**סמינר למידה חישובית:** חיזוי סרטן השד

מגישות: ליאת שגיא 318989654 , פיליס מיאן 327121463

קישור לסט הנתונים ב Kaggle:

<https://www.kaggle.com/datasets/tanshihjen/coimbra-breastcancer>

קישור למחברת שלנו ב Kaggle עם פירוט נוסף והסברים על כל שלב בעבודה:

<https://www.kaggle.com/code/liatsagi/seminar1>

קישור ל Github:

1. **תקציר:**

עבודה זו בחנה את הפוטנציאל של אלגוריתמי למידת מכונה וניתוח נתונים ממאגר Coimbra Breast Cancer לחיזוי סרטן השד. בתחילה נעשה שחזור של אלגוריתם KNN ממאמר שקראנו, תוך שימוש ב-4 תכונות מרכזיות. לאחר מכן נבדקה יעילות של שילוב בין אלגוריתם KNN וטכניקת PCA (Principal Component Analysis) בניבוי סרטן השד. בתוצאות הראשוניות, נמצא כי הדיוק של KNN עם 7 שכנים ורק 4 פיצ'רים הוא 83%, בעוד הדיוק של המאמר הוא 87.5%. לאחר שימוש ב-PCA להפחתת ממדיות הנתונים ובעזרת שני רכיבי PC ו-5 שכנים, התקבל דיוק זהה לזה של המאמר, 87.5%.

העבודה מציגה את יתרונות השימוש בשילוב של טכניקות למידת מכונה ולהפחתת ממדיות בנתונים עבור ניבוי סרטן השד, ומצביע על חשיבות הבחירה המדויקת של הפרמטרים ועל יכולת המודל לזהות את הפרטים החשובים ביותר בנתונים.

1. **מבוא:**

סרטן השד, המחלה הנפוצה ביותר בנשים, מתאפיינת בצמיחה לא מבוקרת של תאים לא תקינים בשד. זיהוי מוקדם חיוני להישרדות, ובדיקות דם, בשילוב עם למידת מכונה, מציעות כלי אבחון יעיל וזמין.

במסגרת עבודה זו, בחנו את הפוטנציאל של אלגוריתמי למידת מכונה: KNN, PCA לחיזוי נוכחות סרטן השד. בחלק הראשון של העבודה התמקדנו בשחזור של תוצאות אלגוריתם KNN כפי שפורט במאמר שקראנו, תוך שימוש במאגר הנתונים Coimbra Breast Cancer, המכיל נתונים של 116 נשים (58 עם סרטן שד ו-58 בריאות) ו-10 בדיקות דם שונות.

בהמשך, בחנו את השימוש בטכניקת PCA (Principal Component Analysis) להפחתת ממדיות הנתונים ולשיפור הביצועים. אנו מציגות תוצאות שיצאו לנו בשילוב בין KNN לPCA בניבוי סרטן השד.

לאחר מכן, אנו משוות את התוצאות שלנו לתוצאות המקוריות של המאמר המחקרי. ולבסוף, אנו מסיקות מסקנות ומציעות המלצות לעבודה נוספת.

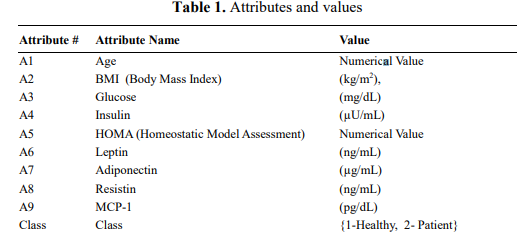
המטרה שלנו היא להעריך את יעילותן של שיטות אלו בחיזוי סרטן השד, תוך שימוש במספר קטן של תכונות רלוונטיות.

1. **סקר ספרות:[**[**1**](https://dergipark.org.tr/en/download/article-file/1390788)**]**
   1. **הקדמה:**

מאמר זה השתמש מאגר הנתונים Coimbra Breast Cancer, המכיל נתונים של 116 נשים (58 עם סרטן שד ו-58 בריאות) ו-10 בדיקות דם שונות וניסה לחזות נוכחות של סרטן השד באמצעות שני אלגוריתמים ספציפיים:

1. K-Nearest Neighbors (K-NN)
2. Naive Bayes.

המחקר מדגיש את הפוטנציאל של למידת מכונה בשיפור אבחון סרטן השד, תוך שימוש בבדיקות דם פשוטות וזמינות. פיתוח כלי אבחון יעיל יתרום רבות לזיהוי מוקדם של המחלה, שהוא הגורם המכריע בהצלת חיים . ע"פ המאמר רק בשנת 2019, ההערכה היא ש-41,760 נשים ו-500 גברים ימותו בגלל סרטן השד לכן גילוי מוקדם של סרטן הוא אחד הגורמים החשובים להישרדות מהמחלה. לכן כל שיטה או כל בדיקה שעוזרת לרופא לאבחן טוב יותר חיוניות לחיי אדם.



