 0.6997 |

(טבלה עם כלל מדדי הערכה שהתקבלו- נספח 5.1.1)

\* יש שונות בין התוצאות ולכן לא ניתן להסתמך על הממוצע. בבחירת החלופה נתייחס לKappa ולF-score בכדי לגשר על כך.

\*\* ערך מחושב ולא מבוסס על הפלט של תוכנת Weka. נספח 5.1.2+ נספח 5.1.3 (נבחר הגדרת טווח סיווג של 0.99)

\*\*\* אלגוריתמים שלא נלמדו במהלך הקורס, אך לצורך השוואת המחקר לספרות, הרצנו גם אותם.

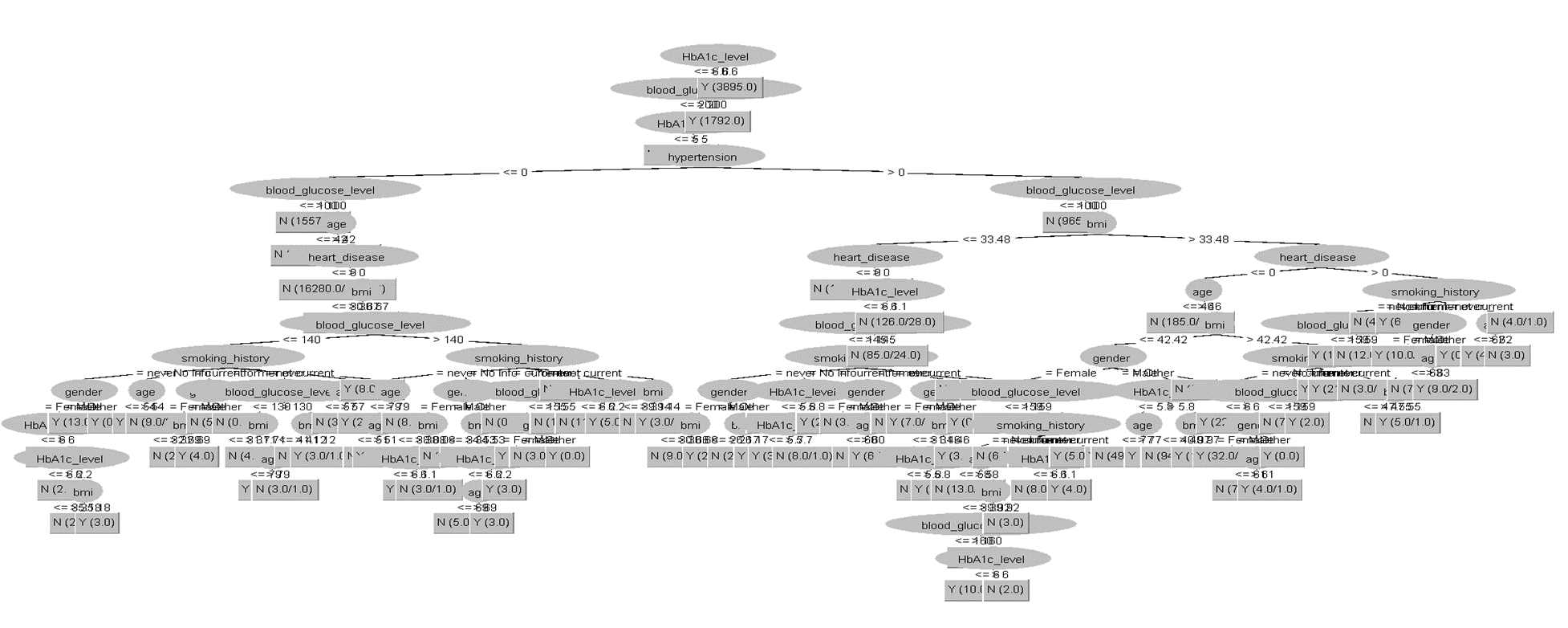
ביצענו בדיקה ידנית של כמה מהמודלים בשביל לבדוק האם התוצאה היא אמינה וניתן להתבסס על מדדי הערכה של המודל. (נספח 5.1.4)

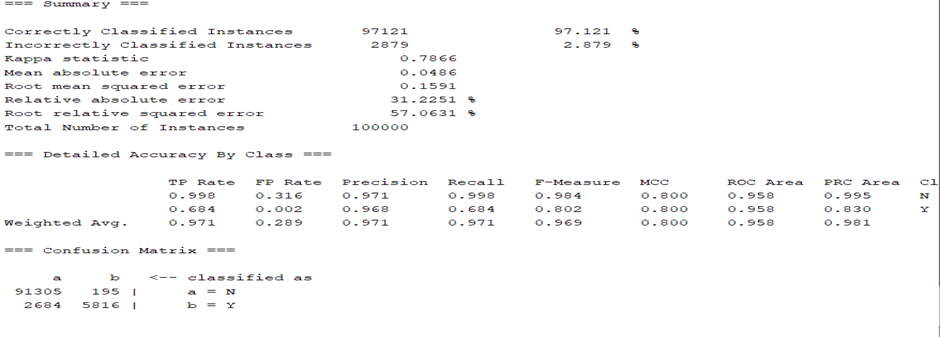
* התוצאה הטובה בשלב הראשון:

בשלב הראשון לאחר סקירת התוצאות ראינו שמודל עץ החלטה מסוג J-48 (Cross Validation) השיג את ההרצה הטובה ביותר, עם שדה מטרה נומינלי ומדדי הערכה F-score= 0.969 וKappa= 0.7866. (סט בדיקה מס' 1 בטבלה נספח 5.1.1).

* שדה מטרה מספרי

מבין כל ההרצות עם שדה מטרה מספרי, הרצה Rep-Tree(CV) הייתה עם הRMSE הכי נמוכה (נספח 5.1.1). למרות זאת, בחישוב הF-score ההרצה קיבלה F-score=0.9028 (נספח 5.1.2), ובהשוואה להרצה של J-48 (CV) בה קיבלנו F-score=0.969 לא ניתן לומר שההרצה הטובה היא עם שדה מטרה מספרי

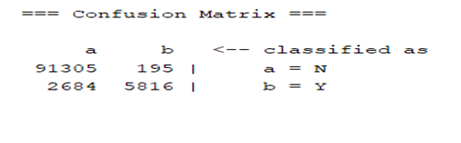
* עץ ההחלטה שהתקבל בתוכנת ה Weka בשלב הראשון:
* פלט הערכים שהתקבל בתוכנת ה Weka בשלב הראשון:



* השלב השני: התערבות בקבוצת המיעוט

בשלב הבא, אנו מבצעים טיפול והתערבות בקבוצת המיעוט. תוצאה זו מציגה כיצד השפעת הקבוצה על המודל משתנה בעקבות ההתערבות, וכיצד המדדים מושפעים בעקבות כך. כאן מתגלה בפנינו הבנה מעמיקה יותר של דינמיקות הקבוצה ושל המודל עצמו.

* לפי הconfusion matrix שהתקבל בפלט עם התוצאה הטובה ביותר, ניתן להבחין בקבוצת מיעוט. 91.5% שהם לא חולים ו8.5% שהם חולים



* דרך לטיפול בקבוצת המיעוט:

לפי דגימת אשכולות לאיזון האוכלוסיות בסט הנתונים שלנו נצמצם בהתאמה את האוכלוסייה של חולי הסוכרת, כך שבכל אחת יהיו 8,500 רשומות של מטופלים חולים ו-8,500 רשומות של מטופלים בריאים, כך שהיחס בין שתי הקבוצות עומד על 50% לכל אחת. נעשה את זה ע"י הפרדה של הרשומות של כל אחת, ביצוע ניתוח אשכולות לכל קבוצה, נמצא את נקודת המינימום הראשונה בשיטת הברך והנגזרת ולבסוף נוציא מWeka את פלט הסיווג ונדגום מכל אשכול באופן אקראי.

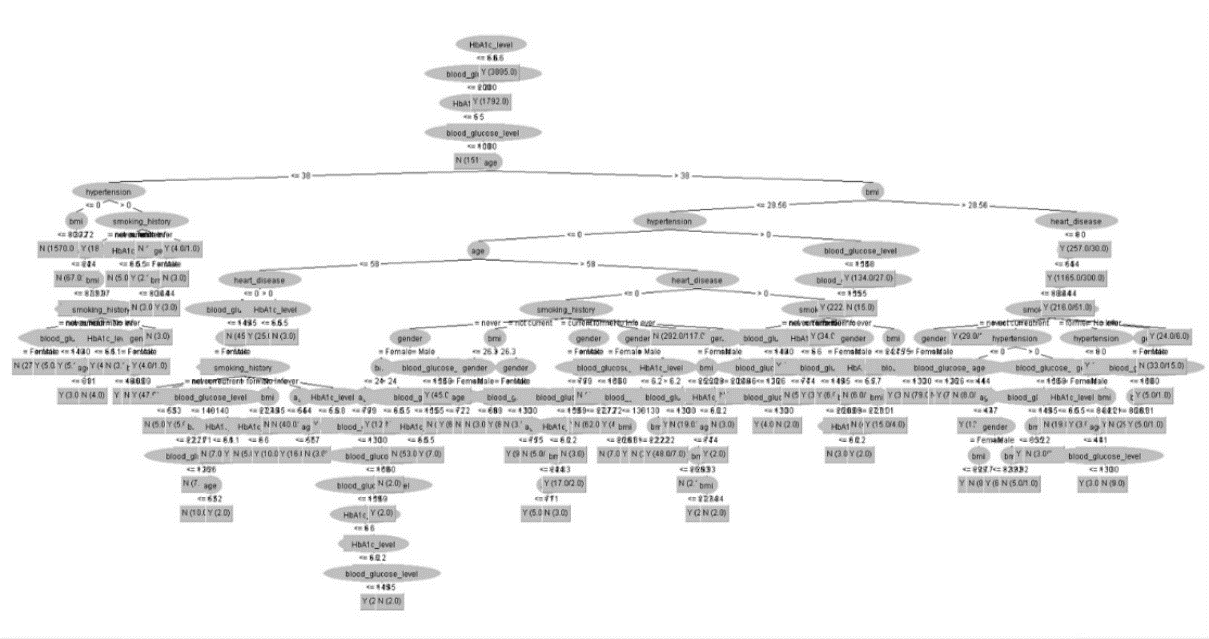
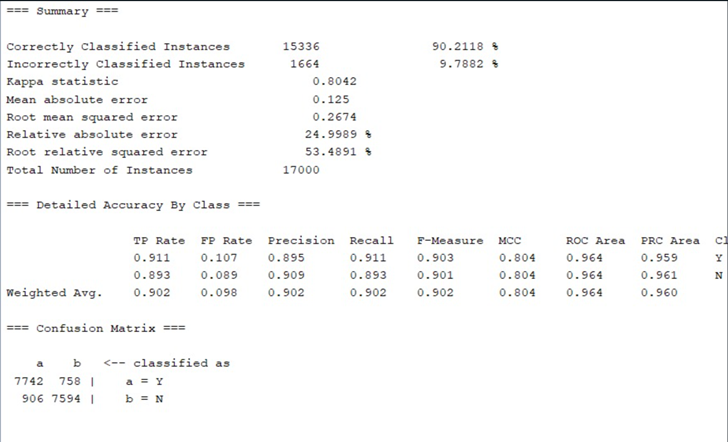
* טבלת התוצאות של השלב השני:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Dataset** | **Algorithm** | **F-score** | **RMSE** | **Kappa** |
| 1 | Diabetes-N | J-48 (Cross Validation) | 0.902 | --- | 0.8042 |
| 2 | Diabetes-N | Logistic regression (Cross Validation ) | 0.882 | --- | 0.7645 |
| 3 | Diabetes-N | NN (Cross Validation) | 0.902 | --- | 0.8041 |
| 4 | Diabetes-O | Rep Tree (Cross Validation) | --- | 0.2637 | --- |
| 5 | Diabetes-O | Random Forest (Cross Validation) | --- | 0.2597 | --- |

* התוצאה הטובה בשלב השני:

בשלב השני לאחר סקירת התוצאות ראינו שמודל עץ החלטה מסוג J-48 (Cross Validation) השיג את ההרצה הטובה ביותר, עם שדה מטרה נומינלי ומדדי הערכה F-score= 0.902וKappa= 0.8042. ריכוז התוצאות באלגוריתמים שונים לאחר טיפול בקבוצת המיעוט- נספח 5.1.5

עץ ההחלטות לאחר טיפול בקבוצות המיעוט נספח 5.1.6

* עץ ההחלטה שהתקבל בתוכנת ה Weka בשלב השני:
* פלט הערכים שהתקבל בתוכנת ה Weka בשלב השני:
* השלב השלישי: התוצאה לאחר טיפול במתווך בשילוב קובצת המיעוט - התוצאה הטובה ביותר

לבסוף, אנו מטפלים במתווך. בתהליך זה, אנו לוקחים בחשבון את ההשפעה שיש למתווך על הקשר בין המשתנים. התוצאה המתקבלת לאחר טיפול במתווך, ובשילוב עם ההתערבות בקבוצת המיעוט, היא תוצאה הסופית, אשר מייצגת את המודל השלם והמדויק ביותר, לאחר שכל הגורמים הובאו בחשבון.

* בדיקת מתווך לפי שינוי שדה המטרה:

יש צורך לבדוק את המשתנים במחקר שלנו, תחילה הרצנו את המודל J-48, כאשר כל פעם הורדנו משתנה אחר בהרצה (תוצאות בנספח 5.4.8) . כפי שנרשם לעיל, המשתנים גלוקוז בדם ורמת ה HbA1c, נשארו המשתנים הכי מהותיים לחיזוי סכרת. עוד ניתן להבחין שהמשתנים מחלות לב ומין, אינם תורמים לדיוק המודל בניבוי סכרת, שכן שהסרה שלהם מעלה את הkappa, ולכן ניתן להגיד שהם אפילו עשויים להוסיף רעש.

