הנדון: <u>ניתוח של מערך נתונים של יעילות שינה</u>

הקדמה

מסמך זה מציג ניתוח מקיף של מערך נתונים של יעילות השינה, תוך התמקדות בטכניקות עיבוד מקדים ובמודלים של סיווג, המיושמים כדי לחזות את עמודת המגדר.

מערך הנתונים מכיל מידע על דפוסי שינה של אנשים, כולל משך השינה, יעילות השינה, שעת השינה, זמן ההתעוררות ושלבי שינה שונים. מטרת מחקר זה הייתה לחקור את מערך הנתונים, לעבד מראש את הנתונים, להפעיל מודלים של סיווג ולהשוות בין התוצאות כדי לזהות את המודל היעיל ביותר ולייצר מסווג שיוכל לחזות את עמודת המגדר בנתונים חדשים המותאמים לנתונים הקיימים לנו במערך הנתונים באופן המדויק ביותר. מסמך זה מספק סקירה מפורטת של מערך הנתונים, מתאר את שלבי העיבוד המקדים ומודלי הסיווג שהופעלו, ומציג השוואה מקיפה של התוצאות.

מבוא

מערך הנתונים כולל מספר עמודות משמעותיות התורמות להבנת דפוסי יעילות השינה. אלו כוללים:

1. משך השינה- משך הזמן הכולל שכל נבדק ישן בשעות. הוא מספק תובנות לגבי משך השינה, שהוא חיוני להערכת איכות השינה.
2. שעת שינה וזמן השכמה- מאפיינים אלה מציינים את לוח השינה הממוצע של האדם, ונותנים מידע על הרגלי שינה והשגרה.
3. שלבי שינה- מערך הנתונים כולל מידע על הזמן המושקע בשלבי שינה שונים, כגון שנת REM , שינה עמוקה ושינה קלה. שלבים אלו מספקים תובנות לגבי איכות ועומק השינה שחווים בני אדם.

   <u>שינה מורכבת מארבעה שלבים-</u>

   א. שנת חלום REM
   ב. שינה עמוקה
   ג. שינה קלה
   ד. שינה שטחית
4. התעוררויות- מספר הפעמים שכל נבדק מתעורר במהלך הלילה. תכונה זו עוזרת להבין הפרעות והפרעות שינה שעשויות להשפיע על יעילות השינה.
5. צריכת קפאין ואלכוהול- מערך הנתונים מתעד את צריכת הנבדק של קפאין ואלכוהול ב-24 השעות שלפני השינה. ידוע כי גורמים אלו משפיעים על איכות ויעילות השינה.
6. סטטוס עישון- מידע האם הנבדק מעשן או לא, המאפשר לבחון קשרים פוטנציאליים בין עישון ויעילות שינה.
7. תדירות פעילות גופנית- תכונה זו מתארת את תדירות הפעילות הגופנית המדווחת על ידי כל נבדק. פעילות גופנית עשויה להיקשר לאיכות שינה טובה יותר, ועמודה זו מאפשרת לחקור את השפעתה על יעילות השינה.

8. יעילות השינה- תכונת יעילות השינה היא מדד לשיעור זמן השהות במיטה שבפועל בשינה, עמודה זו באה לידי ביטוי ביחס ישיר לעמודות השונות בסט הנתונים ובאמצעותן (כלל העמודות למעט עמודת המגדר) ננסה לייצר מודל החוזה באופן הטוב ביותר את עמודת המגדר.
9. מגדר- עמודה המתארת אם מדובר בגבר או באישה- **התכונה שאותה נרצה לחזות.**

ניתוח נתונים חקרני מקיף- EDA

מטרת ה-EDA הייתה להשיג תובנות על מערך הנתונים, לזהות דפוסים ולהבין את הקשרים בין משתנים שונים. חלק זה מתאר את הצעדים שבוצעו במהלך ה-EDA ומספק ניתוח של הממצאים.

<u>שלב 1 : חקר נתונים ותצפיות ראשוניות</u>

עם טעינת מערך הנתונים, העמודות נבדקו כדי להבין את התכונות הקיימות. מערך הנתונים כלל מידע רב התורם להבנת דפוסי יעילות השינה- פירוט העמודות מצורף לעיל. עמודות אלו מילאו תפקיד חיוני בהבנת דפוסי יעילות השינה.

<u>שלב 2 : ערכים חסרים</u>

ערכים חסרים נבדקו כדי לקבוע את מידת החסר במערך הנתונים. המשתנה missing_values שימש כדי לחשב את סך הערכים החסרים, ו missing_data-נוצר כדי להציג את אחוז הערכים החסרים בכל עמודה. ניתוח זה סיפק תובנות לגבי שלמות מערך הנתונים והדגיש פערים פוטנציאליים שיש לטפל בהם במהלך העיבוד המקדים. גילינו שיש מעט ערכים חסרים (הדאטא מכילה 452 שורות על 16 עמודות), סה״כ 4 עמודות עם ערכים חסרים :

20 בעמודה המייצגת התעוררויות, 25 בעמודה המייצגת צריכת קפאין, 14 בעמודה המייצגת צריכת אלכוהול, 6 בעמודה המייצגת תדירות פעילות גופנית- החלטנו לטפל בערכים אלו בתהליך של ניקוי הנתונים ע״י השלמת ערכים ממוצעים שכן גילינו כי מדובר על שורות בעלות מרבית הערכים האחרים מלאים.

<u>שלב 3 : חלוקת הנתונים וחריגים</u>

זוהו עמודות מספריות, והתפלגויותיהן הוצגו באמצעות תרשים קופסאות. זה איפשר הבנה טובה יותר של חלוקת הנתונים, הנטייה המרכזית והנוכחות של חריגים. תצפית ראויה לציון הייתה נוכחותם של ערכים חריגים בעמודות כגון ״אחוז שינה קלה״ ו״צריכת קפאין״.

<u>שלב 4 : טיפול בחריגים</u>

כדי להבטיח את שלמות הנתונים, הוחלפו חריגים בערכי אפס. גישה זו ננקטה כדי להימנע מהטיית תהליך הניתוח והמודלים עקב ערכים קיצוניים שכן ייתכן והוזנו בטעות. באופן ספציפי, ערכי חריגים הוחלפו ב- None בעמודות ״אחוז שינה קלה״ ו״צריכת קפאין״. שלב זה הקל על ניתוח מדויק יותר של מערך הנתונים.

<u>שלב 5 : התפלגות מגדרית</u>

התפלגות המגדר במערך הנתונים נבדקה באמצעות גרפי עמודות. היא סיפקה סקירה כללית של ייצוג הנבדקים גברים ונשים, ואיפשרה הבנה טובה יותר של הרכב המגדר של מערך הנתונים- גילינו כי הגרף שלנו אינו סובל מהתאמת יתר (Overfitting), שכן התפלגות הגברים ונשים שוויונית- 228 גברים, 224 נשים.

<u>שלב 6 : קשר בין משתנים</u>

דיאגרמות פיזור הופעלו כדי לחקור את הקשרים בין משתנים שונים. יש לציין כי נוצרו דיאגרמות פיזור כדי לבחון את הקשר בין ״משך השינה״ ו״יעילות השינה״, וכן בין ״משך השינה״ ל״מגדר״. הדמיות אלו סיפקו

תובנות לגבי מתאמים ודפוסים אפשריים בין משך השינה, יעילות השינה ומגדר, בניתוח הגרפים ניתן לומר כי אינו קיים מתאם בין נתונים אלו.

שלב 7 : ניתוח ספציפי למגדר

ניתוח נוסף נערך על ידי הפרדת מערך הנתונים על סמך מגדר. עמודות הגרף שימשו להשוואת התפלגות משך השינה ויעילות השינה בין גברים ונשים. ניתוח ספציפי למגדר איפשר הבנה מעמיקה יותר של ההבדלים בדפוסי השינה וביעילות המבוססת על מגדר.

שלב 8 : ניתוח רב משתנים

כדי לחקור קשרים מורכבים בין משתנים, יושמו טכניקות ניתוח רב משתנים. לדוגמה, נוצרה מטריצת פיזור כדי להמחיש את הקשר בין יעילות שינה, מצב עישון ואחוז שנת REM. דיאגרמת קו נוספת נוצרה כדי לבחון את הקשר בין גיל, יעילות שינה ומגדר. הדמיות אלו סיפקו תובנות לגבי משחק הגומלין של משתנים מרובים והשפעתם על יעילות השינה. לדוגמה ניתן לראות בדיאגרמת הקו שלאורך חיי אדם (הן גבר והן אישה) ישנה יעילות שינה זהה באופן יחסי הן אצל גברים והן אצל נשים. כמו כן, ע״פ נתונים אלו נראה שאצל גברים בשנות ה-20 המוקדמות יעילות השינה נמוכה מנשים ואילו בשנות ה-30 המוקדמות יעילות השינה של נשים נמוכה יחסית לשל גברים (גם בשנות ה-50 המאוחרות אצל גברים יעילות השינה נמוכה יותר).

שלב 9 : השפעת עישון ואלכוהול

דיאגרמת קופסאות נוצלו כדי לנתח את ההשפעה של מצב העישון וצריכת אלכוהול על יעילות השינה. ניתוח זה נועד לזהות קשרים פוטנציאליים בין גורמי אורח חיים אלו ואיכות השינה. בדיאגרמת השינה המתארת קשר בין צריכת אלכוהול ליעילות שעות שינה, אכן ניתן לזהות שאי צריכת אלכוהול מייצרת את יעילות השינה הגבוהה ביותר, אמנם ניתן לומר ע״פ נתונים אלו שהקשר ביניהם אינו לינארי דבר המקשה עלינו להגיע למסקנה מובהקת בנושא הקשר בין המשתנים. ע״פ התרשים המקשר בין יעילות השינה לסטאטוס מעשן ולא מעשן, ניתן לומר שיעילות השינה אצל לא מעשנים אכן טובה יותר.

סיכום EDA-

תהליך ה-EDA שנערך על מערך הנתונים של יעילות השינה חשף תובנות חשובות לגבי דפוסי השינה, יעילות השינה והגורמים המשפיעים עליהם. באמצעות הדמיות וניתוח סטטיסטי, זוהו דפוסים ונחקרו קשרים בין משתנים. החלפת חריגים בערכים אפסיים הבטיחה את שלמות הנתונים במהלך הניתוח והמודלים הבאים. הניתוח הספציפי למגדר שופך אור על הבדלי דפוסי השינה בפן המגדרי.

מסקנות לגביי טכניקות עיבוד מקדים ומודל חלוקה לאשכולות

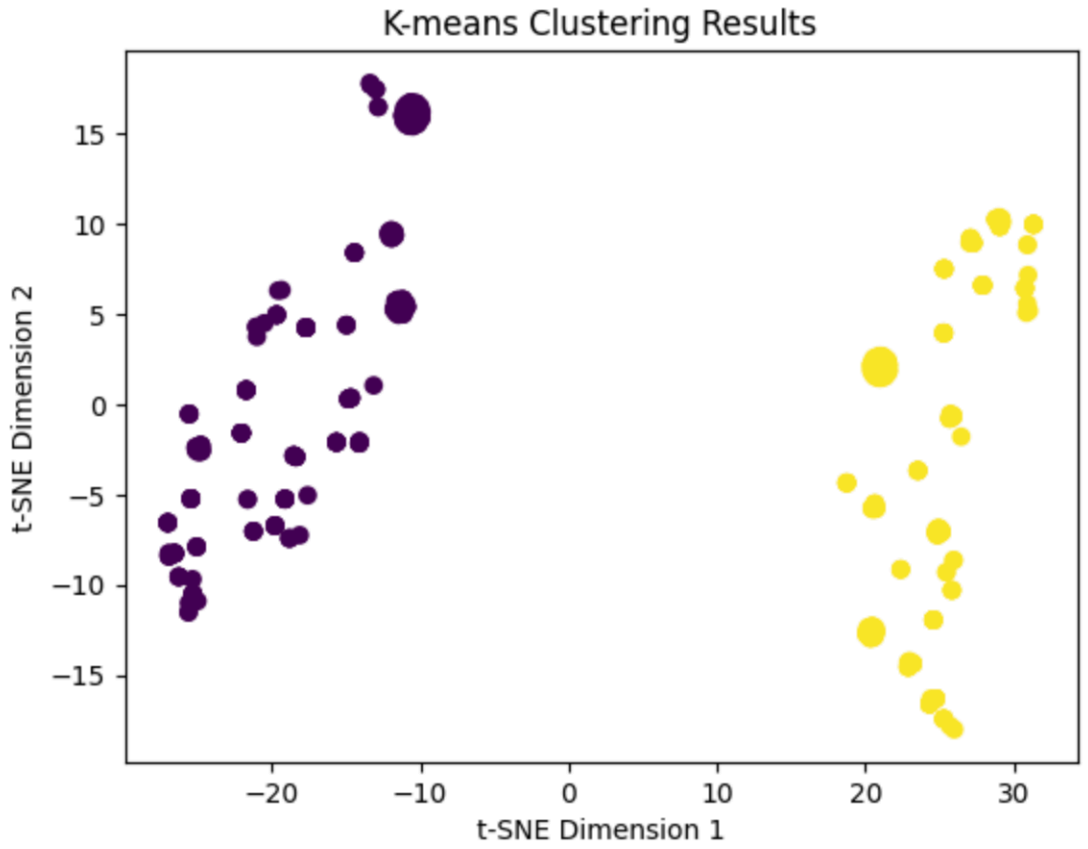
בפרויקט, עשינו שימוש בטכניקות שונות של עיבוד מקדים של נתונים ומודלים של למידת מכונה כדי לנתח את מערך הנתונים. **המטרה העיקרית** הייתה לקבוע את **טכניקת העיבוד המקדים והשילוב המודל הטוב ביותר** שמניב את מדדי הביצועים הגבוהים ביותר.

טכניקות עיבוד מקדים של נתונים כללו קנה מידה מינימלי-מקסימלי, קנה מידה של ציון z , קנה מידה עשרוני ונרמול לפי טכניקות וקטור יחידה. טכניקות אלו יושמו כדי לנרמל את התכונות המספריות של מערך הנתונים לפני אימון המודלים.

כדי להעריך את הביצועים, נעשה שימוש במספר מדדים. מדדים אלו סיפקו לנו תובנות לגבי הדיוק הכולל, דיוק החיזוי החיובי, שיעור החיובי האמיתי ואיזון בין דיוק לזכירה.

כמו כן, בדקנו שילובים שונים של טכניקות ומודלים של עיבוד מקדים, והתוצאות נותחו בקפידה כדי לקבל החלטות מושכלות לגבי השלבים הבאים.

כחלק מתהליך ניסיון שיפור המודל, החלטנו להוסיף סיווג לפי אשכולות של מגדר באמצעות מודל האשכולות K-means.



הגרף מציג ייצוג דו מימדי של הנתונים באמצעות אלגוריתם t-SNE (t-Distributed Stochastic Neighbor Embedding), שהיא טכניקת הפחתת מימדיות. בגרף זה, הצירים אינם מייצגים משתנים ספציפיים ממערך הנתונים. במקום זאת, הם צירים שרירותיים שנוצרו על ידי אלגוריתם t-SNE כדי להמחיש את הנתונים במרחב ממדי מצומצם. אלגוריתם t-SNE הופך נתונים בעלי מימד גבוה למרחב בעל מימד נמוך תוך שמירה

על המבנה המקומי והיחסים בין נקודות הנתונים. הוא משמש לזיהוי אשכולות או דפוסים שעשויים להתקיים בתוך הנתונים.

בניתוח זה, הגרף המתקבל מספק ייצוג חזותי של נקודות הנתונים המקובצות במרחב דו מימדי, ומסייע בזיהוי ובפירוש של דפוסים או קבוצות בתוך מערך הנתונים.

חשוב להתייחס לכך שהתוספת של מודל האשכולות K-means הציגה מימד חדש לניתוח שלנו. **לאחר החלת K-means clustering והוספת תוויות האשכולות כעמודה חדשה למערך הנתונים, ראינו חלוקה ברורה ומושלמת לשני אשכולות מופרדים לחלוטין** (כפי הנראה בגרף לעיל). יתר על כן, בעת ניתוח ההתפלגות המגדרית בתוך כל אשכול, מצאנו שהאשכולות היו מופרדים היטב על סמך מגדר. זה מדגיש את החשיבות של מודל האשכולות K-means בקיבוץ נקודות נתונים דומות יחדיו וחשיפת דפוסים בסיסיים בנתונים. **ובהתאמה ישירה לכך, אכן המודל הנבחר הסופי עם טכניקת הנרמול המתאימה לכל עמודה נתן את המדדים הטובים ביותר לסט הנתונים שלנו עם תוספת העמודה שיצר מודל האשכולות** (נתייחס לכך בסעיף המודל הנבחר).

נתייחס לתוצאות המדדים השונים וננתח את כל האפשרויות והסיבות לשינויים:

1. **נרמול ע״י MIN_MAX**

א. עץ החלטה: הבחירה הראשונית של שימוש בקנה מידה מינימלי-מקסימלי עם מודל עץ ההחלטות הביאה לביצועים נאותים בכל המדדים. עם זאת, הבחנו שהדיוק, הריקול וציון ה F1- היו נמוכים יחסית לאפשרויות אחרות.

ב. יער אקראי: עם המעבר למודל היער האקראי, נצפה שיפור משמעותי בכל המדדים. הדיוק, וציון ה F1- עלו בצורה ניכרת, מה שמצביע על כך שאופי של מודל יער אקראי התאים יותר למערך הנתונים.

ג. naivebayes: למרות תוצאות סבירות שהושגו עם עץ ההחלטות והיער האקראי, מודל naivebayes הניב ביצועים מעט נמוכים יותר בכל המדדים.

ד. KNN: נבדקו ערכים שונים של K (3,5,7) תוך שימוש בקנה מידה מינימלי-מקסימלי. התוצאות הצביעו על כך ש k=5-הפיק את הביצועים הכוללים הטובים ביותר. היא נבחרה כאופציה המועדפת במקרה הזה.

2. **נרמול ע״י Z-score**

א. עץ החלטה: מעבר לקנה מידה של ציון z עם מודל עץ ההחלטות הביא לשיפור הביצועים בכל המדדים.

ב. יער אקראי: בדומה לעץ ההחלטות, יער אקראי עם קנה מידה של ציון z השיג ביצועים גבוהים יותר בהשוואה למקבילו לקנה מידה מינימלי. המדדים השתפרו באופן משמעותי, והדגישו עוד יותר את האפקטיביות של קנה מידה של ציון z בשיפור יכולת הניבוי של המודל.

ג. Naivebayes: הביצועים של Bayes הנאיביים עם קנה מידה של ציון z נשארו דומים לביצועים שלו עם קנה מידה מינימום-מקסימום.

ד. KNN: התוצאות של KNN עם קנה מידה של ציון z היו עקביות עם אפשרות קנה המידה של המינימום המקסימלי.

ה. K-means Clustering: בדומה לתרחיש הקודם, K-means Clustering עם קנה מידה Z-Score הניב ביצועים נמוכים יותר בהשוואה לדגמים אחרים.

3. Decimal scaling

א. עץ החלטה: מודל עץ ההחלטות עם קנה מידה עשרוני הפגין ביצועים דומים לקנה מידה מינימום-מקסימום, והשיג תוצאות נאות בכל המדדים. עם זאת, הוא עדיין נפל בהשוואה לקנה מידה של ציון z במונחים של דיוק כולל וציון F1

ב. יער אקראי: מודל היער האקראי עם קנה מידה עשרוני הציג ביצועים משופרים מעט בהשוואה לקנה מידה מינימלי. אמנם המדדים לא היו שונים באופן משמעותי, אך העלייה ברמת הדיוק, הדיוק, הריקול והציון F1 הייתה בולטת.

ג. Naivebayes: המודל עם קנה מידה עשרוני הראה ביצועים דומים כמו בטכניקות קנה המידה האחרות.

ד. KNN: המודל עם קנה מידה עשרוני הציג תוצאות עקביות.

4. נרמול לפי חלוקה לפחים

א. עץ החלטה: בעת החלת נורמליזציה של חלוקה בינומית על עמודות ספציפיות, ראינו תוצאות דומות לגישת קנה המידה של המינימום המקסימלי. מודל עץ ההחלטות הציג ביצועים נאותים בכל המדדים, אך לא עלה על הביצועים של קנה מידה של ציון z .

ב. יער אקראי: ניצול נורמליזציה של חלוקה בינומית במודל היער האקראי הוביל לשיפור ניכר בכל המדדים. הדיוק, הדיוק, ההיזכרות והציון F1 עלו באופן משמעותי, מה שמצביע על כך שאלגוריתם היער האקראי התאים היטב לטיפול במערך הנתונים בטכניקת נורמליזציה זו.

ג. Naivebayes: בדומה לעץ ההחלטות, מודל Bayes הנאיבי הראה ביצועים עקביים עם נורמליזציה של חלוקה בינומית, אם כי מעט נמוך יותר בהשוואה לטכניקות נורמליזציה אחרות.

ד. KNN: ניסינו עם ערכים שונים של (K=3,5,7) באמצעות נורמליזציה של חלוקה בינומית. התוצאות הצביעו על כך ש-K=5 הפיק את הביצועים הכוללים הטובים ביותר, מה שהפך אותו לאופציה המועדפת עבור KNN עם טכניקת נורמליזציה זו.

כחלק מהתהליך השלם, בדקנו את ערך האנטרופיה עבור כל עמודה וניסינו להריץ שוב את המודלים השונים לאחר בחירת פיצ׳רים, כלומר בחרנו עמודות חדשות בעלות ערך האנטרופיה הנמוך ביותר. הערכנו שלאחר בחירת עמודות בעלות אנטרופיה נמוכה נצליח לשפר את המודל.

עם זאת, ההשפעה של בחירת תכונה על המדדים לא הייתה כצפוי. בניגוד להשערה הראשונית, המדדים הראו למעשה ירידה בביצועים לאחר בחירת תכונה, במיוחד במונחים של דיוק. עניין זה מרמז שייתכן שהתכונות שנבחרו לא ייצגו כראוי את הדפוסים הבסיסיים בנתונים, מה שגרם לירידה ברמת הדיוק החזוי של המודל. בהשוואת המדדים שהתקבלו לאחר בחירת תכונות עם המדדים שהושגו בעת הפעלת המודלים על כל הנתונים, ניכר כי בחירת תכונות לא שיפרה את הביצועים הכוללים.

לאחר ניתוח מדוקדק, ניתן לייחס מספר סיבות לתפקוד נמוך של המודל לאחר בחירת תכונה. ראשית, ייתכן שאלגוריתם בחירת התכונות לא תפס ביעילות את התכונות המבדילות ביותר עבור משימת הסיווג. ייתכן שהגישה מבוססת האנטרופיה במקרה זה לא הייתה מתאימה למערך הנתונים, וכתוצאה מכך הוסרו תכונות חשובות. שנית, הנוכחות של תכונות מתואמות או הרגישות של אלגוריתם היער האקראי שהוכח כמודל הטוב ביותר עבור סט הנתונים שלנו לתת-קבוצת התכונות שנבחרה עשוי להשפיע לרעה על ביצועי המודל.

כמו כן, ייתכן שבחירת התכונות לא שיפרה את התוצאות לאור כך שגילנו במהלך העבודה שהעמודה מגדר הינה עמודה לא מדויקת לחיזוי בסט הנתונים שלנו ואילו היינו בוחרות לחזות את העמודה של יעילות שעות שינה ייתכן שהתוצאות היו משתפרות.

אף על פי זאת, לאור כך שהשגנו תוצאות גבוהות במיוחד 97.8% במודל היער האקראי, עם סוגי נרמול שונים ראינו לנכון לא לשנות שוב את בחירת העמודות ולהשאיר את המודל הטוב שהושג עם סט הנתונים השלם בתוספת עמודת האשכולות החדשה.

המודל הנבחר ומסקנות

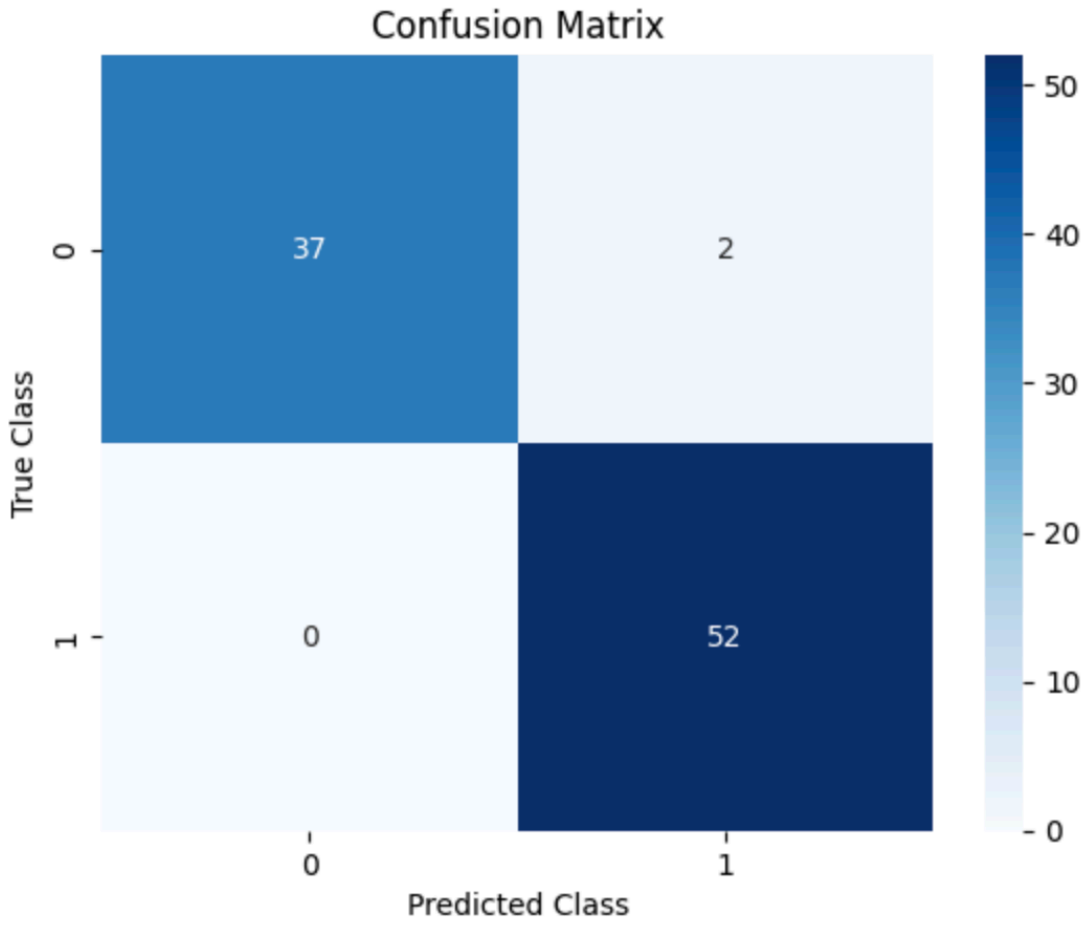
בהתבסס על הניתוח המקיף שביצענו לאורך הפרויקט, המודל הסופי שבחרנו כמודל בעל הביצועים הטובים ביותר הוא **מודל היער האקראי עם נורמליזציה מותאמת אישית לפי עמודות**. מודל זה הניב תוצאות מרשימות, והשיג **דיוק של 97.8% והפגין שיפורים משמעותיים לעומת טכניקות ומודלים מקדימים אחרים שנחקרו במהלך הפרויקט שלנו.**

אלגוריתם היער האקראי, שיטת למידה אנסמבלית, שיחקה תפקיד מרכזי בהצלחת המודל הסופי שלנו. על ידי שילוב של עצי החלטה מרובים, אלגוריתם זה לכד ביעילות דפוסים מורכבים בתוך מערך הנתונים וערך תחזיות מדויקות. על ידי צבירה של התחזיות של עצים מרובים, מודל היער האקראי הפיק תוצאות חזקות ומהימנות.

אחד היתרונות המרכזיים של אלגוריתם Random Forest הוא היכולת שלו להתמודד עם מספר רב של תכונות. בפרויקט שלנו, **זה היה מועיל במיוחד מכיוון שהיה לנו סט מגוון של תכונות לעבוד איתם.** יתרה מכך, **הבחירה האקראית של התכונות במהלך תהליך האימון סייעה בהפחתת ההשפעה של תכונות רועשות או לא רלוונטיות בודדות**, והעלתה את כוח הניבוי הכולל של המודל. על ידי מינוף מנגנון בחירת תכונה זה, מודל ה Random-Forest זיהה והשתמש ביעילות בתכונות האינפורמטיביות ביותר, מה שהוביל לשיפור הביצועים.

במהלך תהליך שיפור המודל, גילינו שעמודות שונות במערך הנתונים שלנו הגיבו בצורה שונה לטכניקות נורמליזציה שונות. כדי לנצל את התובנה הזו, **תכננו פונקציה המיישמת שיטות נורמליזציה ספציפיות על כל עמודה, תוך התחשבות בטכניקת הנורמליזציה שהשיגה את התוצאה הטובה ביותר במהלך שלב הניסויים שלנו.** על ידי שילוב נורמליזציה מותאמת אישית זו, הצלחנו לייעל את ייצוג הנתונים ולספק למודל תשומות משמעותיות יותר וניתנות להשוואה.

המודל הסופי הפגין דיוק יוצא דופן, ריקול וציון F1 וכמו כן הפגין את יכולתו לבצע תחזיות מדויקות לסווג ביעילות מופעים בתוך מערך הנתונים. מטריצת הבלבול (טעויות) ממחישה את הביצועים של המודל, עם זיהוי נכון של 37 נשים מתוך 39 ושל 52 גברים מתוך 52 (הנתונים האלו מייצגים את הביצועים של המודל על סט הנתונים שהוקצה לבדיקה (x_test)), מה שמצביע על מהימנותו בחיזוי מקרים חיוביים.



לסיכום, באמצעות הערכה מקיפה של טכניקות ומודלים שונים של עיבוד מקדים, הפרויקט שלנו זיהה בהצלחה את מסווג היער האקראי עם נורמליזציה ספציפית לעמודה מותאמת אישית **כפתרון האופטימלי עבור מערך הנתונים הנתון (98%).** על ידי מינוף למידת אנסמבל, בחירת תכונות וטכניקות נורמליזציה מותאמות, השגנו תוצאות יוצאות דופן **וסיפקנו מודל אמין לחיזוי מגדר וניתוח יעילות השינה .**

בברכה,
מתן ניצן ואסנת שבתאי

נספחים

א. השוואה בין התוצאות השונות (כולל התוצאה הטובה ביותר) ופירוט העיבוד המקדים והמודל סיווג שבוצע מפורט בהרחבה בקובץ אקסל שהועלה לדרייב וכמו כן יעלה למערכת המודל יחד עם כל הפרויקט והמסמכים השונים.

מצ״ב הקישור –

https://docs.google.com/spreadsheets/d/1JXlyQpO-mgKwAV70LaJ1a2kO_4B8pWWG/edit#gid=1564336516

1													
2	Check	MIN_MAX scalar	Z-Score	Decimal Scaling	Unit Vector Normalization	Bin Boundaries	Bin Means depth	Model	Precision	Recall	F-measure	Accuracy	Confusion Matrix
3	1	All the columns						DecisionTreeClassifier	85%	85%	85%	85%	[35 8] [6 46]
4	2	All the columns						Random Forest classifier	95%	95%	94%	95%	[35 4] [1 51]
5	3	All the columns						naivebayes	82%	82%	82%	82%	[30 9] [7 45]
6	4	All the columns						KNN (k = 5)	87%	87%	87%	87%	[31 8] [4 48]
7	5	All the columns						KNN(k = 3)	81%	80%	80%	80%	[26 13] [5 47]
8	6	All the columns						KNN(k = 7)	86%	86%	86%	86%	[30 9] [4 48]
9	7	All the columns						K-means Clustering	51%	51%	51%	58%	[20 19] [19 33]
10	8		All the columns					DecisionTreeClassifier	87%	87%	87%	87%	[34 5] [7 45]
11	9		All the columns					Random Forest classifier	97%	97%	97%	97%	[36 3] [0 52]
12	10		All the columns					naivebayes	82%	82%	82%	82%	[30 9] [7 45]
13	11		All the columns					KNN (k = 5)	87%	87%	87%	87%	[32 7] [5 47]
14	12		All the columns					KNN(k = 3)	85%	85%	85%	85%	[31 8] [6 46]
15	13		All the columns					KNN(k = 7)	87%	87%	87%	87%	[32 7] [5 47]
16	14		All the columns					K-means Clustering	52%	31%	39%	58%	[12 27] [11 41]
17	15			All the columns				DecisionTreeClassifier	85%	85%	85%	85%	[33 6] [8 44]
18	16			All the columns				Random Forest classifier	97%	97%	97%	97%	[37 2] [1 51]
19	17			All the columns				naivebayes	82%	82%	82%	82%	[30 9] [7 45]
20	18			All the columns				KNN (k = 5)	87%	87%	87%	87%	[33 6] [6 46]
21	19			All the columns				KNN(k = 3)	82%	82%	82%	82%	[31 8] [8 44]
22	20			All the columns				KNN(k = 7)	86%	86%	86%	86%	[33 6] [7 45]
23	21			All the columns				K-means Clustering	37%	49%	42%	42%	[19 20] [33 19]
24	22				All the columns			DecisionTreeClassifier	82%	82%	82%	82%	[33 8] [8 44]
25	23				All the columns			Random Forest classifier	97%	97%	97%	97%	[37 2] [1 51]
26	24				All the columns			naivebayes	75%	74%	74%	74%	[31 8] [16 36]
27	25				All the columns			KNN (k = 5)	68%	67%	67%	67%	[27 12] [18 34]
28	26				All the columns			KNN(k = 3)	72%	72%	72%	72%	[29 10] [16 36]
29	27				All the columns			KNN(k = 7)	66%	66%	66%	66%	[25 14] [17 35]
30	28				All the columns			K-means Clustering	51%	51%	51%	58%	[20 19] [19 33]
31													
32													
33	selectedColumns=['Smoking status', 'Caffeine consumption', 'Alcohol consumption','Exercise frequency', '!												
34	גרסה נוספת עם אוכלוסיה נוכחית - sselectedColumns												
35	Check	MIN_MAX scalar	Z-Score	Decimal Scaling	Unit Vector Normalization	Bin Boundaries	Bin Means depth	Model	Precision	Recall	F-measure	Accuracy	Confusion Matrix
36	1	All the columns						DecisionTreeClassifier	85%	85%	85%	85%	[34 5] [9 43]
37	2	All the columns						Random Forest classifier	90%	90%	90%	90%	[33 6] [4 48]
38	3	All the columns						naivebayes	82%	82%	82%	82%	[29 10] [6 46]
39	4	All the columns						KNN (k = 5)	82%	81%	81%	81%	[33 6] [11 41]
40	5	All the columns						KNN(k = 3)	80%	80%	80%	80%	[31 8] [10 42]
41	6	All the columns						KNN(k = 7)	83%	82%	82%	82%	[32 7] [9 43]
42	7	All the columns						K-means Clustering	33%	33%	33%	42%	[13 26] [27 25]
43	8		All the columns					DecisionTreeClassifier	85%	85%	85%	85%	[34 5] [9 43]
44	9		All the columns					Random Forest classifier	89%	89%	89%	89%	[33 6] [4 48]
45	10		All the columns					naivebayes	82%	82%	82%	82%	[29 10] [6 46]
46	11		All the columns					KNN (k = 5)	83%	82%	82%	82%	[33 6] [10 42]
47	12		All the columns					KNN(k = 3)	84%	84%	84%	84%	[32 7] [8 44]
48	13		All the columns					KNN(k = 7)	82%	82%	82%	82%	[31 8] [8 44]
49	14		All the columns					K-means Clustering	52%	74%	61%	59%	[29 10] [27 25]
50	15			All the columns				DecisionTreeClassifier	84%	84%	84%	80%	[33 6] [9 43]
51	16			All the columns				Random Forest classifier	89%	89%	89%	89%	[33 6] [4 48]
52	17			All the columns				naivebayes	82%	82%	82%	82%	[29 10] [6 46]
53	18			All the columns				KNN (k = 5)	80%	80%	80%	80%	[29 10] [8 44]
54	19			All the columns				KNN(k = 3)	79%	79%	79%	79%	[30 9] [10 42]
55	20			All the columns				KNN(k = 7)	78%	78%	78%	78%	[29 10] [12 42]
56	21			All the columns				K-means Clustering	46%	74%	57%	52%	[29 10] [34 18]
57	22				All the columns			DecisionTreeClassifier	84%	84%	84%	84%	[33 6] [9 43]
58	23				All the columns			Random Forest classifier	88%	88%	88%	88%	[32 7] [4 48]
59	24				All the columns			naivebayes	77%	77%	77%	77%	[29 10] [11 41]
60	25				All the columns			KNN (k = 5)	83%	84%	83%	84%	[31 8] [7 45]
61	26				All the columns			KNN(k = 3)	86%	86%	86%	86%	[33 6] [7 45]
62	27				All the columns			KNN(k = 7)	83%	84%	83%	84%	[31 8] [7 45]
63	28				All the columns			K-means Clustering	21%	28%	24%	23%	[11 28] [42 10]
64													
65													
66													
67	רגדום של עמודות ונרמולים												
68													
69	Check	MIN_MAX scalar	Z-Score	Decimal Scaling	Unit Vector Normalization	Bin Boundaries	Bin Means depth	Model	Precision	Recall	F-measure	Accuracy	Confusion Matrix
70	1	Exercise frequency				Light sleep percentage		DecisionTreeClassifier	85%	85%	84%	85%	[36 9] [5 47]
71	2	Exercise frequency				Light sleep percentage		Random Forest classifier	97%	97%	97%	97%	[37 2] [1 51]
72	3	Exercise frequency				Light sleep percentage		naivebayes	77%	77%	77%	77%	[29 10] [11 41]
73	4	Exercise frequency				Light sleep percentage		KNN (k = 5)	75%	74%	74%	74%	[30 9] [15 37]
74	5	Exercise frequency				Light sleep percentage		KNN(k = 3)	73%	73%	73%	73%	[29 10] [15 37]
75	6	Exercise frequency				Light sleep percentage		KNN(k = 7)	75%	74%	74%	74%	[30 9] [15 37]
76	7	Exercise frequency				Light sleep percentage		K-means Clustering	51%	51%	51%	58%	[20 19] [19 33]
77	8	Sleep efficiency	Smoking status,Bedtime		Age	) percentage, REM sleep p	Deep sleep percentage,Sleep duration, Awakenings	DecisionTreeClassifier	81%	80%	80%	80%	[26 13] [5 47]
78	9	Sleep efficiency	Smoking status,Bedtime		Age	) percentage, REM sleep p	Deep sleep percentage,Sleep duration, Awakenings	Random Forest classifier	80%	80%	80%	80%	[29 10] [8 44]
79	10	Sleep efficiency	Smoking status,Bedtime		Age	) percentage, REM sleep p	Deep sleep percentage,Sleep duration, Awakenings	naivebayes	69%	69%	68%	69%	[20 19] [9 43]
80	11	Sleep efficiency	Smoking status,Bedtime		Age	) percentage, REM sleep p	Deep sleep percentage,Sleep duration, Awakenings	KNN (k = 5)	63%	63%	63%	63%	[23 16] [18 34]
81	12	Sleep efficiency	Smoking status,Bedtime		Age	) percentage, REM sleep p	Deep sleep percentage,Sleep duration, Awakenings	KNN(k = 3)	62%	62%	62%	62%	[22 17] [18 34]
82	13	Sleep efficiency	Smoking status,Bedtime		Age	) percentage, REM sleep p	Deep sleep percentage,Sleep duration, Awakenings	KNN(k = 7)	63%	63%	63%	63%	[23 16] [18 34]
83	14	Sleep efficiency	Smoking status,Bedtime		Age	) percentage, REM sleep p	Deep sleep percentage,Sleep duration, Awakenings	K-means Clustering	40%	49%	44%	47%	[19 20] [28 24]
84	15	Alcohol consumption,Exercise frequency	Caffeine consumption		Smoking status,Age, Deep sleep percent	Sleep duration,Bedtime	sleep percentage,Wakeup time	DecisionTreeClassifier	82%	82%	82%	82%	[36 9] [7 45]
85	16	Alcohol consumption,Exercise frequency	Caffeine consumption		Smoking status,Age, Deep sleep percent	Sleep duration,Bedtime	sleep percentage,Wakeup time	Random Forest classifier	88%	88%	88%	88%	[31 8] [1 49]
86	17	Alcohol consumption,Exercise frequency	Caffeine consumption		Smoking status,Age, Deep sleep percent	Sleep duration,Bedtime	sleep percentage,Wakeup time	naivebayes	62%	60%	61%	60%	[24 13] [23 29]
87	18	Alcohol consumption,Exercise frequency	Caffeine consumption		Smoking status,Age, Deep sleep percent	Sleep duration,Bedtime	sleep percentage,Wakeup time	KNN (k = 5)	69%	68%	68%	68%	[27 12] [17 35]
88	19	Alcohol consumption,Exercise frequency	Caffeine consumption		Smoking status,Age, Deep sleep percent	Sleep duration,Bedtime	sleep percentage,Wakeup time	KNN(k = 3)	64%	64%	64%	64%	[23 16] [19 33]
89	20	Alcohol consumption,Exercise frequency	Caffeine consumption		Smoking status,Age, Deep sleep percent	Sleep duration,Bedtime	sleep percentage,Wakeup time	KNN(k = 7)	69%	68%	68%	68%	[27 12] [17 35]
90	21	Alcohol consumption,Exercise frequency	Caffeine consumption		Smoking status,Age, Deep sleep percent	Sleep duration,Bedtime	sleep percentage,Wakeup time	K-means Clustering	47%	56%	51%	54%	[22 17] [25 27]
91	22		Alcohol consumption			Wakeup time,Smoking status,Deep sleep percentage		DecisionTreeClassifier	84%	84%	83%	84%	[36 9] [6 46]
92	23		Alcohol consumption			Wakeup time,Smoking status,Deep sleep percentage		Random Forest classifier	97%	97%	97%	97%	[36 3] [0 52]
93	24		Alcohol consumption			Wakeup time,Smoking status,Deep sleep percentage		naivebayes	78%	78%	78%	78%	[28 11] [9 43]
94	25		Alcohol consumption			Wakeup time,Smoking status,Deep sleep percentage		KNN (k = 5)	67%	67%	67%	67%	[24 15] [15 37]
95	26		Alcohol consumption			Wakeup time,Smoking status,Deep sleep percentage		KNN(k = 3)	57%	56%	56%	56%	[22 17] [23 29]
96	27		Alcohol consumption			Wakeup time,Smoking status,Deep sleep percentage		KNN(k = 7)	67%	67%	67%	67%	[24 15] [15 37]
97	28		Alcohol consumption			Wakeup time,Smoking status,Deep sleep percentage		K-means Clustering	51%	51%	51%	58%	[20 19] [19 33]
98													
99													
100													
101	תוספת עמודה של k-means אל הראשה הישלמת שלנו ובינתו מחלים על הראשה החדשה												
102	Check	MIN_MAX scalar	Z-Score	Decimal Scaling	Unit Vector Normalization	Bin Boundaries	Bin Means depth	Model	Precision	Recall	F-measure	Accuracy	Confusion Matrix
103	1	All the columns						DecisionTreeClassifier	86%	86%	86%	86%	[32 7] [6 46]
104	2	All the columns						Random Forest classifier	95%	95%	94%	95%	[35 4] [1 51]
105	3	All the columns						naivebayes	82%	82%	82%	82%	[30 9] [7 45]
106	4	All the columns						KNN (k = 5)	87%	87%	87%	87%	[31 8] [4 48]
107	5	All the columns						KNN(k = 3)	81%	80%	80%	80%	[26 13] [5 47]
108	6	All the columns						KNN(k = 7)	86%	86%	86%	86%	[30 9] [4 48]
109	8		All the columns					DecisionTreeClassifier	86%	86%	86%	86%	[34 5] [8 44]
110	9		All the columns					Random Forest classifier	97%	97%	97%	97%	[37 2] [1 51]
111	10		All the columns					naivebayes	82%	82%	82%	82%	[30 9] [7 45]
112	11		All the columns					KNN (k = 5)	87%	87%	87%	87%	[32 7] [5 47]
113	12		All the columns					KNN(k = 3)	84%	84%	83%	84%	[30 9] [6 46]
114	13		All the columns					KNN(k = 7)	87%	87%	87%	87%	[32 7] [5 47]
115	15			All the columns				DecisionTreeClassifier	85%	85%	85%	85%	[31 8] [6 46]
116	16			All the columns				Random Forest classifier	96%	96%	96%	96%	[36 3] [1 51]
117	17			All the columns				naivebayes	82%	82%	82%	82%	[30 9] [7 45]
118	18			All the columns				KNN (k = 5)	87%	87%	87%	87%	[33 6] [6 46]
119	19			All the columns				KNN(k = 3)	82%	82%	82%	82%	[31 8] [8 44]
120	20			All the columns				KNN(k = 7)	86%	86%	86%	86%	[33 6] [7 45]
121	22				All the columns			DecisionTreeClassifier	86%	86%	86%	86%	[32 7] [6 46]
122	23				All the columns			Random Forest classifier	98%	98%	98%	98%	[37 2] [0 52]
123	24				All the columns			naivebayes	75%	74%	74%	74%	[31 8] [16 36]
124	25				All the columns			KNN (k = 5)	68%	67%	67%	67%	[27 12] [18 34]
125	26				All the columns			KNN(k = 3)	72%	72%	72%	72%	[29 10] [16 36]
126	27				All the columns			KNN(k = 7)	66%	66%	66%	66%	[25 14] [17 35]
127													
128													
129													
130	לאחר בחירת המודל הטוב ביותר עם הנמלר הטוב ביותר אלו התוצאות												
131													
132													
133	Check	MIN_MAX scalar	Z-Score	Decimal Scaling	Unit Vector Normalization	Bin Boundaries	Bin Means depth	Model	Precision	Recall	F-measure	Accuracy	Confusion Matrix
134		Exercise frequency	Alcohol consumption		Light sleep percentage,Deep sleep percentage,Smoking status,Wakeup time			Random Forest classifier	98%	98%	98%	98%	[37 2] [0 52]]
135													