מחקר זה שואף לעלות על התוצאות של חקירות קודמות תוך שימוש באותו מערך נתונים.

* 1. **השיטות שמוצגות במאמר:**
     1. **k-NN**: מאמר זה עוסק באלגוריתם k-NN, אלגוריתם סיווג פשוט ויעיל המשמש לזיהוי תבניות, זיהוי דפוסים והכרת דמויות אופטיות.

איך עובד אלגוריתם KNN:

1. קביעת ערך k: תחילה, יש לבחור ערך k, שהוא מספר השכנים הקרובים ביותר שיילקחו בחשבון.
2. חישוב המרחקים: לאחר מכן, מחושב המרחק בין נקודת הנתונים החדשה לכל נקודת נתונים במערך האימון.
3. מציאת k השכנים הקרובים ביותר: נקודות הנתונים k עם המרחקים הקטנים ביותר מנקודת הנתונים החדשה נחשבות לשכנים הקרובים ביותר.
4. סיווג נקודת הנתונים החדשה: התווית של נקודת הנתונים החדשה נקבעת על פי התוויות של k השכנים הקרובים ביותר.

המאמר מתמקד בשני היבטים קריטיים של k-NN:

* + - 1. בחירת ה-k הנכון: מספר השכנים הקרובים המשמשים לסיווג. בחירה זו משפיעה משמעותית על ביצועי האלגוריתם.
      2. מדידת מרחק: חישוב המרחק בין נקודת הבדיקה לשכניה.

הקטע מסביר שלושה מדדי מרחק עיקריים:

* מרחק אוקלידי: הוא המרחק הסטנדרטי המשמש באלגוריתם kNN. הוא מחשב את סכום הריבועים של ההפרשים בין הערכים בכל התכונות.
* מרחק מנהטן: הוא מחשב את סכום ההפרשים המוחלטים בין הערכים בכל התכונות. הוא פחות רגיש להפרשים גדולים מאשר מרחק אוקלידי, אך הוא גם פחות מדויק עבור נתונים לא לינאריים.

* מרחק מינקובסקי: הכללה של המרחק האוקלידי והמנהטן, ניתנת להתאמה בהתאם לפרמטר p.

* + 1. **Naïve Bayes:** הוא אלגוריתם סיווג המשמש למיון נתונים. הוא מבוסס על משפט Bayes, ועל ההנחה ה"נאיבית" שאין תלות בין תכונות האובייקטים המסווגים כאשר כבר ידוע סיווגם. כלומר, לכל תכונה יש השפעה עצמאית על התיוג של הנתון.

איך עובד Naïve Bayes: האלגוריתם מחשב את ההסתברות של כל תכונה לכל תיוג. בנוסף הוא מחשב את ההסתברות של כל תיוג. עבור נתון חדש, האלגוריתם מחשב את ההסתברות של כל תיוג בהינתן התכונות של הנתון. התיוג בעל ההסתברות הגבוהה ביותר הוא התיוג שחיזוי האלגוריתם עבור הנתון.

סיכום פשוט של השיטה:

- (H): מה מנסים לחזות (למשל, נוכחות של סרטן השד).

- אסוף ראיות (X): נקודות נתונים רלוונטיות (למשל, תוצאות בדיקות דם).

המשוואה מתחלקת לארבעה מרכיבים:

* הסתברות אחורית (P(H|X)): הסתברות שההשערה נכונה בהתחשב בהוכחות. (הסיכוי שההשערה (H) נכונה בהתחשב בהוכחות (X))
* סבירות (P(X|H)): ההסתברות להתרחשות הראיות בהינתן ההשערה נכונה. (למשל, באיזו תדירות בדיקות דם ספציפיות מצביעות על סרטן השד?)
* הסתברות קודמת (P(H)): הסתברות שההשערה נכונה ללא קשר לראיות. (למשל, שכיחות כללית של סרטן השד)
* הסתברות שולית (P(X)): הסתברות להתרחשות הראיות ללא קשר להשערה. (לדוגמה, תדירות תוצאות בדיקות הדם הללו באופן כללי)

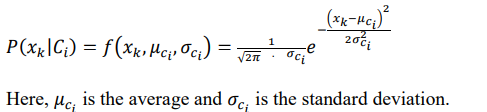
פירוש התוצאה:

* הסתברות אחורית גבוהה מצביעה על הראיות תומכות מאוד בהשערה.
* הסתברות אחורית נמוכה מרמזת שהראיות אינן תומכות חזק בהשערה.