כעת עלינו לבדוק האם קיימת השפעה של משתנים אחרים על שני המשתנים העיקריים שמשפיעים על ניבוי סכרת, רמת גלוקוז בדם ורמת ה- HbA1c.

1. תחילה הרצנו את המודל כאשר רמת ה- HbA1c הוא שדה המטרה וקיבלו סטיית תקן של 1.1722. אח"כ הורדנו כל פעם אחד מהמשתנים ובדקנו את ההשפעה שלהם על המודל כאשר שדה המטרה הוא רמת ה- HbA1c. ניתן ללמוד מהתוצאות (נספח 5.4.9) שכאשר החסרנו את המשתנים BMI ורמת הגלוקוז בדם סטיית התקן רק עלתה ולכן ניתן לומר בוודאות שהמשתנים BMI ורמת הגלוקוז בדם משפיעים על ערך ה- HbA1c.
2. . אחכ הרצנו את המודל כאשר רמת ה- HbA1c הוא שדה המטרה, אבל כל פעם רק עם משתנה אחד נוסף.לפי התוצאות (נספח 5.4.10) ניתן לקבוע שרמת הגלוקוז בדם הוא בכל השפעה מהותית על רמת ה- HbA1c בהשוואה לשאר המשתנים במודל.
3. . כעת הרצנו את המודל כאשר רמת הגלוקוז בדם הוא שדה המטרה וקיבלו סטיית תקן של 52.478. אח"כ הורדנו כל פעם אחד מהמשתנים ובדקנו את ההשפעה שלהם על המודל כאשר שדה המטרה הוא רמת הגלוקוז בדם. ניתן ללמוד מהתוצאות (נספח 5.4.11) כאשר החסרנו כל אחד מהמשתנים האחרים לא ניכר שינוי מהותי בסטיית התקן, ולכן ניתן לקבוע כי אין משתנה אחד שמוביל להשפעה משמעותית על רמת הסוכר בדם במודל הנוכחי.
4. . לאחר שבחנו בנפרד את המשתנים יש לבדוק את המודל בניבוי סכרת עם משתנים אלו בלבד. ביצענו 3 הרצות כאשר אבחון סכרת הוא שדה המטרה, הראשונה לקחה בחשבון את שני המשתנים העיקריים (רמת גלוקוז בדם ורמת ה- HbA1c), השנייה התייחסה אך ורק לרמת ה- HbA1c, והשלישית התייחסה רק לרמת הגלוקוז בדם. מהתוצאות (נספח 5.4.12) ניתן ללמוד שלרמת ה- HbA1c יש השפעה גבוהה יותר באבחון סכרת לפי תוצאות הF-score, אך הKappa מלמדת אותנו שאין די במשתנים אלו לאבחון סכרת.

כלומר ניתן להגיד שרמת ה- HbA1, משמש כמתווך חלקי בין רמת הגלוקוז בדם לאבחון סכרת. כלומר, רמת הסוכר בדם משפיעה על רמת ה HbA1c, שבתורו משפיע על אבחון הסוכרת. עם זאת, עדיין יש קשר ישיר בין רמת הגלוקוז בדם לאבחון סכרת שאינו תלוי ברמת ה- HbA1c.

* טבלת התוצאות של השלב השלישי:

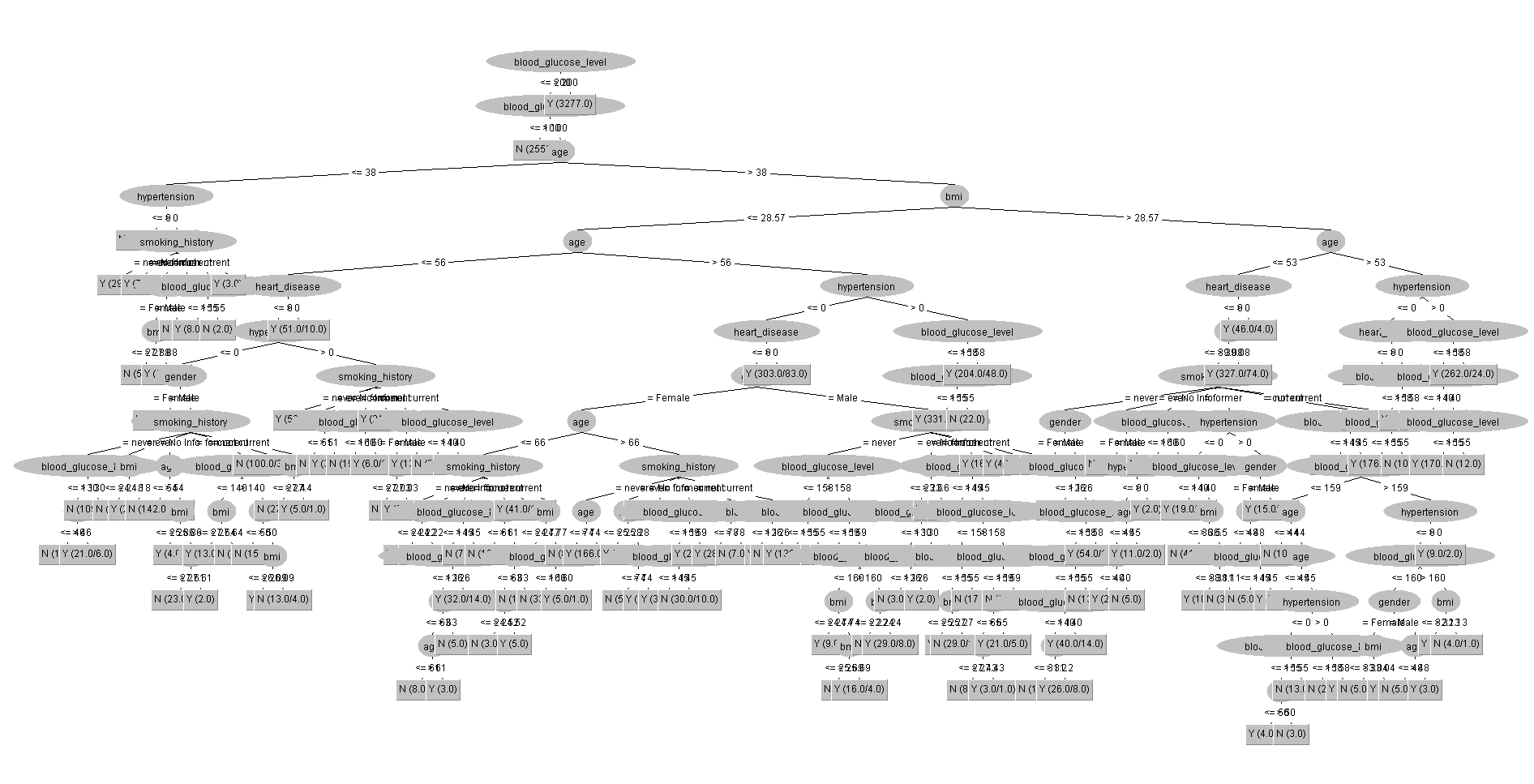
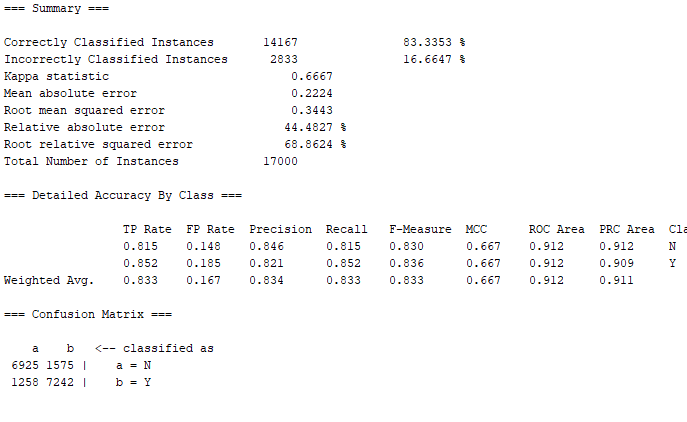
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Dataset** | **Algorithm** | **F-score** | **RMSE** | **Kappa** |
| 1 | Diabetes-N | J-48 (Cross Validation) | 0.833 | --- | 0.6667 |
| 2 | Diabetes-N | NN (Cross Validation ) | 0.831 | --- | 0.6627 |
| 3 | Diabetes-N | Logistic regression (Cross Validation) | 0.815 | --- | 0.63 |
| 4 | Diabetes –O | Linear Regression (Cross Validation) | --- | 0.3675 | --- |
| 5 | Diabetes –O | NN (Cross Validation ) | 0.787\*\* | 0.3946 | 0.55\*\* |
| 6 | Diabetes –O | Rep-Tree (Cross Validation ) | \*\*0.8255 | 0.3469 | \*\*0.65 |
| 7 | Diabetes –O | Random Forest CV | --- | 0.3474 | --- |
| 8 | Diabetes –O | IBK CV (KNN) | --- | 0.4609 | --- |
| 9 | Diabetes-N | SMO CV (SVM) | 0.818 | --- | 0.6353 |

\*\* ערך מחושב ולא מבוסס על הפלט של תוכנת Weka. נספח 5.1.2.2+ נספח 5.1.2.3. בNN טווח בסיווג הוא 0.3 ו REP טווח הסיווג הוא 0.5

* התוצאה הטובה בשלב השלישי

בשלב השלישי לאחר סקירת התוצאות ראינו שמודל עץ החלטה מסוג

J-48 (Cross Validation) השיג את ההרצה הטובה ביותר, עם שדה מטרה נומינלי ומדדי הערכה F-score= 0.833וKappa0.6667.

* עץ ההחלטה שהתקבל בתוכנת ה Weka בשלב השלישי:
* פלט הערכים שהתקבל בתוכנת ה Weka בשלב השלישי:
  1. פלט מפורט של ההרצה הטובה ביותר:

לאחר סקירת התוצאות ראינו שבלב השלישי לאחר טיפול בקבוצת המיעוט והורדת המשתנה המתווך המודל הטוב ביותר הוא עץ החלטה מסוג **J-48 (Cross**

**Validation)** אשר השיג את ההרצה הטובה ביותר, עם שדה מטרה נומינלי ומדדי הערכה F-score= 0.8333 וKappa= 0.6667. (ראה בשלב השלישי את פלט הערכים שהתקבל בתוכנת הWeka)

* 1. הרצה עם שדה מטרה מספרי

מבין כל ההרצות עם שדה מטרה מספרי, הרצה Rep-Tree(CV) הייתה עם הRMSE הכי נמוכה .למרות זאת, בחישוב הF-score ההרצה קיבלה

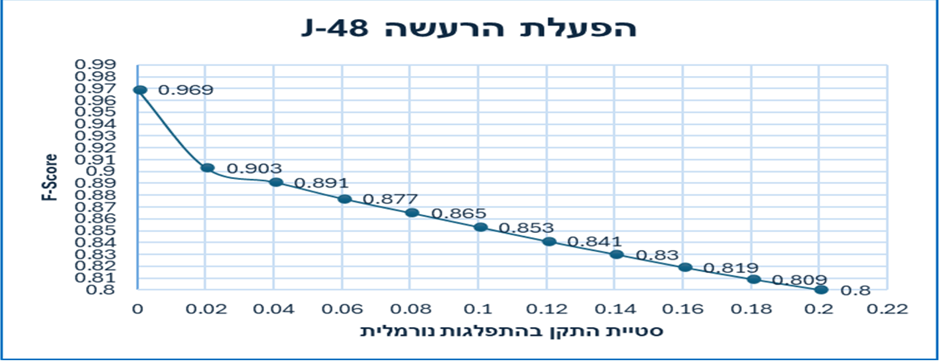
F-score=0.8255. ובהשוואה להרצה של J-48 (CV) בה קיבלנו

F-score=0.833.

* 1. בדיקת יציבות המודל
* בשלב הראשון:

מבדיקת חשיבות המשתנים, עולה כי משתנים HbA1c\_level ו Blood\_glucose\_level בעלי ההשפעה הגבוהה ביותר במודל (נספחים 5.4.1+5.4.2+5.4.3+5.4.4) , ולכן עליהם הופעל הרעש.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlation** | **GainRatio** | **InfoGain** | **OneR** | **Evaluator** |
| 0.4007 | 0.16043 | 0.18723 | 95.395 | HbA1c\_level |
| 0.4196 | 0.0823 | 0.16163 | 94.777 | Blood\_glucose\_level |
| 0.258 | 0.01664 | 0.05873 | 91.5 | age |
| 0.1978 | 0.05079 | 0.01949 | 91.5 | hypertension |
| 0.1717 | 0.05695 | 0.01365 | 91.5 | heart\_disease |
| 0.0654 | 0.00663 | 0.01418 | 91.5 | smoking\_history |
| 0.0376 | 0.00105 | 0.00103 | 91.5 | gender |
| 0.2144 | 0.01306 | 0.03363 | 91.385 | bmi |



המסקנה- המגמה היא אחידה, וככל שאנחנו מגדילים את עוצמת הרעש הF-score יורד, בדיוק כפי שהיינו מצפים.- כלומר המודל יציב. תוצאות הפעלת הרעש מפורטות בפורמט של טבלה- נספח 5.4.5

* בשלב השני (לאחר טיפול בקבוצת המיעוט):

מבדיקת חשיבות המשתנים, עולה כי משתנים HbA1c\_level ו Blood\_glucose\_level בעלי ההשפעה הגבוהה ביותר במודל (נספחים 5.4.1+5.4.2+5.4.3+5.4.4) , ולכן עליהם הופעל הרעש

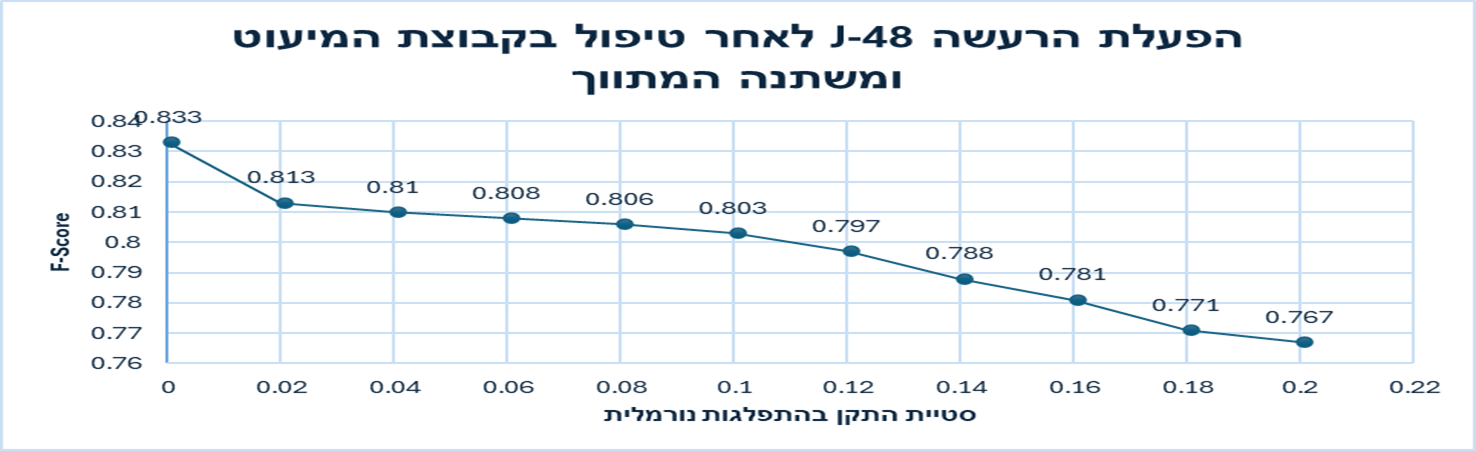
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlation** | **GainRatio** | **InfoGain** | **OneR** | **Evaluator** |
| 0.5955 | 0.30553 | 0.43326 | 72.747 | HbA1c\_level |
| 0.535 | 0.18465 | 0.38309 | 69.718 | Blood\_glucose\_level |
| 0.4784 | 0.06245 | 0.19928 | 71.312 | age |
| 0.2545 | 0.07994 | 0.04951 | 59.182 | hypertension |
| 0.2034 | 0.07329 | 0.03215 | 55.841 | heart\_disease |
| 0.1143 | 0.02147 | 0.04828 | 60.624 | smoking\_history |
| 0.0736 | 0.00396 | 0.00391 | 53.653 | gender |
| 0.3439 | 0.04109 | 0.10938 | 62.776 | bmi |



לאחר הטיפול בקבוצת המיעוט, והרצה של המודלים נבדקה בשנית יציבות המודל- נספח 5.1.7 המגמה הינה אחידה, כלומר המודל יציב.

* בשלב השלישי (לאחר הורדת טיפול בקבוצת המיעוט והורדת המשתנה המתווך):
* מבדיקת חשיבות המשתנים, עולה כי משתנים Blood\_glucose\_level והמשתנה age הם בעלי ההשפעה הגבוהה ביותר במודל (נספחים 5.4.1+5.4.2+5.4.3+5.4.4) , ולכן הופעלו עליהם הרעש

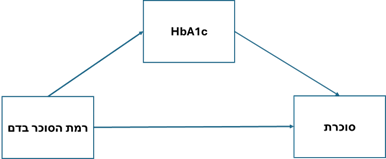
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlation** | **GainRatio** | **InfoGain** | **OneR** | **Evaluator** |
| 0.535 | 0.18465 | 0.38309 | 69.718 | Blood\_glucose\_level |
| 0.4784 | 0.06245 | 0.19928 | 71.312 | age |
| 0.2545 | 0.07994 | 0.04951 | 59.182 | hypertension |
| 0.2034 | 0.07329 | 0.03215 | 55.841 | heart\_disease |
| 0.1143 | 0.02147 | 0.04828 | 60.624 | smoking\_history |
| 0.0736 | 0.00396 | 0.00391 | 53.653 | gender |
| 0.3439 | 0.04109 | 0.10938 | 62.776 | bmi |



לאחר הטיפול בקבוצת המיעוט ובמשתנה המתווך, והרצה של המודלים נבדקה בשנית יציבות המודל- נספח 5.1.7.1 המגמה הינה אחידה, כלומר המודל יציב.

1. **דיון בתוצאות**:
   1. מודל והידע שהתקבל:

* מודל עץ החלטות J-48(Cross Validation) הוא המודל הטוב ביותר לחיזוי סכרת.
* רשת נוירונים (NN) קיבלה כמעט מדדי הערכה זהים לעץ החלטות ולכן לא ניתן להגיד בוודאות שרשת נוירונים אינה מודל טוב לחיזוי.
* בחלק הראשון וכן בחלק השני ע"פ המודל גילינו שרמת ה HbA1cו גלוקוז בדם הם המשתנים הטובים ביותר לחיזוי סכרת. כאשר המשתנים BMI ורמת הסוכר בדם משפעים על רמת הHbA1c, ואילו על רמת הגלוקוז בדם אינו מושפע מאף אחד מהמשתנים השונים.
* רמת ה-HbA1c ורמת הסוכר בדם שניהם מייצגים מצבים פיזיולוגיים הקשורים לניהול רמות הגלוקוז בגוף, שהם מרכזיים להבנת התפתחות הסוכרת. HbA1c מייצג רמות סוכר מצטברות לאורך זמן, בעוד שרמת הסוכר בדם משקפת את המצב המיידי של הגלוקוז בגוף. לכן, השפעתם המשותפת על אבחון סוכרת היא משמעותית והכרחית למודל. בבדיקה האם משתנים אלו מספיקים לאבחון סכרת, למדנו שאין בהם די לצורך האבחנה, שכן גם הf-score והKappa ירדו בצורה משמעותית.
* ייתכן וניתן לקבוע השפעה סיבתית ישירה בין רמת הגלוקוז בדם ורמת ה- HbA1cעל אבחון סכרת. רמת הסוכר בדם יכולה לגרום לסוכרת דרך מנגנונים פיזיולוגיים כמו עמידות לאינסולין או פגיעה בתאי הלבלב, ואילו רמת HbA1c מייצג את המצב הממושך של שליטה על הסוכר, מה שמשקף את הסיכון הכרוני לפיתוח סוכרת.
* קיימת האפשרות שהמשתנה, רמת ה- HbA1, משמש כמתווך חלקי בין רמת הגלוקוז בדם לאבחון סכרת. כלומר, רמת הסוכר בדם משפיעה על רמת ה HbA1c, שבתורו משפיע על אבחון הסוכרת. עם זאת, עדיין יש קשר ישיר בין רמת הגלוקוז בדם לאבחון סכרת שאינו תלוי ברמת ה- HbA1c.



* תוצאות המודל בחלק הראשון ושני היו יחסית טובות ולכן עלינו לוודא שהתוצאה אינה מבוססת על משתנה ספציפי- משתנה מתווך. ולכן החלק השלישי הורדנו את המשתנה המתווך ביחד עם טיפול קבוצת המיעוט וקיבלנו את המודל המדויק ביותר.
* לאחור טיפול במשתנה המתווך (HbA1) ע"י הורדת המשתנה גילנו כי גיל ורמת הגלוקוז בדם הם המשתנים החשובים ביותר
* ניתן לראות לאורך שלוש השלבים המודל הטוב ביותר הוא j48 .
  1. המודל והידע ביחס לספרות:
* תוצאות המחקר בהשוואה לספרות- מאמר 1:
* במאמר 1 ניתן לראות כי למשתנה BMI יש משקל רב בחיזוי סכרת ואילו תוצאות המודל כמעט ולא השתנו ברגע שהורדנו את המשתנה BMI מסט הנתונים. מה שמלמד אותנו שBMI אינו בעל השפעה גבוהה לפי המחקר שלנו. BMI בספרות נבדק כצמוד להיקף המותנים, שככל הנראה משפיע על חשיבות המשתנה, מה שאצלנו לא קיים במחקר.
* רמת הגלוקוז בדם נמצא כמשתנה בעל השפעה במחקר שלנו, הדבר תואם למה שמלמדת הספרות במאמר 1.
* בספרות ניתן היה לראות כי חשיבות המשתנים שונה בין המגדרים השונים, בדקנו את חשיבות המשתנים כאשר המדגם מכיל רק נשים או רק גברים והתוצאות היו כמעט זהות בין שניהם. גם בספרות וגם במחקר שלנו החלוקה בין המגדרים כמעט ושווה (נספח6.2.1) ולכן לא ניתן להסביר את ההבדלים באמצעות גודל המדגם או קבוצת מיעוט, ואילו מה שיכול להסביר את הפערים זה השונות בין המשתנים הנוספים שככל הנראה אינם שונים בהרבה בין המינים.

* תוצאות המחקר בהשוואה לספרות- מאמר 2:
* המאמר קבע כי עץ החלטות הינו המודל הטוב ביותר לחיזוי סכרת, בדיוק כפי שעולה מתוצאות המחקר שלנו. אך עדיין מדדי הערכה של המודל בספרות ובמחקר שלנו שונים, רמת הדיוק אצלנו במחקר היא 83.3% ואילו בספרות 77.87% . ניתן להסביר זאת בגלל שסט הנתונים שלנו שונה הן מבחינת גודל והן מבחינת משתנים נבדקים, נוסיף לזה שבמאמר לא מפורט האלגוריתם הנבדק בספרות.
* בבדיקת חשיבות המשתנים על מודלים שונים במאמר 2, משתנה BMI היה בין המשתנים עם הכי פחות השפעה- מה שמתאים למה שיצא לנו במודל. במהלך המחקר בדקנו את התוצאות כאשר מנטרלים את המשתנה BMI, אין שינוי מהותי בתוצאות ולכן ניתן לקבוע שאין למשתנה השפעה גדולה.
* לעומת זאת, מאמר 2 מלמד אותנו שגיל הוא משתנה הכי חשוב במודל, וכן שלב השלישי שלנו גיל הוא היה משתנה חשוב יחד עם רמת הגלוקוז בדם.
* בספרות הconfusion matrix מלמד כי אין קבוצת מיעוט, ואילו אצלנו היה צורך לטפל בקבוצת המיעוט וכתוצאה מכך מדדי הערכה שלנו נפגעו. ניתן להסביר את הפער באמצעות גודל המדגם והרכב המשתנים במחקר.
* בנוסף בספרות לא ניתן היה לדעת כלל על משתנה מתווך בניגוד למודל שלנו כן היה לנו משתנה מתווך.
* תוצאות המחקר בהשוואה לספרות- מאמר 3:
* המאמר בחן אלגוריתמים שונים בכריית נתונים, כאשר המסקנה הייתה שהמודלים הטובים ביותר לחיזוי סכרת הם SVM, KNN וRF (נספח 3.3). הרצה של מודלים אלו במסגרת המחקר שלנו על סט הנתונים שלנו, קיבלו RMSE נמוכים מעץ החלטות Rep-Tree CV, ולכן תוצאות המחקר שלנו לא משתנות, ועץ החלטות נשאר הדרך הטובה ביותר לחיזוי המחלה במודל שלנו. ניתן להסביר את הפער בשוני בהרכב המשתנים בין המחקרים.
* מאמר 3 מציג 2 סטים עם משתנים זהים אך במדינות שונות, כאשר באף אחד מהם עץ החלטות הוא לא האלגוריתם עם רמת הדיוק הגבוהה ביותר. אולם בסט הנתונים של גרמניה ניתן לראות הAccuracy שקיבלו מאוד קרוב למה שהמודל שלנו השיג, אבל עם Kappa הרבה יותר גבוהה, כלומר המשתנים שנבחנו במאמר יותר יעילים בניבוי סכרת.
* הרצנו את סט הנתונים בתוכנת הweka באותם האלגוריתמים שביצענו על סט הנתונים שלנו (נספח 6.2.2), קיבלנו תוצאות שונות ממה שהספרות מציגה ולכן ניתן להניח שהריצו את האלגוריתמים עם פרמטרים שונים ממה שאנחנו הרצנו בתוכנה. כאשר הרצנו את סט הנתונים מתוך הספרות, מצאנו שהמודל הטוב ביותר לחיזוי סכרת הוא עץ החלטות J-48 (נספח 6.2.3) בדיוק כפי שעולה מהמחקר שלנו- מה שמדגיש את יציבות המודל שלנו. לצורך הדיון חישבנו באופן ידני את הf-score בשביל לוודא שאנו מסוגלים להגיע להסביר את התוצאה שהתקבלה בתוכנת הWeka (נספח 6.2.4).
* הרצה של סט הנתונים בספרות לימד אותנו שסט נתונים שונה עם אותם משתנים בדיוק, מקבל מדדי הערכה שונים. מה שיכול להעיד על יציבות המודל, שהוא רגיש לשינויים בערכי המשתנים וכמות הרשומות. ההרצות בוצעו פעם אחת בצורה של ערבוב בין 2 הסטים של הנתונים ((Dataset Pima/ hospital Frankfurt ופעם אחת סט נתונים אחד אחרי השני, ניתן לראות שמדדי הערכה השתנו בין ההרצות מה שמעיד על כך שיש משמעות לדיוק במודל בסדר הנתונים בקלט.
* בהרצה של סט הנתונים מתוך הספרות, ניתן להבחין בקבוצת מיעוט, כפי שיצא לנו במחקר. לא ניתן למצוא לכך אזכור במאמר, ולכן ישנה האפשרות שהתוצאות שהמאמר מציג הן לאחר טיפול בקבוצת המיעוט.
* במאמר עץ ההחלטות קיבל תוצאות שונות בשני הסטים של הנתונים, כאשר הסט של ההודים הציג רמת דיוק וkappa נמוכים מאוד, ואילו הסט של בית החולים הגרמני הציג רמת דיוק וkappa גבוהים מהמודל שלנו. עדיין הפערים בין האלגוריתמים האחרים היו מהותיים וקבעו שעץ ההחלטות אינו המודל הטוב ביותר לחיזוי.
* סיכום 3 המאמרים:
* לפי הספרות טווח הערכים הרצוי הוא

: Accuracy מכיוון שבמאמר 2 הערך היה 77.87% ובמאמר 3 הערך היה 74% וגם 98.7% - בחרנו במקרה שלנו ב 83.5%.

Recall: מכיוון שבמאמר 2 הערך היה 80.68% ובמאמר 3 הערך היה 74% וגם 98% - בחרנו במקרה שלנו ב 84.2%.

Precision: מכיוון שבמאמר 2 הערך היה 80.75% ובמאמר 3 הערך היה 74% וגם 98.7% - בחרנו במקרה שלנו ב 84.5%.

Kappa: מכיוון שבמאמר 2 הערך היה 55% ובמאמר 3 הערך היה 48% וגם 97% וגם 98% - בחרנו במקרה שלנו ב 74.5%.

* לפי המודל המתקבל שלנו קיבלנו:

Accuracy של 83.3%, Recall של 83.3, Precision של 83.4%ו-Kappa של 66.67%

* כלומר Accuracy Recall Precision קבלנו כמעט כמו בספרות, ואילו Kappa קיבלנו הרבה פחות מהספרות.
* הצגת החידוש והתרומה של המודל שהתקבל ביחס לספרות:
* המודל שלנו מציג מספר חידושים ותרומות משמעותיות ביחס לספרות הקיימת בתחום סיווג חולי סוכרת:
* המודל שלנו השיג תוצאות מרשימות, עם
* Accuracy של 83.3%, Recall של 83.3, Precision של 83.4%ו-Kappa של 66.67%.
* התמודדות עם חוסר ודאות במתודולוגיה: חקר הספרות לא פירט באילו סוגי עצי החלטות השתמשו, ולכן נאלצנו לבצע ניסויים עם מספר סוגי עצי החלטות כדי לבחון את הביצועים.
* בחינת משתנים משמעותיים: המודל שלנו בחן את החשיבות של משתנים שונים כמו רמת HbA1c ,רמת הגלוקוז בדם וגיל, כדי לבדוק את השפעתו על הביצועים.
* שימוש במאגר נתונים נרחב: בניגוד למחקרים רבים שהשתמשו במאגרים קטנים יותר, המודל שלנו נבנה על בסיס מאגר נתונים רחב ומקיף של 100,000 רשומות שלאחר מכן טופל בקבוצת המיעוט והגיע ל 17,000 רשומות.
* גישה חדשנית לנורמליזציה ותקנון הנתונים: ניתוח הנתונים בוצע הן על מאגרי נתונים מנורמלים ומתקננים והן על מאגרי נתונים רגילים, מה שהביא לשיפור הביצועים של המודל
  1. שימוש במודל והידע שהתקבל:

לאור התובנות החדשות שהתקבלו מהמודל שלנו, ניתן להציע מספר שימושים אפשריים:

* מערכות זיהוי מוקדם: שימוש במודל לזיהוי מוקדם של חולי סוכרת במערכות בריאות.
* תמיכה בהחלטות רפואיות: שילוב המודל במערכות תומכות החלטות לצורך אבחון וטיפול בסוכרת.
* מחקר רפואי: שימוש במודל במחקרים רפואיים נוספים להבנת הקשרים בין משתנים שונים לסיכון לחלות בסוכרת.
* חינוך והסברה: שימוש בתוצאות המודל לצורך חינוך והסברה על גורמי הסיכון לסוכרת.
  1. מגבלות העבודה:

ישנם מספר מגבלות שעלינו להיות מודעים אליהם.

* מגבלות המודל: המודל עשוי להיות מותאם יתר על המידה לנתונים עליהם הוא אומן, ולכן ייתכן שלא יפעל באותה רמת דיוק על נתונים חדשים
* מגבלות העבודה: חוסר פירוט בספרות בנוגע לשיטות הספציפיות בהם נעשה שימוש הקשה על השוואה ישירה.
* מגבלות הנתונים: למרות היקפו הגדול של מאגר הנתונים, ישנם משתנים נוספים, שהוגדרו כחשובים בספרות אך לא היו קיימים בסט הנתונים שלנו.
* מגבלות הבעיה: סיווג חולי סוכרת הינה בעיה מורכבת ומגוונת, וייתכן שישנם גורמים נוספים המשפיעים על הסיכון לחלות בסוכרת שלא נבחנו במחקר זה.

לסיכום, למרות שהספרות לא פירטה באילו סוגי עצי החלטות השתמשו, מה שהקשה עלינו לבצע השוואה מדויקת, הצלחנו להשיג תוצאות מרשימות באמצעות ניסויים עם מספר סוגי עצי החלטות. הפערים המשמעותיים בתוצאות מעידים על כך שהמודלים בהם השתמשנו הצליחו לספק ביצועים מעולים בהשוואה למחקרים קודמים. השיפור במדדים עשוי לנבוע מהשימוש במאגר נתונים רחב יותר ושיטות נורמליזציה ותקנון מתקדמות.

1. **רשימת מקורות**

* מאמר 1:

Balkau, B., Lange, C., Fezeu, L., Tichet, J., de Lauzon-Guillain, B., Czernichow, S., ... & Eschwège, E. (2008). Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes care, 31(10), 2056-2061

* מאמר 2:

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99:

* מאמר 3:

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. Bulletin of Electrical Engineering and Informatics, 12(3), 1728-1737:

* סט הנתונים בספרות מתוך מאמר 3:

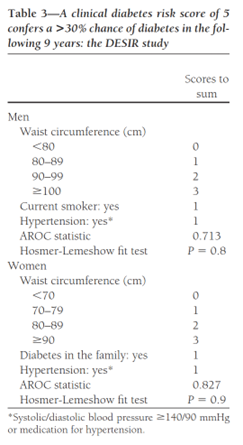


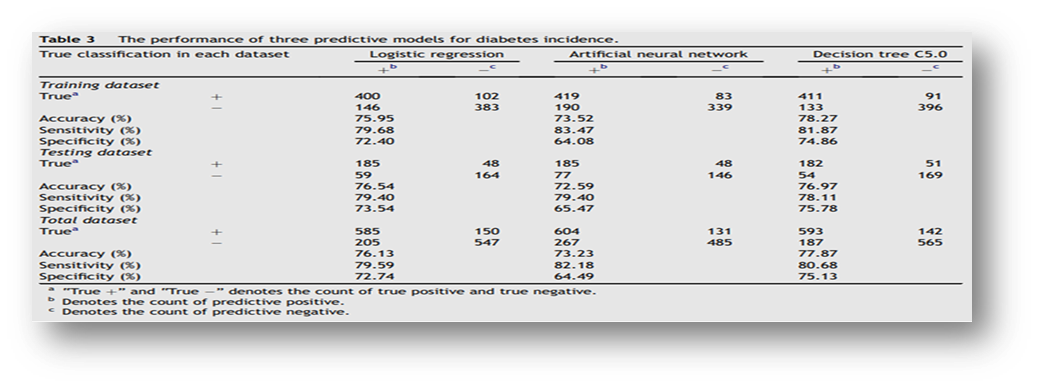
https://www.kaggle.com/datasets/johndasilva/diabetes?resource=download

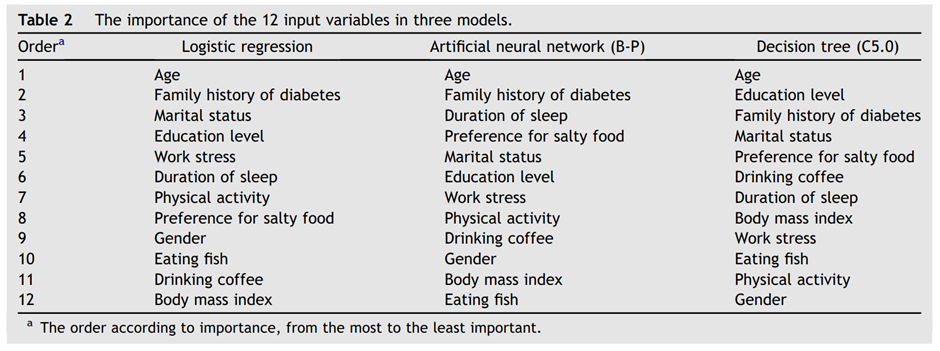
https://www.kaggle.com/uciml/pima-indians-diabetes-database

1. **נספחים:**

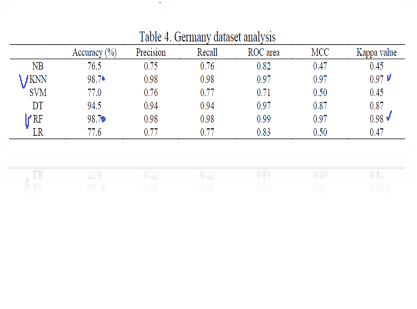
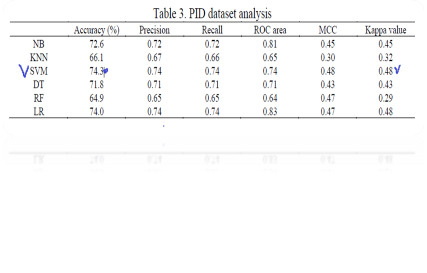
* נספח 3.1: המשתנים המשפיעים על המודל בחלוקה לגברים ונשים



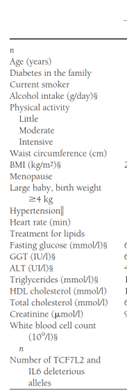
* נספח 3.2.1: הנתונים של 3 מודלים לחיזוי סכרת
* נספח 3.2.2: חשיבות המשתנים במודלים השונים



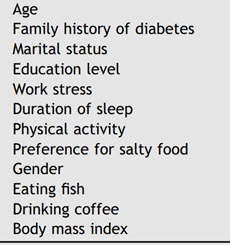
* נספח 3.3: אלגוריתמים שונים בלמידת מכונה לזיהוי סכרת



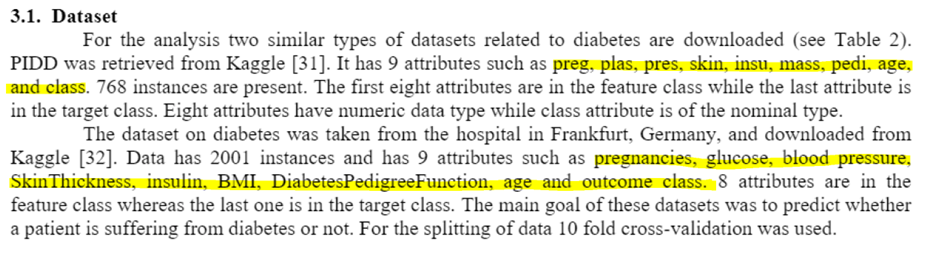
* נספח 4.1.1:משתנים במאמר מס' 1



* נספח 4.1.2: משתנים במאמר מס' 2



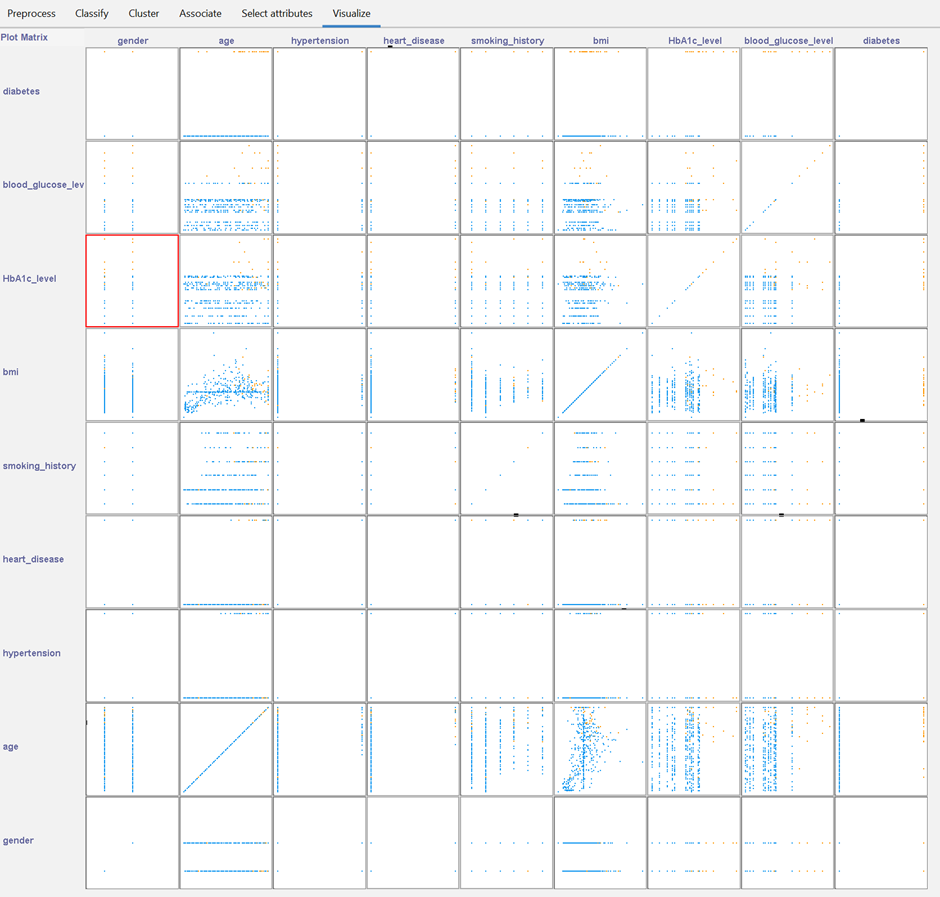
* נספח 4.1.3: משתנים במאמר מס' 3



* נספח 4.2: סקירת המשתנים בספרות- ההשפעה שלהם והזמינות שלהם במחקר

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ניהול הפער | זמינות | השפעה ב% | צורות ייצוג חלופיות | טווח ערכים | שם המשתנה |
| בעל השפעה רק על חצי מאוכלוסיית המדגם ולכן לא נתייחס אליו | אינו זמין | מנבא הטוב ביותר אצל נשים בלבד | 80-90,70-80, יותר מ90 | 70-100 | היקף מותניים |
|  | זמין | מנבא חשוב | 0/1 | ≥ 140/90 | יתר לחץ דם |
|  | זמין | מנבא חשוב אצל גברים בלבד | (0,1)(1,0) | 0/1 | עישון |
| מאמר אחד טען כי ההשפעה היא רק על חצי מאוכלוסיית המדגם, במאמר השני הרוב היה נשים ולכן לא ניתן לקבוע שהגברים הם גם חלק מההשפעה | אינו זמין | מנבא חשוב אצל נשים בלבד/ השפעה גבוה |  | 0/1 | היסטוריה משפחתית של סוכרת |
|  | זמין | מנבא הטוב ביותר | 0-199 | ≥ 7.0 | גלוקוז בצום |
|  | זמין | מנבא הטוב ביותר | מעל 25 או מתחת ל 25 | 18.5-30 | BMI |
| בעל השפעה רק על חצי מאוכלוסיית המדגם ולכן לא נתייחס אליו | אינו זמין | בעל השפעה קלה אצל נשים בלבד | 200-500,15-200 | 150-500 | טריגליצרידים |
| נמצא כלא בעל השפעה על הניבוי, ולכן לא נתייחס אליו | אינו זמין | אינו בעל השפעה | (0,1)(1,0) | 0/1 | פולימורפיזמים גנטיים |
|  | זמין | נמוך | 0/1 | גבר / אישה | מין |
|  | זמין | גבוה מאוד | 21-81 | 20-39, 40-59, 60-79, 80+ | גיל |
| ניתן להשתמש בסטטיסטיקות | אינו זמין | גבוה | (0,0)(1,1)(0,1)(1,0) | נשוי/לא נשוי/אלמן/אחר | מצב משפחתי |
| ניתן להשתמש בסטטיסטיקות | אינו זמין | גבוה | 0/1 | עד 9 שנים, בין 9-12 שנים, יותר מ 12 שנים | רמת השכלה |
| בעל השפעה נמוכה, אבל נוכל להשתמש בBMI כמדד שיכול להעיד על משך השינה של הפרט | אינו זמין | נמוך | ייצוג באחוזים מתוך 24 שעות | עד 5 שעות או בין 5-8 שעות או מעל 8 שעות | משך השינה |
| בעל השפעה נמוכה, אבל נוכל להשתמש בBMI כמדד שיכול להעיד על רמת פעילות גופנית גבוהה או נמוכה | אינו זמין | נמוך | 0/1 | פחות מ-30 דקות ביום או 3 ימים בשבוע 30 דקות או יותר ביום לפחות 3 ימים בשבוע | פעילות גופנית |
| משתנה סובייקטיבי שתלוי בפרשנות הנבדק ולאו דווקא מעיד על מצב אמיתי | אינו זמין | בינוני | 0,1,2 | גבוה, בינוני, נמוך | מצב לחץ בעבודה |
| בעל השפעה נמוכה | אינו זמין | נמוך | 0/1 | פחות מפעם בשבוע או יותר מפעם בשבוע | שתיית קפה |
| בעל השפעה נמוכה | אינו זמין | נמוך מאוד | 0/1 | פחות מ-3 פעמים בשבוע או לפחות 3 פעמים בשבוע | אכילת דגים |
| בעל השפעה נמוכה, אך נוכל להשתמש באינדיקציה של יתר לחץ דם אשר יכולה להעיד על העדפה לאוכל מלוח | אינו זמין | נמוך מאוד | כמות נתרן (מ"ג) | 0/1 | העדפה לאוכל מלוח |
| ניתן להסתמך על סטטיסטיקות | אינו זמין | לא ידוע | 0-3,4-8, יותר מ8 | 0-17 | מספר הריונות |
| ניתן להסתמך על המשתנה יתר לחץ דם | אינו זמין | לא ידוע | פחות מ80/80, יותר מ80 | 0-122 | לחץ דם דיאסטולי (מ"מ כספית) |
| לא ניתן לגשר על פער | אינו זמין | לא ידוע | נמוך/בינוני/גבוה | 0-99 | עובי העור (מ"מ) |

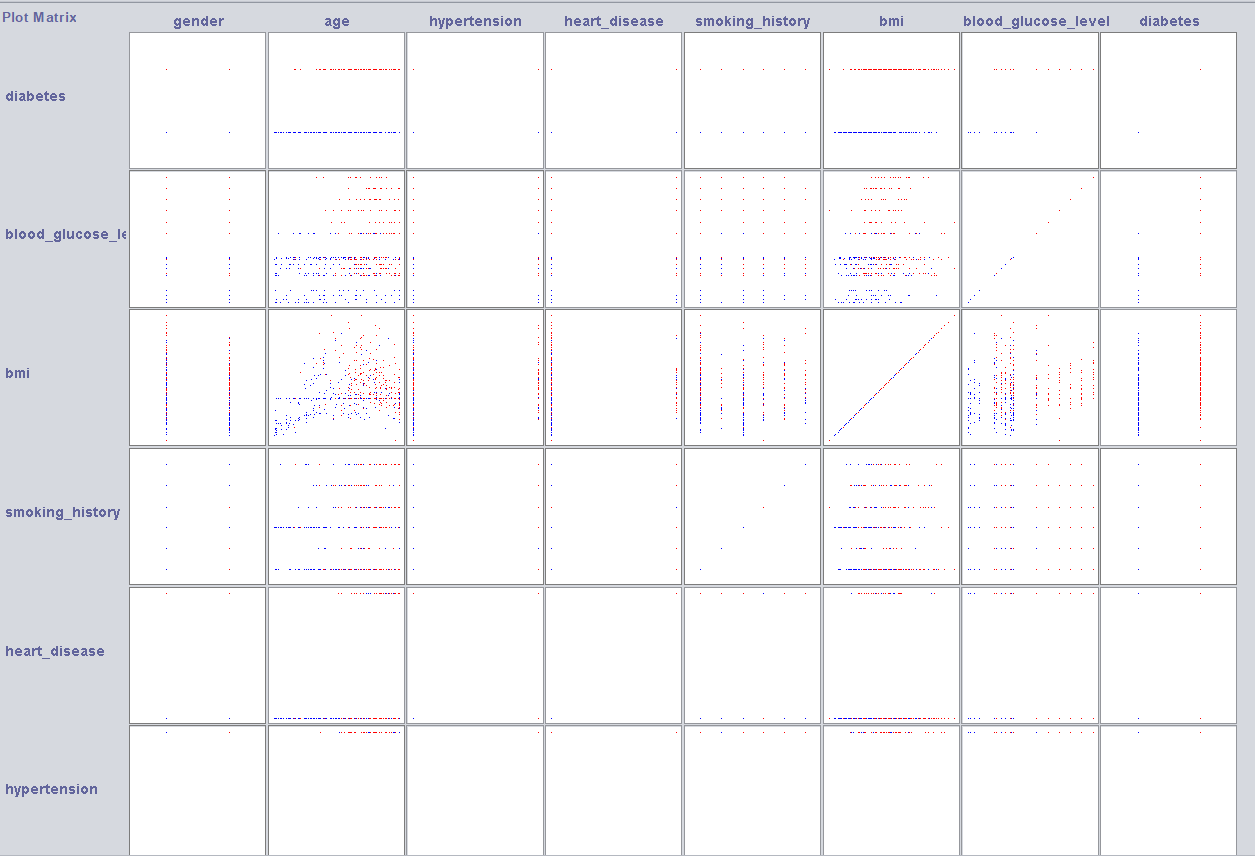
נספח 4.3: בדיקת קשרים או תלות בין המשתנים

בשלב הראשון:

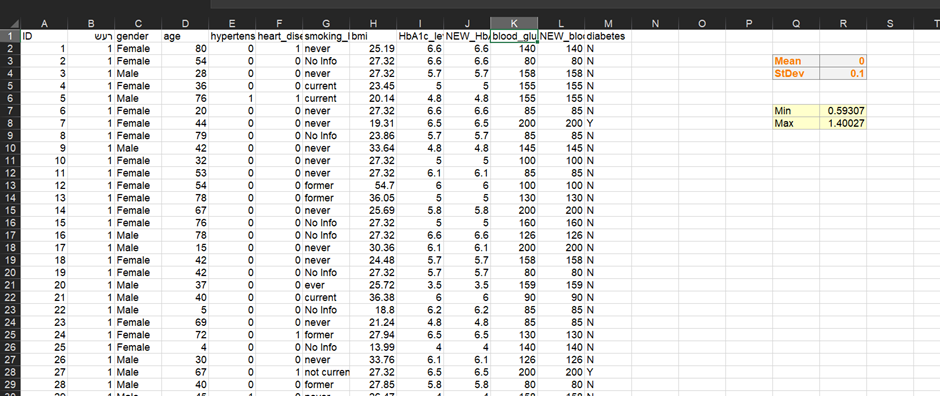
בשלב השני ( לאחר טיפול בקבוצת המיעוט):

תמונה שמכילה טקסט, צילום מסך, מספר, מקביל

התיאור נוצר באופן אוטומטי

בשלב השלישי ( המודל הטוב ביותר- לאחר טיפול במתווך ובקובצת המיעוט):

* נספח 4.6: דוגמא לביצוע ניתוח רגישות באמצעות הפעלת רעש

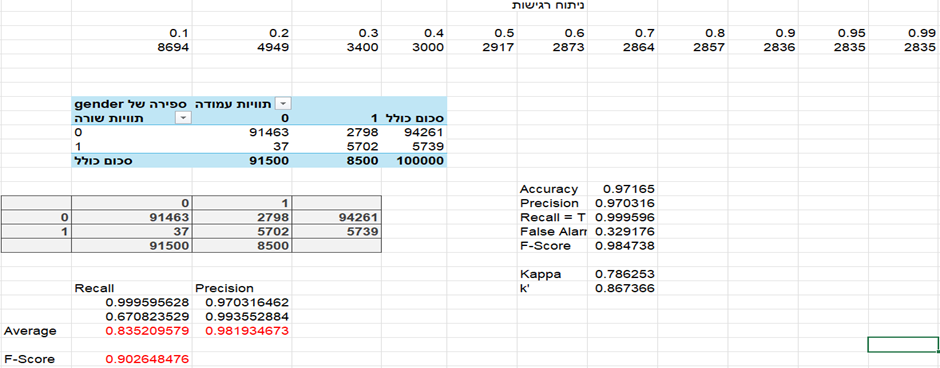


* נספח 5.1.1: מדדי הערכה שהתקבלו בתוכנת הWeka בשלב הראשון

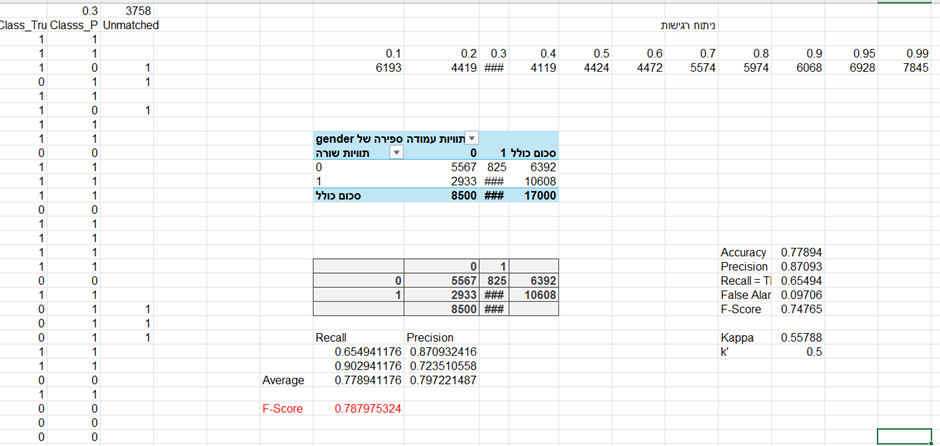
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **Dataset** | **Algorithm** | **F-score** | **RMSE** | **Kappa** | **Accuracy** | **Recall** | **Precision** |
| 1 | Diabetes-N | J-48 (CV) | 0.969\* | --- | 0.7866 | 0.971 | 0.971 | 0.971 |
| 2 | Diabetes-N | J-48 (Training 66%) | 0.967\* | --- | 0.7826 | 0.970 | 0.970 | 0.971 |
| 3 | Diabetes-N | J-48 (Training 66%)- Preserve order for % split | 0.971\* | --- | 0.7904 | 0.973 | 0.973 | 0.974 |
| 4 | Diabetes-N | NN (CV) | 0.969\* | --- | 0.7858 | 0.971 | 0.971 | 0.971 |
| 5 | Diabetes-N | Logistic regression (CV) | 0.957\* | --- | 0.7082 | 0.960 | 0.960 | 0.958 |
| 6 | Diabetes-N | SMO CV (SVM) | 0.957\* | --- | 0.6997 | 0.961 | 0.961 | 0.960 |
| 7 | Diabetes -N  Without BMI | J-48 (CV) | 0.969\* | --- | 0.7871 | 0.972 | 0.972 | 0.973 |
| 8 | Diabetes -N  Without HbA1c\_level | J-48 (CV) | 0.935\* | --- | 0.5465 | 0.946 | 0.946 | 0.946 |
| 9 | Diabetes -N  Without blood\_glucose\_level | J-48 (CV) | 0.946\* | --- | 0.6145 | 0.953 | 0.953 | 0.953 |
| 10 | Diabetes -N  Without HbA1c\_level and blood\_glucose\_level | J-48 (CV) | 0.880\* | --- | 0.055 | 0.915 | 0.915 | 0.883 |
| 11 | Diabetes –O | Linear Regression (CV) | --- | 0.2251 | --- | --- | --- | --- |
| 12 | Diabetes –O | NN (CV) | --- | 0.1884 | --- | --- | --- | --- |
| 13 | Diabetes –O | Rep-Tree (CV) | 0.9026\*\* | 0.1558 | 0.7862\*\* | 0.9716 | 0.8352 | 0.9819 |
| 14 | Diabetes –O  (Normalized 0-1) | Linear Regression (CV) | --- | 0.2251 | --- | --- | --- | --- |
| 15 | Diabetes –O  (Normalized 0-1) | NN (CV) | --- | 0.1847 | --- | --- | --- | --- |
| 16 | Diabetes –O  (Normalized 0-1) | Rep-Tree (CV) | --- | 0.1558 | --- | --- | --- | --- |
| 17 | Diabetes –O )Standardized( | Linear Regression (CV) | --- | 0.2251 | --- | --- | --- | --- |
| 18 | Diabetes –O  )Standardized( | NN (CV) | --- | 0.1864 | --- | --- | --- | --- |
| 19 | Diabetes –O  )Standardized( | Rep-Tree (CV) | 0.9028\*\* | 0.1555 | 0.7865\*\* | 0.9717 | 0.8351 | 0.9824 |
| 20 | Diabetes –O (0,1000) | Rep-Tree (CV) | --- | 225.0537 | --- | --- | --- | --- |

\* יש שונות בין התוצאות ולכן לא ניתן להסתמך על הממוצע. בבחירת החלופה נתייחס לKappa ולF-score בכדי לגשר על כך.  
\*\* ערך מחושב ולא מבוסס על הפלט של תוכנת Weka. נספח 5.1.2+ נספח 5.1.3

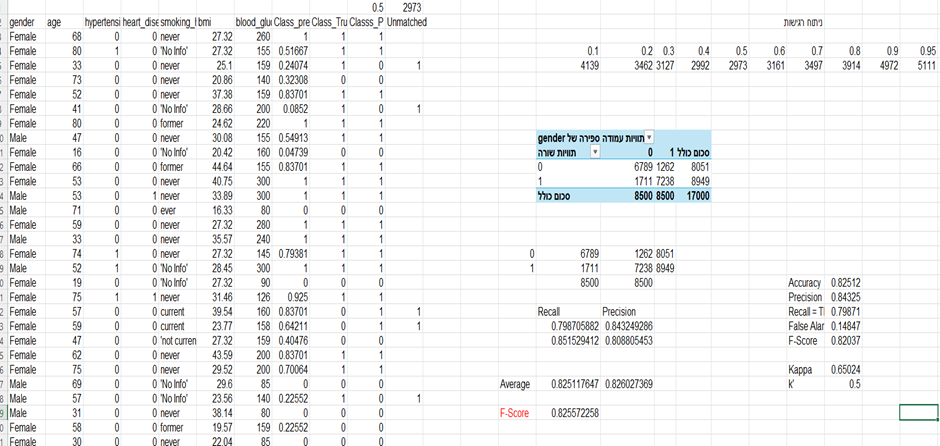
* נספח 5.1.2: חישוב מדדים לשדה מטרה מספרי לאלגוריתם Rep-Tree



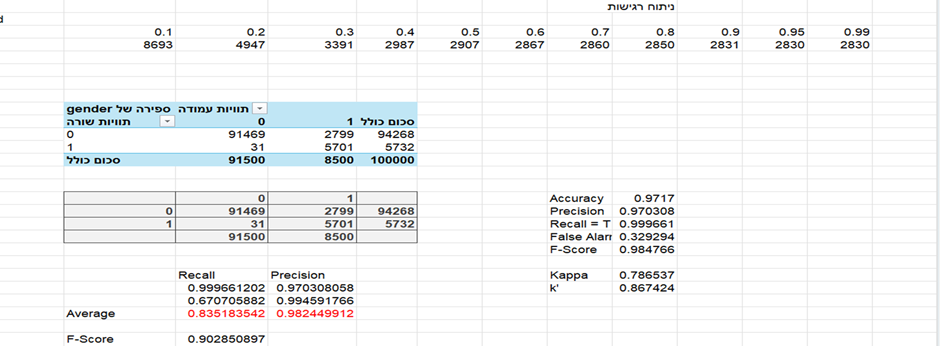
* נספח 5.1.2.2 חישוב מדדים לשדה מטרה מספרי לאלגוריתם NN



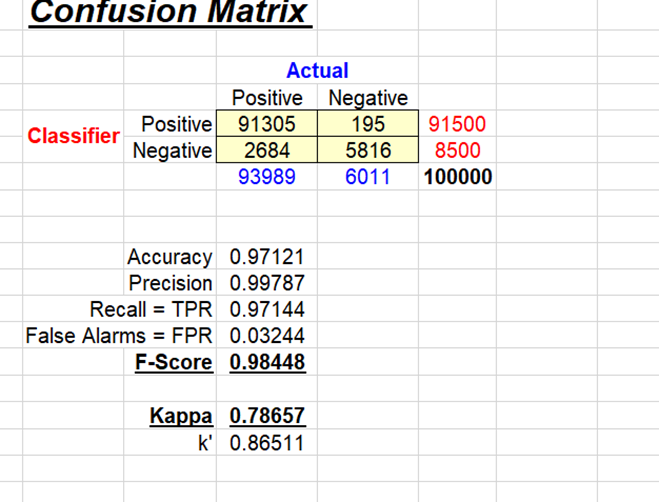
* נספח 5.1.2.3: חישוב מדדים לשדה מטרה מספרי לאלגוריתם Rep-Tree



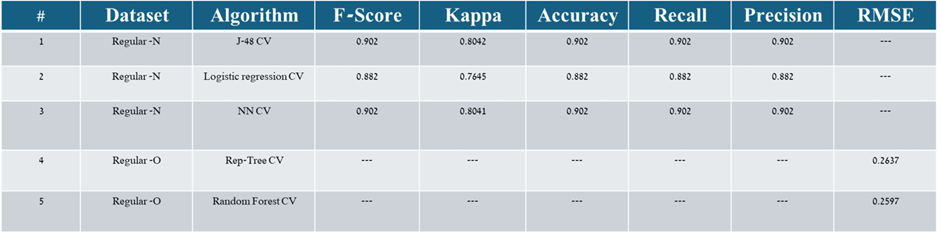
* נספח 5.1.3: חישוב מדדים לשדה מטרה מספרי לאלגוריתם Rep-Tree- לאחר תקנון



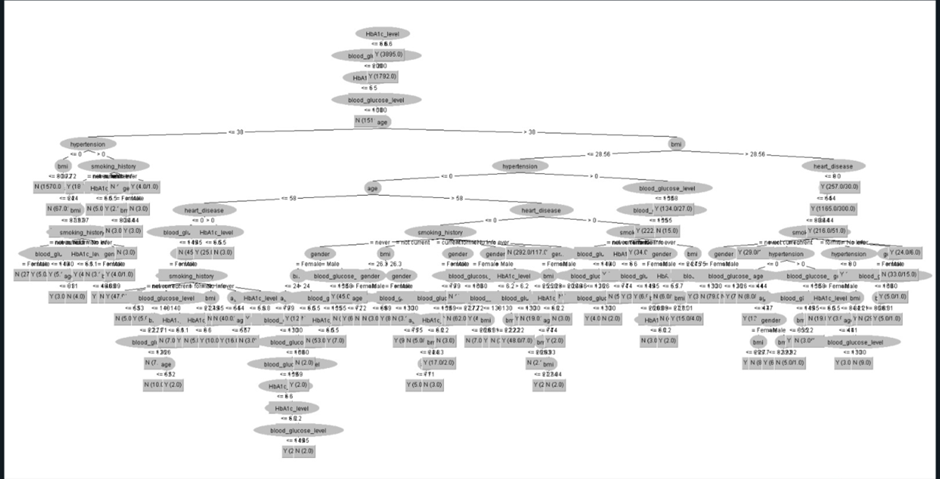
* נספח 5.1.4: דוגמה לבדיקת הConfusion Matrix אל מול התוצאות בWeka



* נספח 5.1.5: ריכוז תוצאות לאלגוריתמים שונים לאחר טיפול בקבוצת המיעוט



* נספח 5.1.6: עץ ההחלטות לאחר טיפול בקבוצת המיעוט



* נספח 5.1.7: בדיקת יציבות המודל לאחר טיפול בקבוצת המיעוט

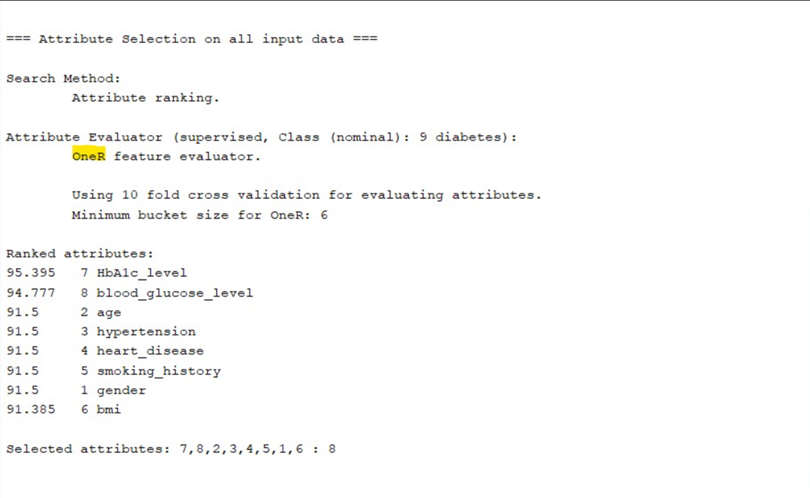


* נספח 5.1.7.1 בדיקת יציבות המודל לאחר טיפול בקבוצת המיעוט ובמשתנה המתווך

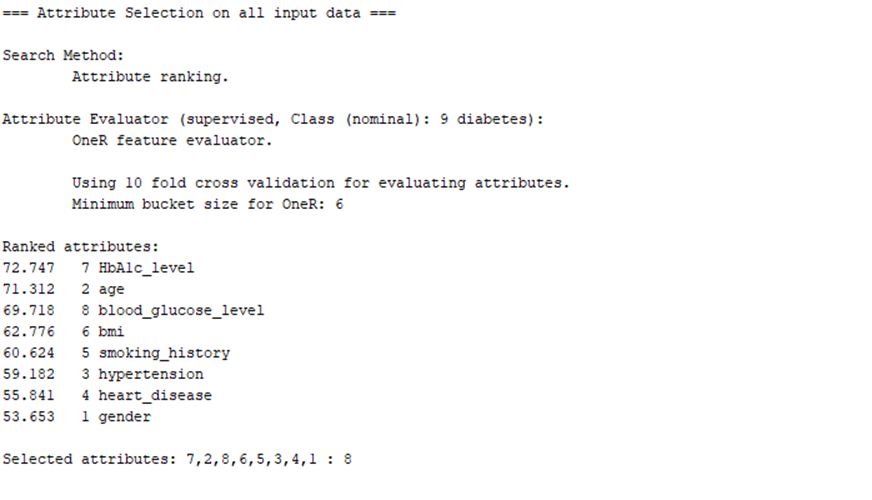


* נספח 5.4.1: בדיקת OneR של המשתנים

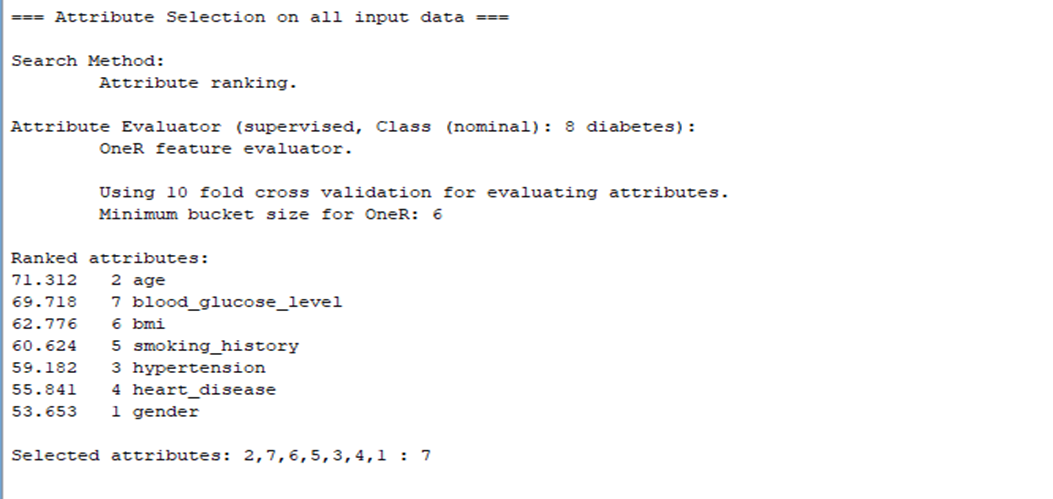
שלב הראשון:



שלב השני – לאחר טיפול בקבוצת המיעוט:

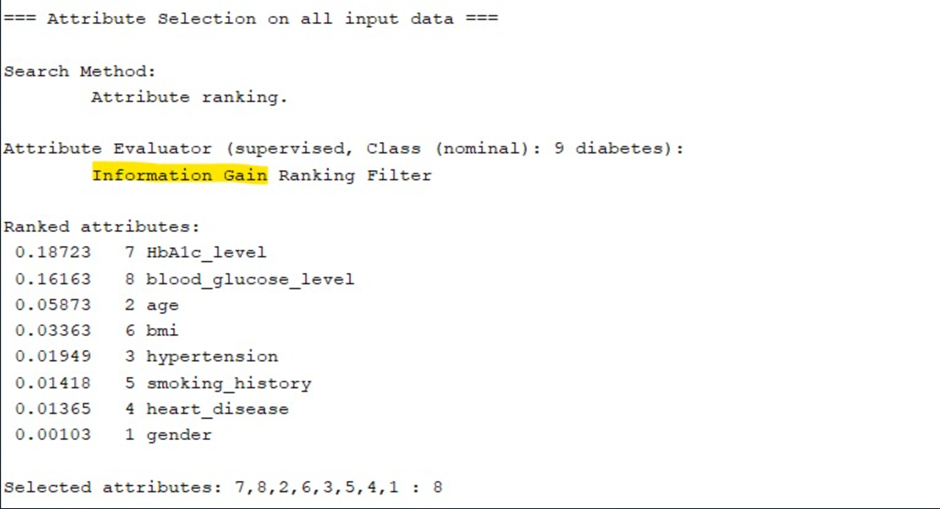


שלב השלישי – לאחר טיפול במתווך וקבוצת המיעוט:

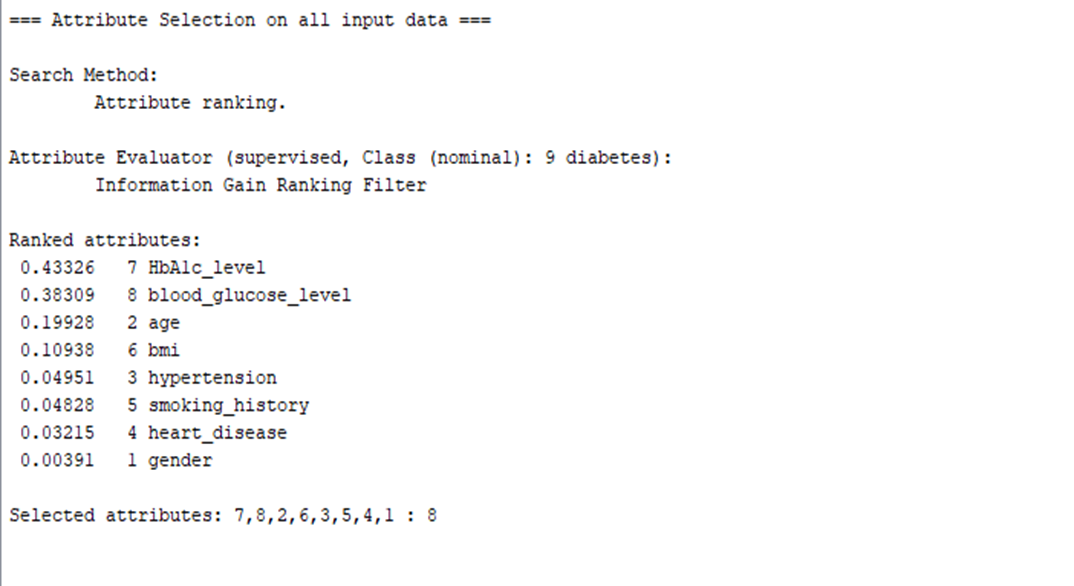


* נספח 5.4.2: בדיקת Information Gain של המשתנים

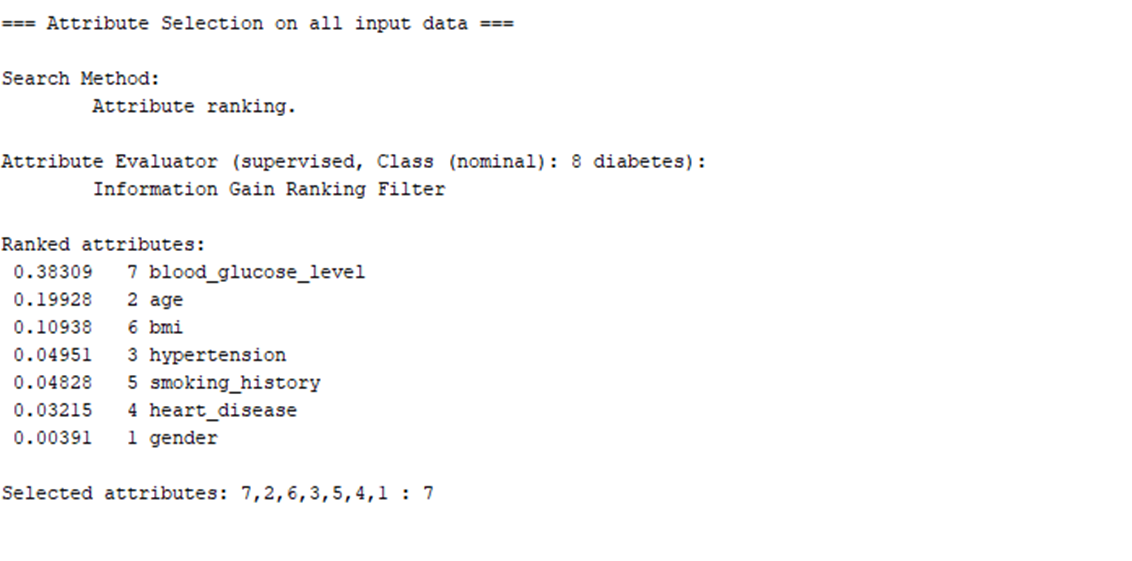
השלב הראשון:



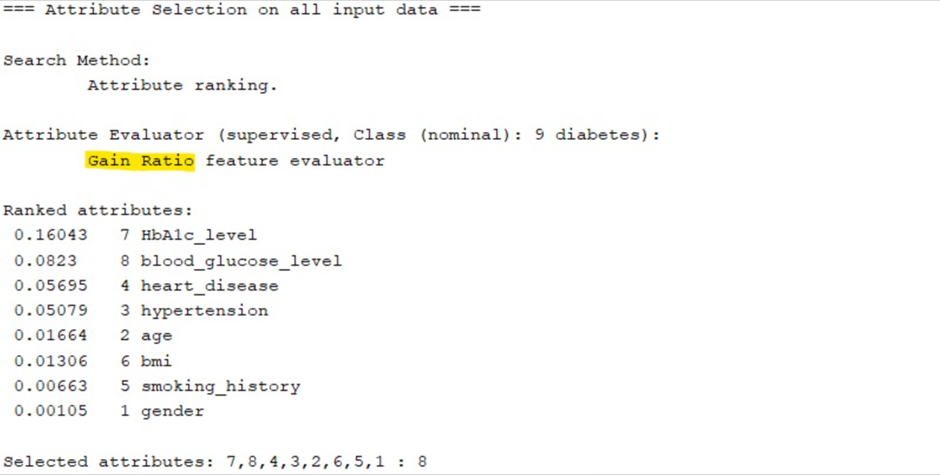
השלב השני – לאחר טיפול בקבוצת המיעוט:



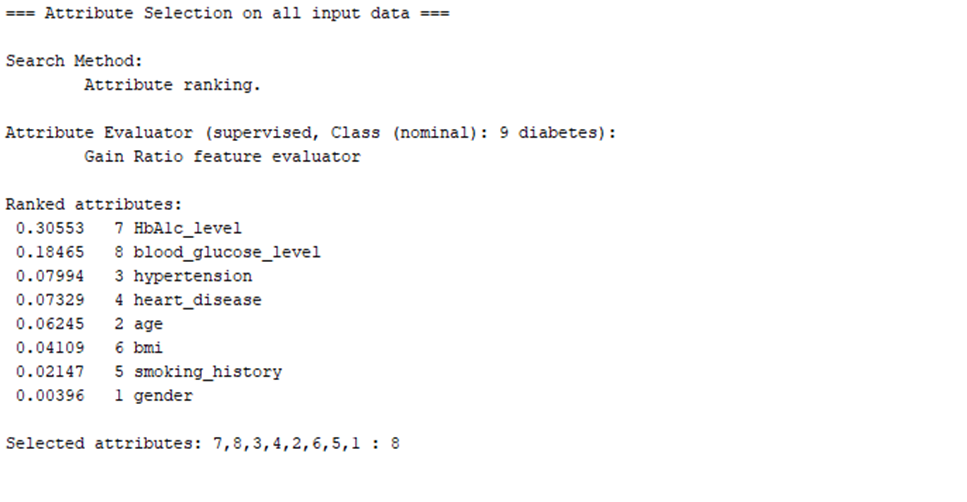
השלב השלישי – לאחר טיפול במתווך ובקבוצת המיעוט:



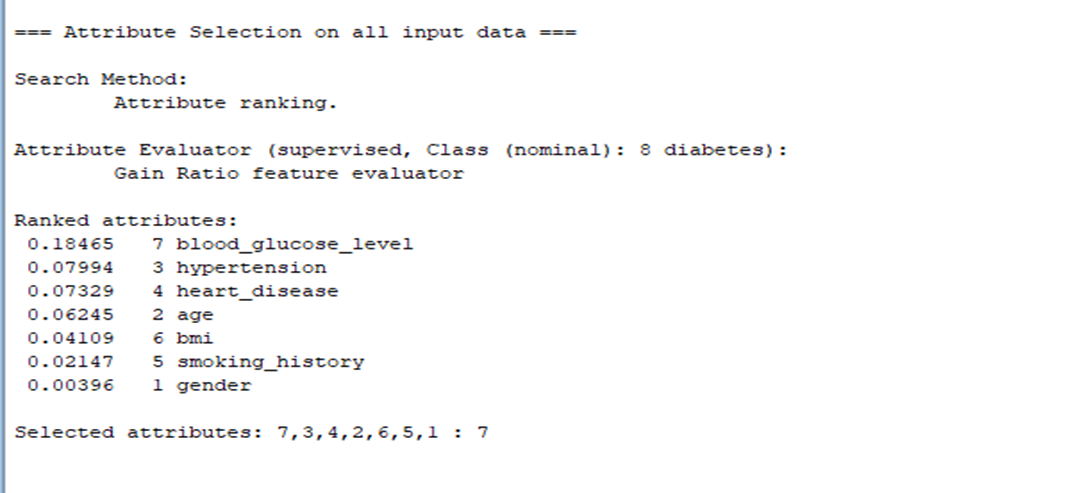
* נספח 5.4.3: בדיקת Gain Ratio של המשתנים

השלב הראשון:

השלב השני – לאחר טיפול בקבוצת המיעוט:

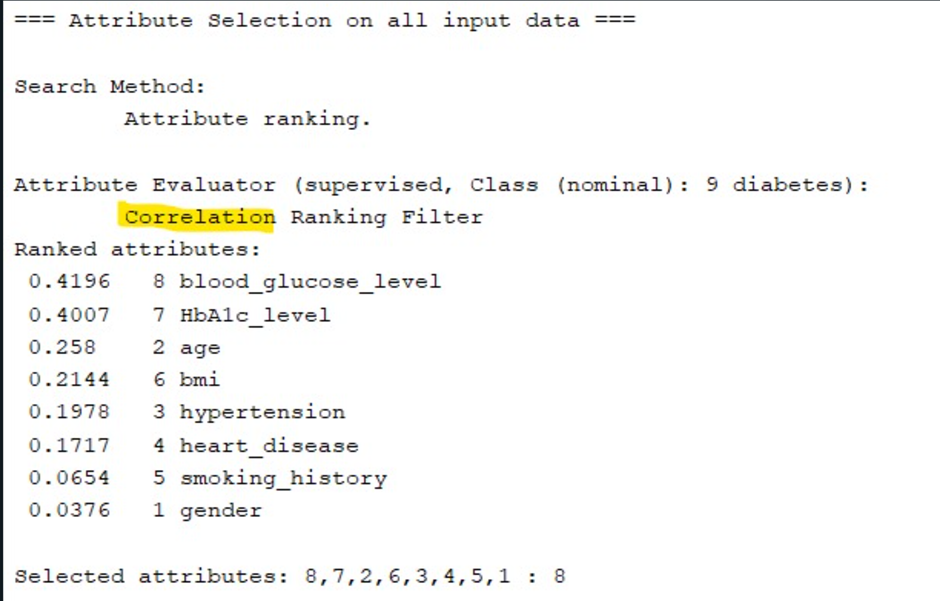


השלב השלישי – לאחר טיפול במתווך ובקבוצת המיעוט:

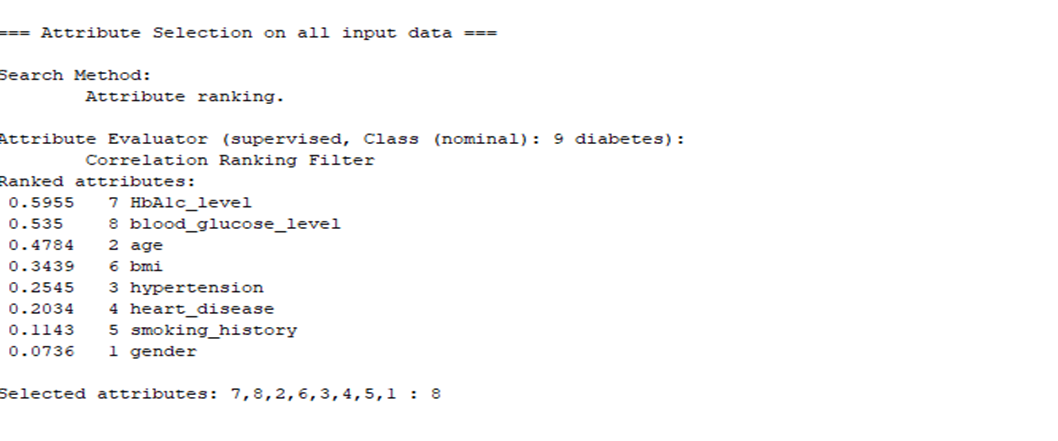


* נספח 5.4.4: בדיקת Correlation של המשתנים

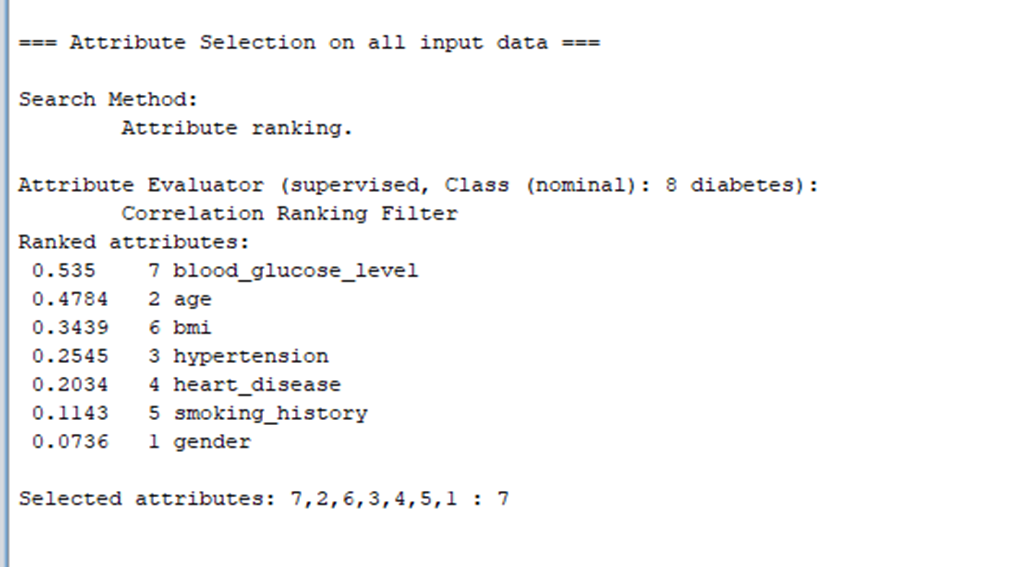
השלב הראשון:



השלב השני – לאחר טיפול בקבוצת המיעוט:



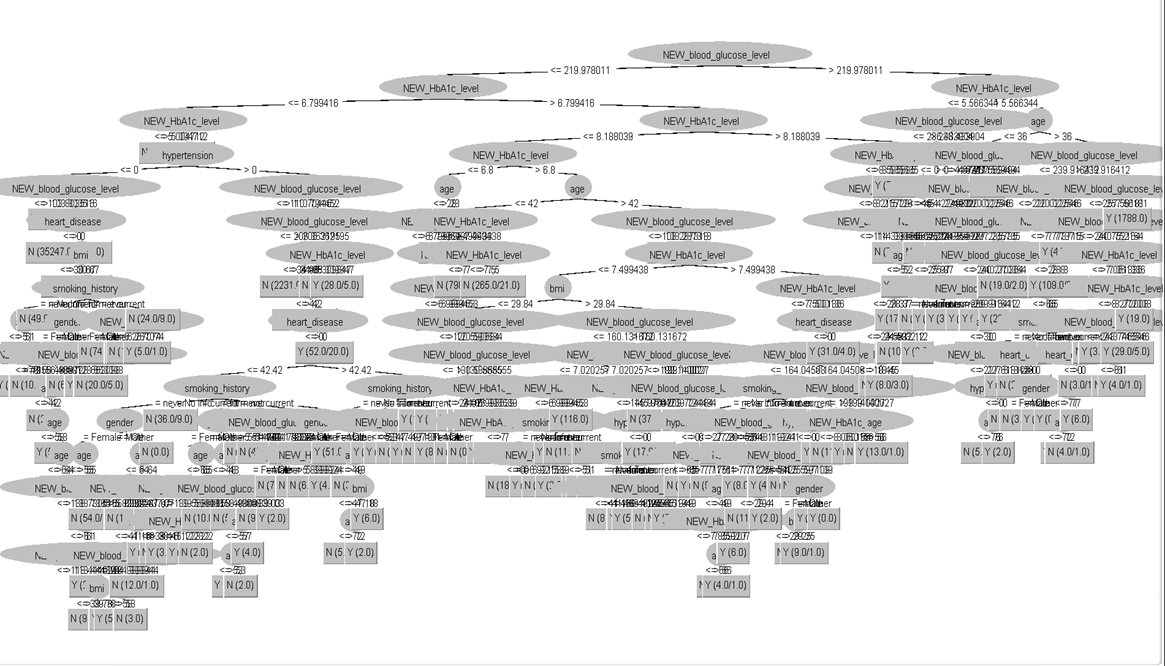
השלב השלישי – לאחר טיפול בקבוצת המיעוט ומשתנה המתווך:



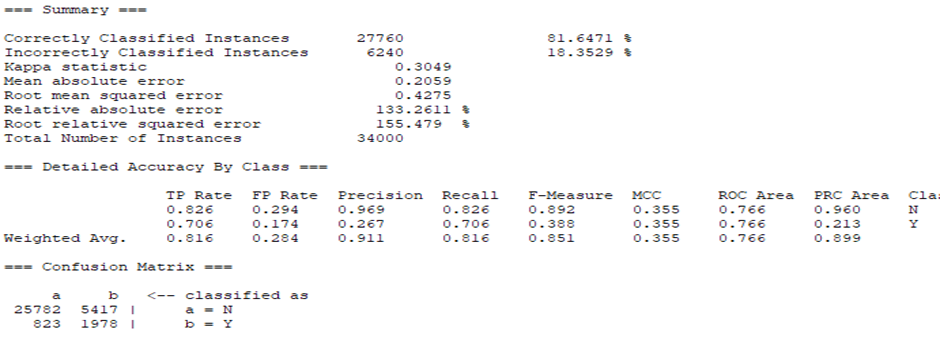
* נספח 5.4.5 טבלת בדיקת רעש על משתנים HbA1c\_level ו blood\_glucose\_level



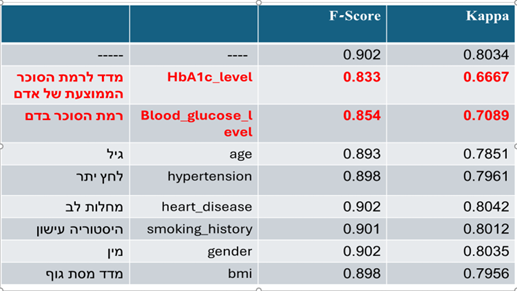
* נספח 5.4.6 מודל העץ לאחר הפעלת רעש של 0.1 בJ-48 (CV) של השלב הראשון:



* נספח 5.4.7 המודל לאחר הפעלת רעש של 0.1 ב J-48 (CV) של השלב הראשון:



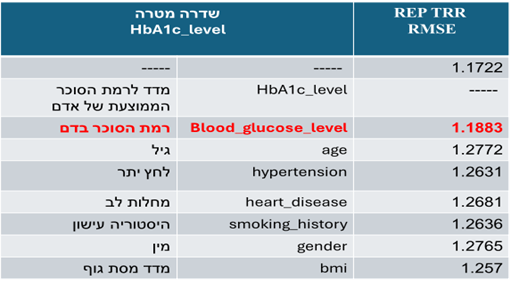
* נספח 5.4.8 שינוי במדדי הערכה בהחסרה של משתנה



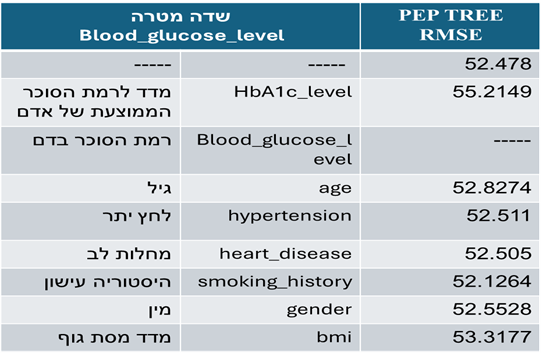
* נספח 5.4.9 השפעת המשתנים על משתנה רמת ה-HbA1c



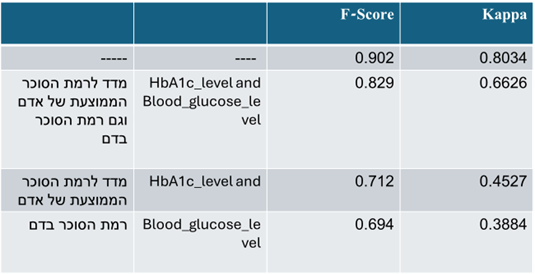
* נספח 5.4.10 השפעת משתנה בודד על משתנה רמת ה-HbA1c



* נספח 5.4.11 השפעת המשתנים על משתנה רמת הגלוקוז בדם



* נספח 5.4.12 מדדי הערכה במודל בהתייחסות למשתנים שנמצאו המשפיעים ביותר



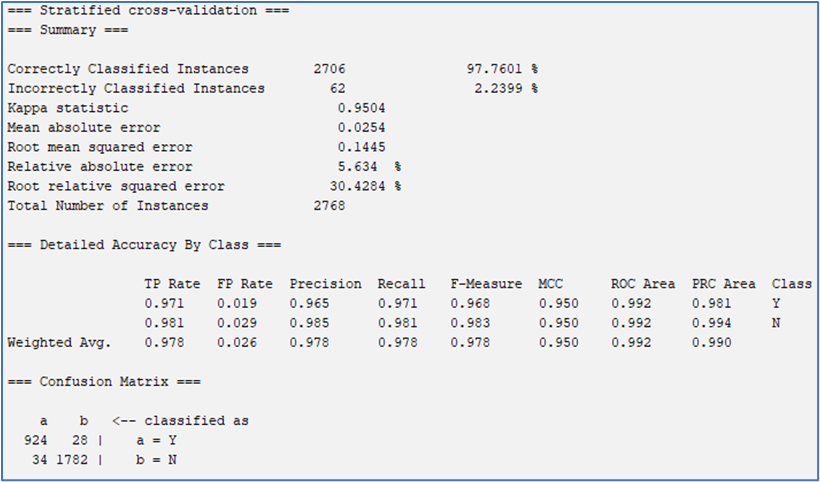
* נספח 6.2.1 חלוקת המדגם בין המגדרים



* נספח 6.2.2 הרצה של האלגוריתמים שנבדקו על סט הנתונים בספרות:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Dataset** | **Algorithm** | **F\_Score** | **Kappa** | **RMSE** |
| 1 | India and Germany | J-48 (Cross Validation) | **0.978** | **0.9504** |  |
| 2 | India and Germany (Unsorted) | J-48 (Cross Validation) | 0.975 | 0.9445 |  |
| 3 | India and Germany | J-48 (Training 66%) | 0.937 | 0.8629 |  |
| 4 | India and Germany (Unsorted) | J-48 (Training 66%) | 0.956 | 0.9032 |  |
| 5 | India and Germany | NN (C.V.) | 0.809 | 0.5766 |  |
| 6 | India and Germany | Logistic Regression (C.V.) | 0.771 | 0.4817 |  |
| 7 | India and Germany without Glucose | J-48 (Cross Validation) | 0.972 | 0.9383 |  |
| 8 | India and Germany without Glucose and BMI | J-48 (Cross Validation) | 0.968 | 0.9294 |  |
| 9 | India and Germany without Glucose and BMI and Age | J-48 (Cross Validation) | 0.967 | 0.9267 |  |
| 10 | India and Germany | Linear Regression (C.V.) |  |  | 0.4011 |
| 11 | India and Germany (Normelized 0-1) | Linear Regression (C.V.) |  |  | 0.3998 |
| 12 | India and Germany | NN (C.V.) |  |  | 0.4385 |
| 13 | India and Germany | Rep-Tree (C.V.) | **0.904** |  | **0.2603** |

* נספח 6.2.3 פלט ההרצה הטובה ביותר על סט הנתונים שבספרות:



* נספח 6.2.4 בדיקת מדדי הערכה בהשוואה לתוכנה:

