אוניברסיטת
אריאל
בשומרון

סיכום וניתוח ממצאים בנושא חיזוי CKD באמצעות למידת מכונה

מגישים:
דור אינגבר
אביה גמרא
איתי בקנשטיין

הקדמה

מחקר זה עוסק בחיזוי מחלת כליות כרונית (CKD) באמצעות שיטות למידת מכונה. המטרה היא להשתמש בנתונים רפואיים שונים כדי לבנות מודלים שיחזו האם מטופל יסבול מ-CKD. המחקר משתמש בטכניקות שונות של עיבוד נתונים, ניתוח סטטיסטי ולמידת מכונה כדי להשיג מטרה זו.

הערה: אנו ממלצים לעבור על Python notebook בעיקר כיוון שהכתיבה והצגת תהליך העבודה מוצגת באופן מלא, מסודרת ומפורטת באופן משמעותי. קישור

סיכום ממצאים

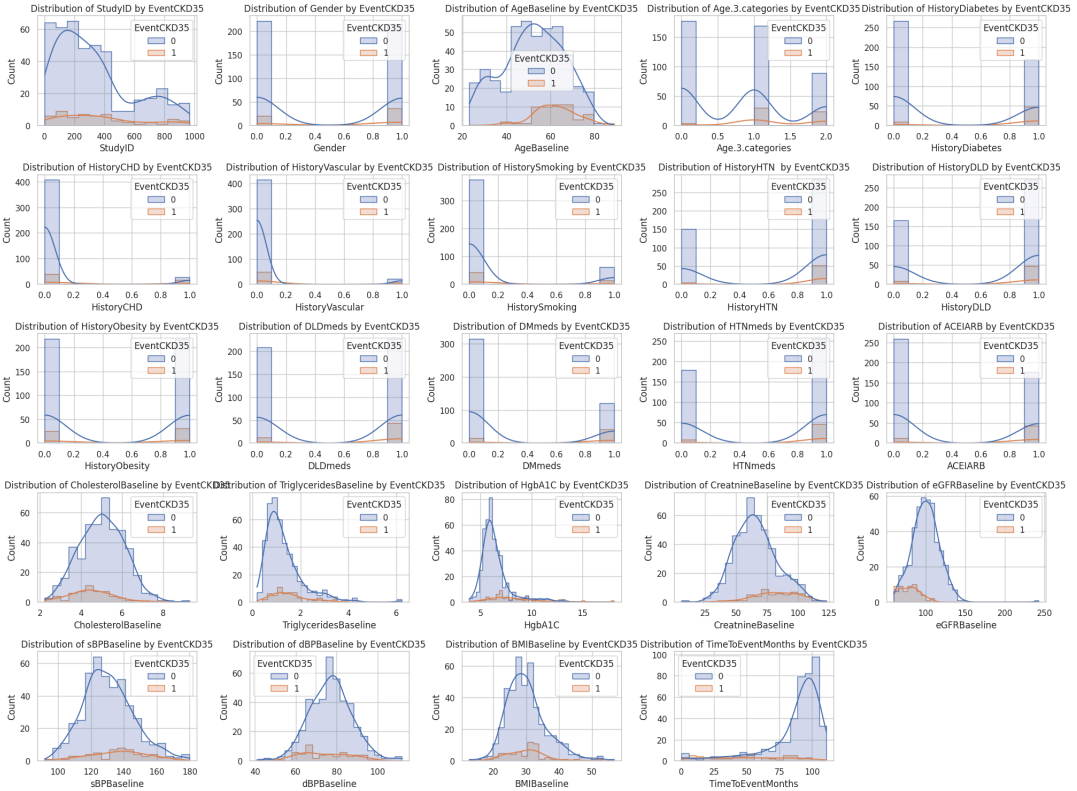
1. עיבוד נתונים:

·ערכים חסרים היו מינימליים למעט בעמודות Triglycerides Baseline ו-HbA1C אשר הושלמו באמצעות הערך הממוצע.

2. ניתוח נתונים חוקר:

שכיחות מחלת כליות כרונית (CKD) נמוכה בסט הנתונים.
ניתוח משתנים שונים הצביע על גורמי סיכון פוטנציאליים ל-CKD.

אוניברסיטת
אריאל
בשומרון



3. ניתוח משתנים:

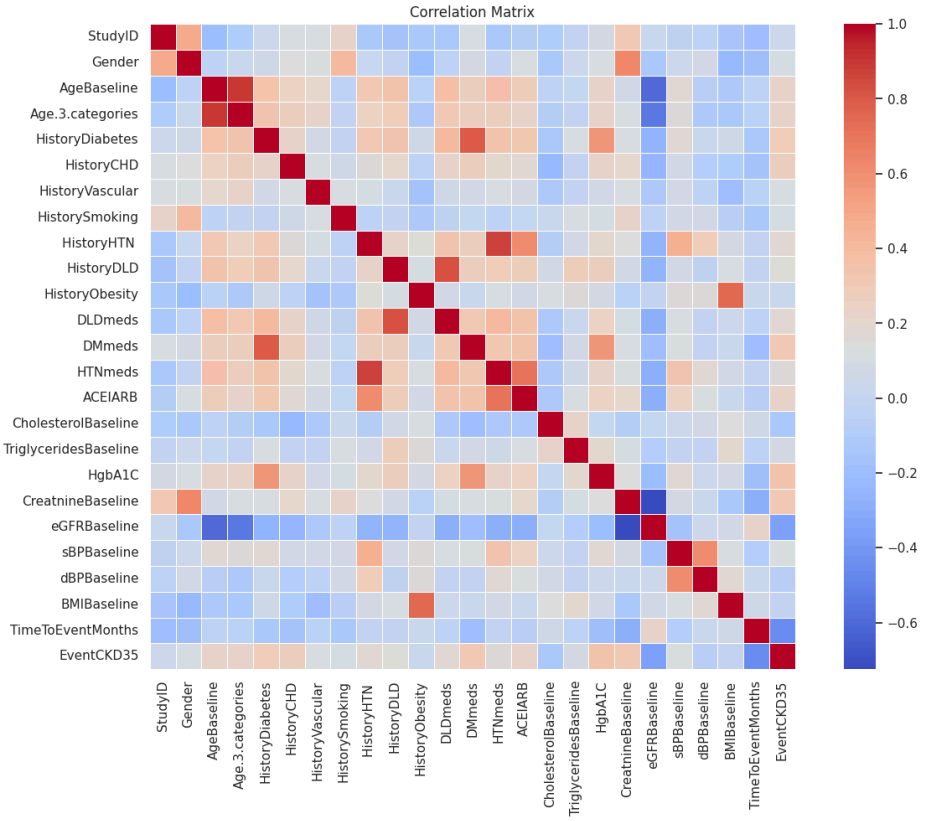
- **גיל:** אנשים מבוגרים נוטים יותר לפתח CKD.
- **מגדר:** נשים נוטות יותר ל-CKD.
- **סוכרת:** יש קשר חיובי ל-CKD.
- **יתר לחץ דם (HTN):** מגדיל את הסבירות ל-CKD.
- **תרופות (ACETARB DLD):** אנשים הנוטלים תרופות אלו נמצאים בסיכון גבוה יותר ל-CKD.

משתנים נומריים:

- **HgbA1C:** ערכים גבוהים יותר בקרב חולי CKD, מה שמעיד על קשר אפשרי לרמות סוכר גבוהות בדם.
- **Creatinine Baseline:** ערכים גבוהים יותר בקרב חולי CKD, אינדיקטור חשוב למחלת כליות כרונית.
- **eGFR Baseline:** ערכים נמוכים יותר בקרב חולי CKD, מה שמעיד על תפקוד כלייתי ירוד.
- **sBP Baseline ו-dBP Baseline:** ערכי לחץ דם סיסטולי ודיאסטולי גבוהים יותר בקרב חולי CKD.

4. ניתוח מטריצת המתאם:

- סיפק תובנות לגבי חוזק וכיוון הקשרים בין המשתנים.



הקשירים בין המשתנים במטריצה מספקים תובנות על חוזק וכיוון הקשר ביניהם:

- **מתאמים חיוביים חזקים**:
  - AgeBaseline ו-Age.3.categories: קטגוריות גיל מתואמות גבוה עם גיל המשתתפים.
  - CreatnineBaseline ו-eGFRBaseline: רמות קריאטינין גבוהות מתואמות עם eGFR נמוך.
  - DMmeds ו-HistoryDiabetes: שימוש בתרופות לסוכרת מתואם גבוה עם היסטוריה של סוכרת.
  - HTNmeds ו-HistoryHTN: שימוש בתרופות ללחץ דם גבוה מתואם עם היסטוריה של יתר לחץ דם.
- **מתאמים חיוביים בינוניים**:
  - HistoryObesity ו-BMI: היסטוריה של השמנה מתואמת עם מדד מסת גוף.
  - DLDmeds ו-HistoryDLD: שימוש בתרופות לדיסליפידמיה מתואם עם היסטוריה של דיסליפידמיה.
  - ACEIARB ו-HistoryHTN: שימוש ב-ACEI או ARB מתואם עם היסטוריה של יתר לחץ דם.
- **מתאמים שליליים**:
  - eGFRBaseline ו-AgeBaseline: ערכי eGFR נוטים לרדת עם העלייה בגיל.
  - CreatnineBaseline ו-AgeBaseline: רמות קריאטינין נוטות לעלות עם העלייה בגיל.

מטריצת המתאם מספקת תמונה על קשרים משמעותיים בין משתנים שונים, המסייעת להבנת הקשרים הבריאותיים והדמוגרפיים.

אוניברסיטת
אריאל
בשומרון

5. מידול והערכה:

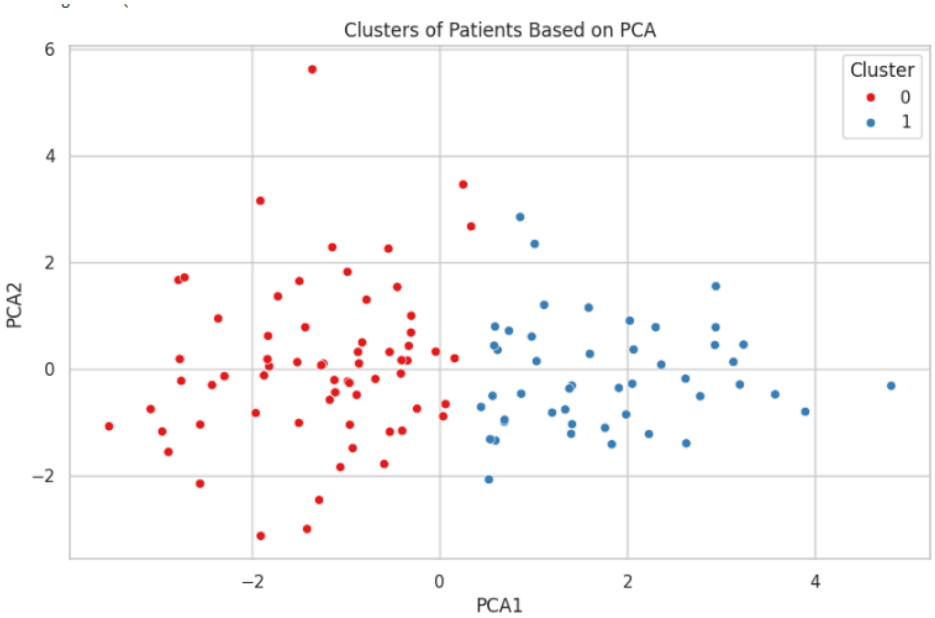
	Best Params	AUC-ROC	Precision	Recall	F1-Score	CV AUC-ROC
Logistic Regression	{'C': 0.1, 'solver': 'lbfgs'}	0.966667	0.875	0.875	0.875	0.87875
Random Forest	{'max_depth': 10, 'min_samples_split': 2, 'n_e...	0.954167	1.0	0.75	0.857143	0.9025
Neural Network	{'activation': 'tanh', 'alpha': 0.001, 'hidden...	0.95	0.7	0.875	0.777778	0.885
XGBoost	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.9	0.714286	0.625	0.666667	0.89375

בהתאם לתוצאות, המודל המומלץ ביותר הוא **Logistic Regression**, מכיוון שהוא מציג את הביצועים הגבוהים ביותר בכל המדדים המרכזיים (AUC-ROC, Precision, Recall, F1-Score). מודל זה מתאים במיוחד כאשר יש חשיבות גבוהה לאיזון בין Precision ל-Recall.

עם זאת, **Random Forest** גם מציג ביצועים טובים מאוד ויכול להיות אופציה מצוינת במקרים בהם חשוב לשמר דיוק גבוה ביותר (Precision) ולמנוע False Positives.

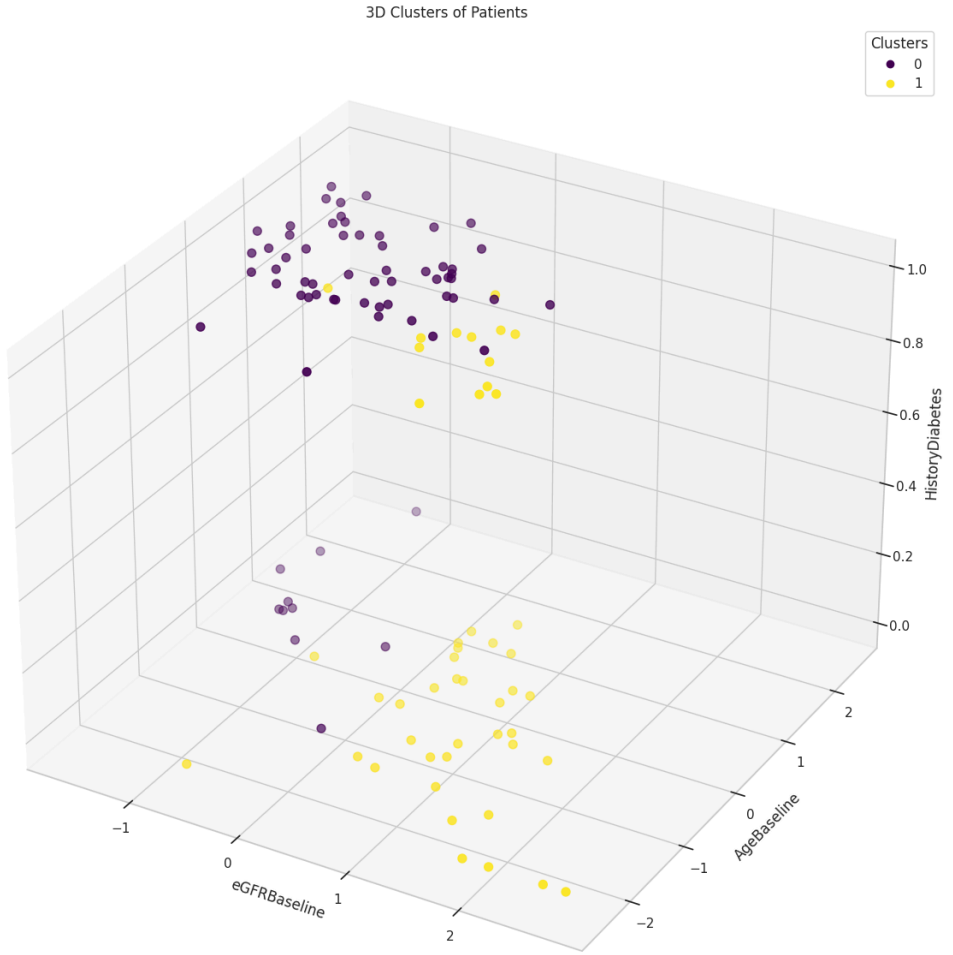
6. ניתוח קבוצות:

בשלב הבא הרצנו מודל קלאסטרינג בכדי לסווג את המטופלים. לאחר ניתוח הביצועים ההמלצה הייתה לקבוע את K=2 U



בניתוח המשתנים המשמעותיים ביותר ניתן לראות את החלוקה לקבוצות המטופלים ואת היחס בי המטופלים החולים, שנוטים להיות במיקום גבוהה יותר בגרף לבין המטופלים הבריאים הנוטים להיות ממוקמים בתחלק

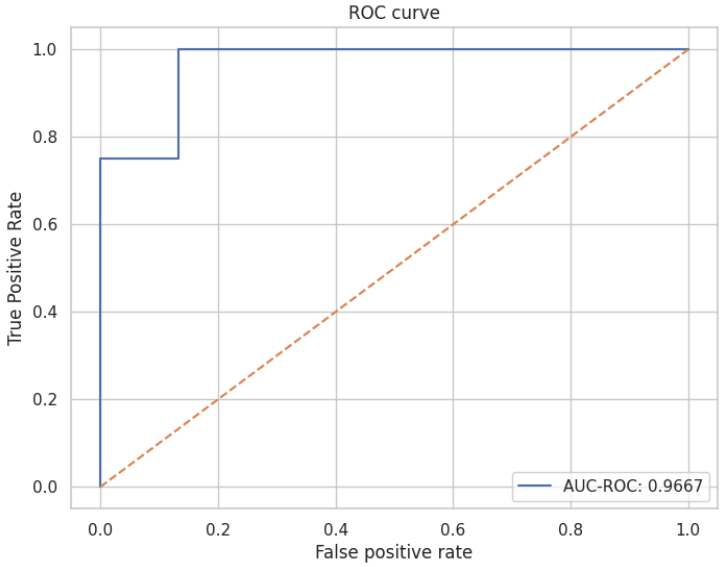
התחתון של הגרף.



לאחר מכן סיננו את כלל המטופלים השייכים לקבוצה 0, הנוטים להיות בעלי סבירות גבוהה יותר לחלות והרצנו פונקציה המבצעת מודל חיזוי לסבירות המחלה

7. תוצאות ניתוח ביצועי המודל

גרף ROC

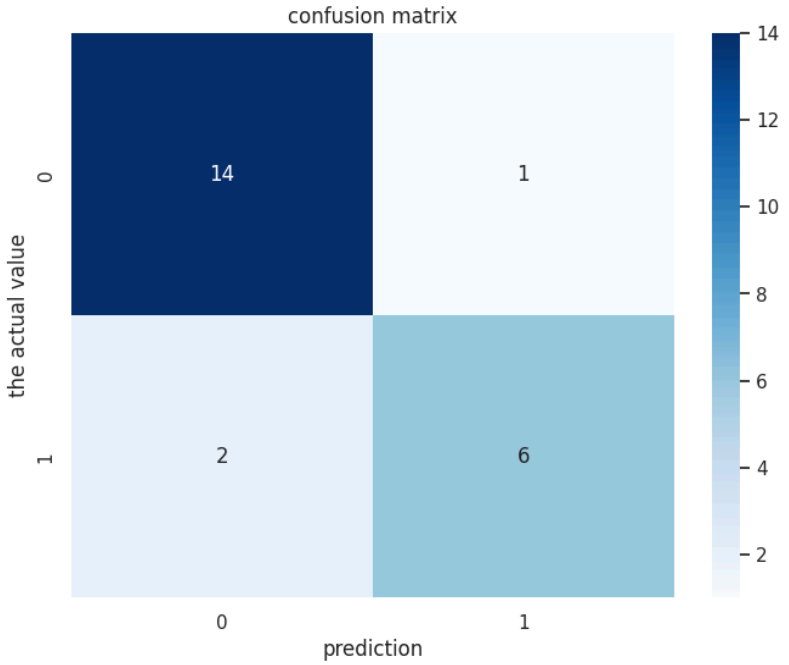


הגרף מציג את עקומת ROC (Receiver Operating Characteristic) עבור המודל:

- **True Positive Rate** (ציר ה-Y): שיעור החיוביים הנכונים מתוך כלל המקרים החיוביים (Recall).
- **False Positive Rate** (ציר ה-X): שיעור החיוביים השגויים מתוך כלל המקרים השליליים.
- **AUC-ROC**: השטח תחת עקומת ה-ROC, הערך הוא 0.9667. ערך זה מעיד על כך שהמודל מציג ביצועים מצוינים במבחן ההפרדה בין הקטגוריות.

Summary of Model Results

- **AUC-ROC**: 0.9667 - מצביע על כך שהמודל טוב מאוד בהפרדה בין הקטגוריות.
- **Precision**: 0.8571 - מצביע על כך ש-85.71% מהתחזיות החיוביות של המודל הן נכונות.
- **Recall**: 0.7500 - מצביע על כך שהמודל מזהה 75% מהמקרים החיוביים.
- **F1 Score**: 0.8000 - ממוצע הרמוני של Precision ו-Recall, מצביע על איזון טוב בין שני המדדים.
- **Actual**: הערך האמיתי של המשתנה המוסבר.
- **Predicted**: התחזית של המודל.
- **Probability**: ההסתברות שהמודל חישב עבור הקטגוריה החיובית.



מסקנות

המודל מציג ביצועים טובים מאוד לפי המדדים השונים (AUC-ROC, Precision, Recall, F1 Score). המודל מצליח להפריד בצורה טובה בין המקרים החיוביים לשליליים ומציג תוצאות מרשימות. התחזיות הראשוניות מראות כי המודל מנבא ברוב המקרים נכון, עם אחוז נמוך יחסית של טעויות.

תובנות שיפורים אפשריים והמלצות

אחת מהבעיות המרכזיות שנמצאו במהלך המחקר היא כמות הנתונים המועטה, אשר כללה רק כ-500 מטופלים. כמות קטנה זו מקשה על בניית מודלים איכותיים ואמינים. הבעיה של מעט הנתונים מובילה גם לקושי בזיהוי קבוצות שונות של מטופלים, מה שמכריח לבצע ניתוח שטחי ולא מעמיק.

בנוסף, קיים חוסר איזון קיצוני בין המטופלים שיש להם CKD לבין אלו שלא, מה שמוביל לכך שהשיטות בהן השתמשנו היו יכולות להיות שונות אם הדאטה היה מאוזן יותר. כדי להתמודד עם חוסר האיזון, ניתן להשתמש בטכניקות כמו SMOTE (Synthetic Minority Over-sampling Technique) על מנת לשפר את יכולת הזיהוי של מחלקות מיעוט.

שיפורים אפשריים:

1. **הנדסת תכונות**: תוספת תכונות חדשות או שינויים בתכונות קיימות עשויים לשפר את ביצועי המודל.
2. **טיפול בנתונים לא מאוזנים**: שימוש בטכניקות כמו SMOTE או טכניקות אחרות לשיפור האיזון בדאטה.

3. **כיוונון היפרפרמטרים**: כיוונון נוסף של היפרפרמטרים במודלים השונים עשוי לשפר את ביצועי המודל במדדים כמו Precision, Recall וביצועים כלליים.

כיווני מחקר נוספים

בנוגע לכיווני מחקר נוספים, ניתן להרחיב על מספר רעיונות חדשים הקשורים ללמידת מכונה וניתוח, חיזוי וזיהוי תובנות עומק במצבים רפואיים שונים. לדוגמה:

1. **גישות למידת מכונה חלופיות**:
   - **למידה בלתי מפוקחת**: חקר טכניקות אשכולות (clustering) לזיהוי דפוסים ללא תוויות מוגדרות מראש.
   - **למידה חצי מפוקחת**: ניצול נתונים מתויגים ולא מתויגים לשיפור ביצועי החיזוי.
   - **שיטות אנסמבל**: שילוב מספר מודלים לשיפור הדיוק והחוסן.
2. **טכניקות מתקדמות**:
   - **למידה עמוקה**: חקר מודלים של למידה עמוקה (deep learning) אשר עשויים לספק ביצועים טובים יותר במערכי נתונים גדולים.
   - **בינה מלאכותית מובנת**: פיתוח מודלים המספקים תובנות לגבי תהליך קבלת ההחלטות על מנת להבין טוב יותר את גורמי הסיכון ל-CKD.
3. **מחקרים לאורך זמן**:
   - ביצוע מחקרים לאורך זמן על מנת לעקוב אחר התפתחות CKD ושיפור מודלים חיזוי בהתבסס על נתונים זמניים.

הצעה למחקר נוסף: ביצוע מחקר נוסף שיתמקד בהגדלת סט הנתונים על ידי איסוף נתונים ממקורות נוספים או בשיתוף פעולה עם מוסדות רפואיים נוספים. כמו כן, שילוב נתוני תמונה (כמו תמונות MRI או אולטרסאונד) יכול לשפר את יכולות החיזוי של המודל באמצעות שימוש בטכניקות של עיבוד תמונה ולמידת מכונה.

המשך המחקר והשיפורים המוצעים יכולים להוביל לשיפור משמעותי במודלים לחיזוי CKD ובכך לסייע בשיפור הטיפול והמעקב אחר מטופלים.