נוסחה להסתברות אחורית:

Gaussian Naive Bayes classifier:

סיווג Gaussian Naïve Bayes (GNB) הוא סוג של Naïve Bayes המניח שהתפלגות התכונות עבור כל תיוג היא נורמלית (גאוסית). GNB הוא יעיל יותר מ-Naïve Bayes רגיל כאשר התכונות מתפלגות באופן נורמלי.

חישוב: משתמש בפונקציית צפיפות הסתברות ספציפית כדי להעריך את ההסתברות של נקודת נתונים () השייכת למחלקה ().

נוסחה:

* 1. **תוצאות המאמר:**

מאמר זה התעמק בפוטנציאל של שני אלגוריתמים KNN ו- Naive Bayes, בחיזוי סרטן השד באמצעות גישה ממוקדת המתמקדת בסמנים ביולוגיים ספציפיים. בעוד ששתי השיטות הניבו תוצאות טובות, החוזקות והחולשות שלהן הופיעו בבדיקה מעמיקה יותר.

מה שמייחד את המחקר זה הוא שבמקום להשתמש בכל מערך הנתונים כמו רוב המחקרים הקודמים, השתמשו בארבעת הסמנים הביולוגיים הראשונים: resistin, גלוקוז, גיל ו-BMI כסמנים ביולוגיים כדי לחזות את נוכחות סרטן השד. וכך השיגו תוצאות טובות ועדינות יותר ממחקרים קודמים.

במחקר השתמשו ב-80% מהנתונים לאימון ו-20% לבדיקות

|  |  |  |
| --- | --- | --- |
| Negative 2 | Positive 1 |  |
| 1 | 11 | Positive 1 |
| 10 | 2 | Negative 2 |

תוצאות המחקר:

|  |  |  |
| --- | --- | --- |
| Metric | KNN | Naive Bayes |
| Accuracy | 87.50% | 79.00% |
| Sensitivity | 84.62% | 75.00% |
| Specificity | 91.00% | 83.00% |
| AUC | 87.76 | 79 |

:KNN algorithm confusion matrix

:Naive Bayes algorithm confusion matrix

|  |  |  |
| --- | --- | --- |
| Negative 2 | Positive 1 |  |
| 2 | 10 | Positive 1 |
| 9 | 3 | Negative 2 |

כפי שניתן לראות, KNN השיגה ביצועים טובים יותר מ-Naive Bayes בכל המדדים. זה מצביע על כך ש-KNN היא שיטה מדויקת ואמינה יותר לאיתור סרטן השד באמצעות ארבעת הסמנים הביולוגיים הראשונים (רסיסטין, גלוקוז, גיל ו-BMI).

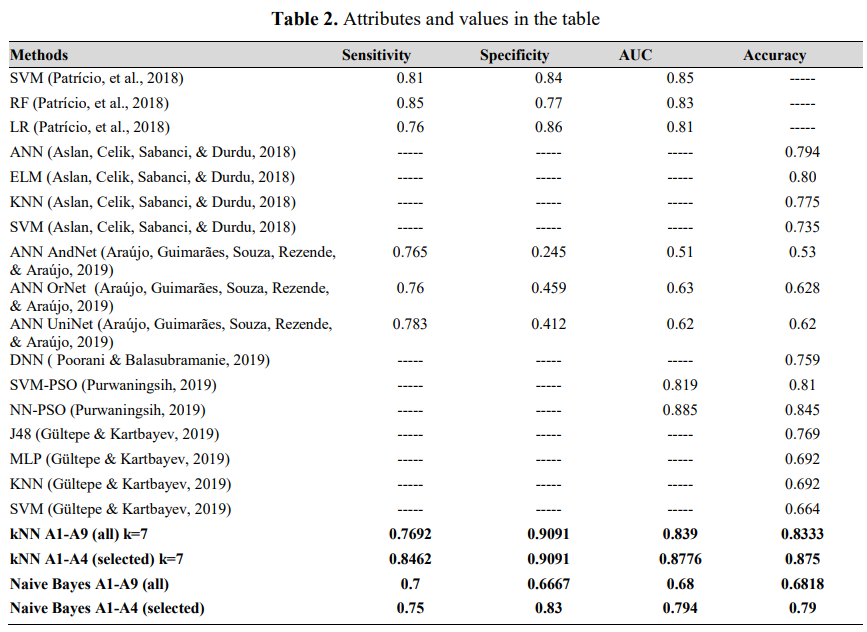
ממצאים אלו מספקים ראיות לכך שמודלים המשלבים רזיסטין, גלוקוז, גיל ו-BMI עשויים להיות כלי רב עוצמה לגילוי סרטן השד. להלן הסבר על המדדים:

KNN:

* KNN שלטה טוב מאוד בדיוק, וזיהתה נכון סרטן שד ב-87.5% מהמקרים. משמעות הדבר מבחינה בין אנשים בריאים לבין אלה עם המחלה.
* יכולת הזיהוי הביאה תוצאות טובות הן ברגישות והן בספציפיות. KNN זיהה במדויק מקרים סרטניים ב- 84.62% מהזמן, תוך מזעור תוצאות כוזבות (אנשים בריאים המסווגים בטעות כסרטניים) עם ספציפיות של 91%.
* ערך k הטוב ביותר: 7

**Naive Bayes**

* בעוד שהדיוק לא היה בדיוק ברמה של KNN, Naive Bayes השיג 79%. זה מצביע על כך שהוא תפקד היטב בסיווג הכולל.
* בספציפיות הגיע ל-83%. משמעות הדבר היא שהוא ממזער תוצאות חיוביות כוזבות, ומספק ביטחון לאנשים שעלולים להיות מודאגים מאבחון שגוי.

טבלת השואה בין תוצאות המאמר לתוצאות מאמרים קודמים לו:

התוצאות מראות כי השיטות השונות הציגו ביצועים שונים, אך באופן כללי, הן הצליחו להשיג ביצועים טוביים. KNN ו-Native Bayes היו השיטות הטובות ביותר.

* 1. **מסקנות:**  
     שימוש בסמנים ביולוגיים ספציפיים: התמקדות בגיל, בגלוקוז, ברסיסטין וב-BMI במקום במערך הנתונים כולו מראה הבטחה לפישוט ולשיפור הזיהוי.

ביצועים חזקים של KNN: השגת 90% ספציפיות, 84% רגישות ודיוק של 87.5% בולטת, במיוחד בהשוואה למחקרים אחרים.

קלות יישום: kNN ו-Naive Bayes הם אלגוריתמים פשוטים יחסית, מה שהופך אותם לפשוטים עבור יישומים מעשיים.

פוטנציאל לגילוי מוקדם: יישום שיטות אלה כדי להבחין בין מקרים שפירים לממאירים עשוי להיות בעל ערך עבור רופאים בגילוי סרטן השד מוקדם יותר.

1. **שיטות בהם השתמשנו בעבודה:**
   1. **KNN** – בעבודה זו בחרנו לנסות לשחזר את תוצאות אלגוריתם KNN (שהוא אלגוריתם סיווג פשוט ויעיל) כפי שצוין במאמר. בנוסף בחרנו להשתמש באלגוריתם זה כדי לנסות לחזות שיפור בנתונים לאחר הורדת הממד ע"י PCA. ניתן לקרוא על האופן בו האלגוריתם KNN עובד בהרחבה בסקר ספרות (סעיף 3.2.1).
   2. **PCA (Principal Component Analysis)** - היא שיטה המשמשת להפחתת ממדים וניתוח נתונים גדולים. השיטה מאפשרת לפשט את ערכי המשתנים בנתונים למערכת קטנה יותר, בלי לאבד מידע חשוב.

השיטה עובדת בחמישה שלבים עיקריים:

1. סטנדרטיזציה: טווח הערכים של כל משתנה מותאם כך שלכולם יהיה ממוצע של 0 ושונות של 1.
2. מטריצת הקובריאנס: מחושבת כדי לזהות קשרים (קורלציות) בין המשתנים.
3. וקטורים עצמיים וערכי עצמיים של מטריצת הקובריאנס: מחושבים כדי למצוא את הרכיבים הראשיים.
4. וקטור תכונות: מחליט באילו רכיבים ראשיים להשתמש.
5. טרנספורמציה: הנתונים מותאמים לצירים החדשים שנוצרו על ידי הרכיבים הראשיים.

הרכיבים הראשיים הם משתנים חדשים שנוצרים כשילוב לינארי של המשתנים המקוריים. המשתנים החדשים הללו מורכבים בצורה שבה הם לא קורולטים ומכילים את רוב המידע מהמשתנים המקוריים. תהליך זה מאפשר להפחית את ממדי הנתונים מבלי לאבד הרבה מהמידע המקורי.

PCA משמש בעיקר ל:

* פישוט המידע**:** מאפשר לנתח נתונים בקלות רבה יותר מכיוון שיש פחות משתנים.
* חיסכון במשאבים**:** אלגוריתמי למידת מכונה יכולים לעבד נתונים קטנים יותר מהר וביעילות.
* שיפור ביצועים**:** לעיתים, הפחתת מימדיות יכולה לשפר את הביצועים של אלגוריתמי למידת מכונה.

בסמינר השתמשנו בשיטה זו כדי לצמצם את מימדי הנתונים, וכך לנסות לשפר את החיזוי

1. **תוצאות:** 
   1. **שיחזור אלגוריתם KNN מהמאמר-**

השתמשנו באותו מאגר נתונים שצוין במאמר, ממנו לקחנו רק את 4 הפיצ'רים הראשונים שצוינו במאמר וביצענו על הנתונים את אלגוריתם KNN עם 7 שכנים (בדיוק כמו שנכתב במאמר) להלן התוצאות:

:KNN algorithm confusion matrix

|  |  |
| --- | --- |
| תוצאות שלנו | תוצאות המאמר |
| |  |  |  | | --- | --- | --- | |  | Positive | Negative | | Positive | 11 | 1 | | Negative | 3 | 9 | | |  |  |  | | --- | --- | --- | | Negative 2 | Positive 1 |  | | 1 | 11 | Positive 1 | | 10 | 2 | Negative 2 | |

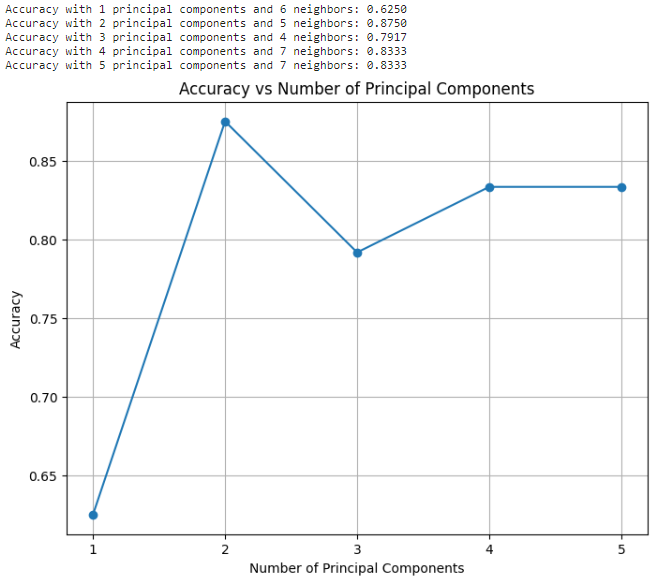
התוצאות שלנו לעומת תוצאות המאמר:

|  |  |  |
| --- | --- | --- |
|  | תוצאות המאמר | תוצאות שלנו |
| דיוק (Accuracy) | 87.50% | 83% |
| רגישות (Recall/Sensitivity) | 84.62% | 75% |
| AUC | 87.76% | 83% |
| ספציפיות (Specificity) | 91.00% | 91.60% |

פירוט התוצאות:

* על פי תוצאות המאמר, הדיוק (Accuracy) הוא 87.50%, בעוד שבתוצאות שיצאו לנו, הדיוק הוא 83.00%. כלומר, הדיוק שלנו נמוך ב-4.5% מזה של המאמר. ייתכן שהסיבה לכך היא הבדלים בנתונים או בפרמטרים אחרים של האלגוריתם.
* בנוגע לרגישות (Recall/Sensitivity), התוצאות מציינות כי במאמר יש רגישות של 84.62%, ובתוצאות שיצאו לנו רק 75.00%. משמעות הדבר, שהאלגוריתם שלנו מפספס יותר מקרים חיוביים (חולות סרטן) מאשר האלגוריתם במאמר.
* בנושא הספציפיות (Specificity), התוצאות מציינות כי במאמר יש ספציפיות של 91.00%, ובתוצאות שיצאו לנו קיבלנו ספציפיות של 91.60%. משמעות הדבר, שהאלגוריתם שלנו טוב יותר בזיהוי חולים בריאים (שליליים).
* - AUC תוצאות המאמר מציינות 87.76%, בעוד שהתוצאות שלנו הן 83.00%. כלומר, ערך ה-AUC שלנו נמוך ב-4.76% מזה של המאמר. ערך זה מודד את היכולת של האלגוריתם להבחין בין חולים בריאים לחולים, ולכן ירידה זו מעידה על ירידה ביכולת ההבחנה.
* כאשר אנו מביטים ב confusion matrix, רואים כי יש הבדלים בין התוצאות שלנו ובין התוצאות של המאמר. ניתן לראות כי האלגוריתם שלנו זיהה נכונה 11 חולות סרטן (חיובי חיובי) ו-9 נשים בריאות (שלילי שלילי). עם זאת, הוא פספס 3 חולות סרטן (שלילי חיובי) וסיווג בטעות אישה בריאה אחת כחולה (חיובי שלילי)

התוצאות שלנו מצביעות על ביצוע טוב של אלגוריתם KNN עם 7 שכנים על סט הנתונים, אך הן מעט פחות טובות מתוצאות המאמר. למרות זאת, שתי התוצאות הן תוצאות טובות המצביעות על כך שאלגוריתם KNN עם 7 שכנים ו4 פיצ'רים יכול להוות כלי יעיל לחיזוי סרטן השד

* 1. **תוצאות אלגוריתם KNN לאחר ביצוע PCA עם 1-5 רכיבי PC -**

שלב 1 - בחירת מספר רכיבי PC:

הגרף מראה כיצד הדיוק של מסווג KNN משתנה ככל שמספר רכיבי PC גדל. ציפינו שהדיוק יגדל ככל שנוסיף יותר רכיבי PC מכיוון שהם לוכדים יותר שונות בנתונים, ומספקים מידע נוסף לסיווג. עם זאת, ייתכן שיש נקודה שבה הוספת רכיבים נוספים אינה משפרת משמעותית את הדיוק, או אפילו מפחיתה אותו עקב רעש או התאמה יתר.

כמו במקרה שלנו, ניתן לראות שהדיוק מגיע לשיא של 87.5% עם 2 רכיבי PC ו-5 שכנים ולאחר מכן יורד. לכן נתמקד בתוצאות אלו ונשווה אותן לתוצאות המאמר. להלן התוצאות:

:KNN algorithm confusion matrix

|  |  |
| --- | --- |
| תוצאות שלנו | תוצאות המאמר |
| |  |  |  | | --- | --- | --- | |  | Positive | Negative | | Positive | 11 | 1 | | Negative | 2 | 10 | | |  |  |  | | --- | --- | --- | | Negative 2 | Positive 1 |  | | 1 | 11 | Positive 1 | | 10 | 2 | Negative 2 | |

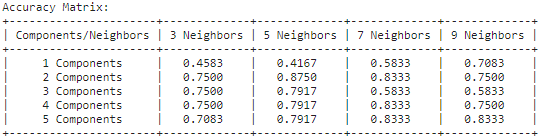
|  |  |  |
| --- | --- | --- |
| מדד | תוצאות המאמר | תוצאות שלנו לאחר PCA |
| דיוק (Accuracy) | 87.50% | 87.5% |
| רגישות (Recall/Sensitivity) | 84.62% | 83.3% |
| AUC | 87.76% | 87.5% |
| ספציפיות (Specificity) | 91.00% | 91.60% |

ניתוח התוצאות:

* דיוק- הדיוק שלנו זהה לזה של המאמר (87.5%). זה אומר שאנחנו מצליחות לזהות את אותה כמות של מקרים חיוביים ושליליים.
* רגישות- הרגישות נמוכה מעט מזו של המאמר (83.3% לעומת 84.62%). משמעות הדבר, שאנו מפספסים מעט יותר מקרים חיוביים (חולות סרטן) מאשר האלגוריתם במאמר.
* AUC- ה-AUC שלנו דומה לזה של המאמר (87.5% לעומת 87.76%). ערך זה מודד את היכולת של האלגוריתם להבחין בין חולים בריאים לחולים, ולכן ירידה קלה זו אינה משמעותית.
* ספציפיות - הספציפיות שלנו מעט גבוהה מזו של המאמר (91.6% לעומת 91.00%). משמעות הדבר, שאנו טובים יותר בזיהוי חולים בריאים (שליליים).
* confusion matrix - ניתן לראות כי האלגוריתם שלנו זיהה נכונה 11 חולות סרטן (חיובי חיובי) ו-10 נשים בריאות (שלילי שלילי). עם זאת, הוא פספס 2 חולות סרטן (שלילי חיובי) וסיווג בטעות 2 נשים בריאות כחולות (חיובי שלילי). בדיוק כמו במאמר.

התוצאות לאחר PCA מצביעות על ביצוע טוב של אלגוריתם KNN עם 2 רכיבי PC ו-5 שכנים. הן דומות מאוד ואפילו טובות יותר מתוצאות המאמר, מה שמצביע על כך ש-PCA יכול לשמש כטכניקה יעילה להפחתת מימדיות בנתונים מבלי לפגוע משמעותית בביצועים.

שלב 2 - בחירת מספר רכיבי KNN:



הטבלה מציגה את ציוני הדיוק שהושגו על ידי k-NN עבור 5 שילובים של מספר רכיבים עיקריים ומספר השכנים. הסבר קצר על התוצאות:

* כאשר נעשה שימוש רק ברכיב אחד, הדיוק נמוך יחסית עבור כל הערכים של k (מספר שכנים). זה צפוי מכיוון שרכיב בודד עשוי שלא ללכוד מספיק מידע כדי לסווג במדויק את הנתונים.
* ככל שמספר הרכיבים גדל (מ-2 ל-5), הדיוק בדרך כלל משתפר, במיוחד עבור ערכים נמוכים יותר של k (למשל, 3 או 5 שכנים).
* עם יותר רכיבים, ל k-NN יש גישה למידע נוסף על מבנה הנתונים, מה שמוביל לביצועי סיווג טובים יותר.
* עם זאת, שימוש ביותר מדי רכיבים עלול גם להכניס רעש או מידע לא רלוונטי, מה שיכול להפחית את הביצועים.
* בחירת מספר השכנים (k) משפיעה גם על דיוק הסיווג. במקרים מסוימים, הגדלת מספר השכנים מובילה לביצועים טובים יותר, בעוד שבאחרים, מספר קטן יותר של שכנים מתפקד טוב יותר.

תוצאות הטבלה מראות את החשיבות של בחירת היפר פרמטרים (כגון בחירת המספר המתאים של שכנים) בהשגת ביצועי סיווג טובים ובחירת מספר רכיבים. ניתן להסיק מכך, שחשוב למצוא איזון בין כמות המידע המאוחסן (מספר הרכיבים) למורכבות המודל (מספר שכנים) כדי למנוע התאמת יתר או התאמה של הנתונים.

1. **מסקנות**:

* תוצאות האלגוריתם KNN עם 7 שכנים ו-4 תכונות מצביעות על יעילותו בחיזוי סרטן השד, עם דיוק של 83%. אמנם התוצאות קצת נמוכות יותר מאלו של המאמר המקורי (87.5%), אך הן עדיין גבוהות מאוד ומוכיחות את היעילות של האלגוריתם.
* בחירת מספר רכיבי PC ומספר השכנים הם פרמטרים חשובים בתהליך הסיווג באלגוריתם KNN. התוצאות מראות כי יש עליה וירידה בדיוק בהתאם לשינוי במספר רכיבים ומספר השכנים.
* תוצאות האלגוריתם KNN לאחר ביצוע PCA מראות כי השימוש בטכניקת ה-Dimensionality Reduction על ידי PCA יכול להוביל לשיפור בביצועים. במיוחד כאשר משתמשים במספר רכיבי PC קטן.
* ההבדלים הקטנים בין תוצאות המאמר לבין התוצאות שלנו מראים על יכולת המודל לשמור על יעילות ודיוק גבוהים, גם לאחר ביצוע PCA. זה מעיד על חשיבות הבחירה המדויקת של פרמטרים ועל היכולת של המודל לזהות את הפרטים החשובים ביותר בנתונים.
* התוצאות שהושגו מדגישות את הפוטנציאל של אלגוריתמי למידת מכונה, ובפרט KNN, לחיזוי סרטן השד. בחירת פרמטרים אופטימלית, שימוש בטכניקות "Dimensionality Reduction" כגון PCA, ובחירת תכונות רלוונטיות, הן גורמים מרכזיים להשגת דיוק גבוה וביצועים מיטביים.

1. **מגבלות**:

* הסט נתונים שהשתמשנו בו מורכב מ-116 דוגמאות בלבד, ומתוכן כמעט חצי הדוגמאות מייצגות אנשים חולים. זה אינו מייצג מדויק של האוכלוסייה בשל העובדה שרוב האנשים באוכלוסייה אינם חולים, ולכן כמות זו של דוגמאות שלא מייצגות את הרוב של האוכלוסייה שהיא מאוד נמוכה. בשל כך, יתכן כי התוצאות שנקבל יהיו מוטעות או לא מייצגות באופן מדויק את התופעות באוכלוסייה הכללית.
* המאמר שקראנו מתייחס לשימוש בשני אלגוריתמים בלמידת מכונה, אך לא מתאר באופן מפורט אילו היפר-פרמטרים נעשה בהם שימוש בכל אלגוריתם. בנוסף, המאמר לא מספק מידע על העיבוד המקדים שנעשה על הנתונים לפני השימוש באלגוריתמים, ולא מפרט איך נבחרו הפיצ'רים בהם השתמש, חוסר מידע הזה עשוי להשפיע על דיוק התוצאות שקיבלנו בהשוואה לתוצאות שצוינו במאמר.

1. **הצעות להמשך עבודה:**

* ניסוי שחזור אלגוריתם Naive Bayes המצוין במאמר והשוואה לתוצאות לאחר ביצוע PCA
* ניסוי אלגוריתמים אחרים של למידת מכונה על רכיבי PCA והשוואת הביצועים בניהם.
* ניסוי אותם אלגוריתמים בהם השתמשנו על dataset יותר גדול

7.1 הסבר קצר למה שחזור אלגוריתם Naive Bayes המצוין במאמר והשוואה לתוצאות לאחר ביצוע PCA יכול להביא לתוצאות טובות:

1. הפחתת ממדיות: PCA מפחית את מספר התכונות במערך הנתונים, ובכך מקל על עיבוד וניתוח הנתונים. יתרון זה חשוב במיוחד עבור Naive Bayes, כיוון שסיווג נאיבי יעיל יותר עם מספר קטן יותר של תכונות.
2. ההנחה ה"נאיבית" (אין תלות בין תכונות): Naive Bayes מסתמך על ההנחה של עצמאות תכונה, כאשר כל תכונה תורמת באופן עצמאי להסתברות של מחלקה מסוימת. PCA יכול לעזור להבטיח הנחה זו על ידי יצירת רכיבים עיקריים חסרי קורלציה.
3. הפחתת רעש: PCA מסנן רעש ותכונות לא רלוונטיות על ידי לכידת השונות המשמעותית ביותר בנתונים, ומאפשר ל-Naive Bayes להתמקד בהיבטים החשובים של הנתונים.
4. **מקורות מידע:**

המאמר ששחזרנו:

[[1] ABDULKAREEM, A. H., & KASAPBAŞI, M. C. (2020). ENHANCING DETECTION METHOD OF BREAST CANCER USING COIMBRA DATASET. İstanbul Ticaret Üniversitesi Teknoloji Ve Uygulamalı Bilimler Dergisi, 3(1), 51-59.](https://dergipark.org.tr/en/download/article-file/1390788)

מקורות מידע בהם נעזרנו בקוד/ במציאת מידע על האלגוריתמים בהם השתמשנו:

2. <https://saturncloud.io/blog/what-is-sklearn-pca-explained-variance-and-explained-variance-ratio-difference/>

3. <https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html>

4. <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

5. <https://www.kaggle.com/code/avikumart/pca-principal-component-analysis-from-scratch>

6. <https://rstudio-pubs-static.s3.amazonaws.com/471974_9c6250108efc497789ee5840f24b0db4.html#data-analysis>

7. <https://erdogant.github.io/pca/pages/html/Algorithm.html#normalizing-out-pcs>

8. <https://www.kaggle.com/code/abdelrasoul/knn-pca-ml-83>

9. <https://www.kaggle.com/code/tanshihjen/module-eda-with-fasteda>

10. <https://www.kaggle.com/code/saswattulo/coimbra-breastcancer-prediction-with-91-6-accurac>

11. <https://www.kaggle.com/code/benyaminghahremani/coimbra-breast-cancer-classification-feature-eng>

12. <https://medium.com/analytics-vidhya/dimensionality-reduction-principal-component-analysis-d1402b58feb1>

כאן ניתן לראות את המחברת שלנו ב Kaggle עם פירוט נוסף והסברים על כל שלב בעבודה:

<https://www.kaggle.com/code/liatsagi/seminar1>

קישורים למצגות שהצגנו במהלך הסמסטר:

מצגת 1: <https://www.canva.com/design/DAF65So4Xv4/M_LtpLrC17u-JK9oLYQoww/edit?utm_content=DAF65So4Xv4&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>

מצגת 2: <https://www.canva.com/design/DAF-LvoIJJI/c0EVN2IER8bAV5Wei0W3rw/edit?utm_content=DAF-LvoIJJI&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>

קישור ל Github: