תוכן עניינים

חלק א' – ניתוח נתונים משותפים

1. טיפול בדאטה ותיקונו

במספר לא מבוטל של קבצים היו מגוון שגיאות – בסכמה של הטבלה, בשמות של הקבצים, בערכים שבטבלאות וכדו' – אשר דרשו מאיתנו לטפל בהן ולתקן אותן, על מנת לקבל פורמט אחיד בקבצים, שיאפשר לנו לנתח את הנתונים כראוי. דוגמאות למספר טעויות: שמות עמודות לא תקין, אי-איפוס זמן בכל קובץ (מה שהוביל לזמנים גדולים מאוד בחלק מהקבצים) או איפוס הזמן באמצע התיעוד ועוד. **אופן הטיפול והתיקון של הדאטה –**

א. קבצים עם סכמה לא נכונה הכילו את הרשומה עם מספר הצעדים המוערך, "Estimated", או לחלופין היו שתי שורות רווח בין הנתונים היבשים של הקובץ לבין נתוני התאוצות; מה שאילץ אותנו לאכוף בצורה יותר קונקרטית מאיזה רשומה להתחיל לקרוא את נתוני התאוצות.

ב. קבצים עם זמן לא ממוין/מצטבר (כלומר, התיעוד לא התחיל מזמן אפס בכל קובץ): בהתחשב בכך שעשינו אגרגציה לרשומות כך שכל קובץ הצטמצם לכדי רשומה אחת (כפי שיפורט בחלק של **עיבוד הדאטה**), לא הייתה משמעות לעמודה של זמן, ולכן השארנו קבצים כאלה, בתנאי שנתוני התאוצה נראו לנו תקינים.

ג. נתוני תאוצה חסרים (תאים ריקים בטבלה)/קפיצות חריגות במהלך התיעוד – בחלק מהקבצים, היו רשומות עם נתונים חסרים או לחלופין קיצוניים מאוד. כמענה לכך, חיפשנו את הרשומות הדומות ביותר לרשומות אלה והשלמנו את הערכים החסרים/החלפנו את הערכים הקיצוניים לפי הערכים של הרשומות הדומות.

ד. נתוני זבל – היו קבצים עם המון רשומות שמכילות ערכי זבל בצורה של מספרים ומחרוזות, לדוגמא "10.67-1.88". לא הצלחנו למצוא דרך לטפל בבעיה זו ועל כן נאלצנו (בצער רב) לזרוק קבצים אלה.

ה. תיעודים לא אמינים – היו שני סוגים של תיעודים לא אמינים. האחד, היה כמות צעדים לא ריאלית ביחס לזמן המתועד, למשל תיעוד של 110 צעדים לבאורה ב-5 שניות. מדובר בנתונים שלא תואמים את המציאות ואף סותרים את היכולות הפיזיות של האדם ולכן זרקנו קבצים עם תיעודים כאלה. הסוג השני של תיעוד לא אמין היה ביצועים לא עקביים של אותו האדם, כלומר, תיעוד שבו הוא רץ יותר צעדים בפחות זמן ביחס לתיעוד אחר שלו. בשני המקרים הנ"ל נאלצנו לזרוק את הקבצים על מנת לא לפגוע בלמידה של המודלים.

- הקו המנחה שלנו היה לנסות לתקן כמה קבצים שאפשר, על מנת להשאיר כמה שיותר דאטה עבור המודלים, ולזרוק קבצים רק כאשר לא הייתה לנו ברירה אחרת.

2. עיבוד הדאטה והנדסת פיצ'רים

כפי שלמדנו בקורס, על מנת לקבל אחידות בדאטה ולנטרל את השונות וההטיה בניתוח הנתונים ואימון המודל שעלולה לנבוע מאופן קיבוע ה-IMU על הגוף, ביצענו טרנספורמציה על נתוני התאוצה ב-3 הצירים והפכנו אותם לנורמה האוקלידית ע"פ הנוסחה – $N = \sqrt[2]{a_x^2 + a_y^2 + a_z^2}$. בנוסף לכך, עבור שני המודלים, ביצענו אגרגציה והוספת פיצ'רים כך שכל קובץ הוצג ע"י רשומה בודדת. בהקשר של הפיצ'רים, הרעיון שעמד מאחורי יצירת חלק מהם התבסס על הממצאים בחלק של **ניתוחים סטטיסטיים**, כפי שיפורט בהמשך. נתאר תחילה את הפיצ'רים שחילצנו מהנתונים פר קובץ לאחר ביצוע הטרנספורמציה לעיל –

1) **ממוצע הנורמות:** חישוב ממוצע הנורמות.
2) **סטיית התקן של הנורמות:** חישוב סטיית התקן של הנורמות.

3) **סוג פעילות:** אינדוקס סוג הפעילות כך ש-1 מייצג ריצה ו-0 הליכה.

4) **Steps with lower bound:** הערכת כמות הצעדים שנלקחו, כאשר התחלת הצעד אופיינה ע"י נורמה שחוצה סף עליון מסוים (פרמטר חופשי שמשלב את הממוצע ואת סטיית התקן של הנורמות) וסוף הצעד אופיין ע"י נורמה שיורדת מתחת לסף תחתון כלשהו (משתנה חופשי ג1ם כן, שמחושב לפי הסף העליון).

5) **Manual peak detection:** דרך נוספת להערכת כמות הצעדים; ספירת כמות נקודות המקסימום הלוקליות, שגם חוצות את אותו הסף העליון שהוזכר לעיל.

6) **Signal find peaks:** כמות נקודות המקסימום ביחס לגודל סביבה מסוימת של נקודות (פרמטר חופשי לבחירתנו).

7) **Peak diff:** ממוצע ההפרשים בין פיקים שחצו את הסף העליון שהוזכר לעיל.

8) **Above mean:** כמות הנורמות שהיו גדולות מהממוצע.

הפיצ'רים שהשתמשנו בהם בכל אחד מהמודלים הם –

א. <u>עבור המודל המסווג הליכה/ריצה</u>: הממוצע וסטיית התקן של הנורמות.

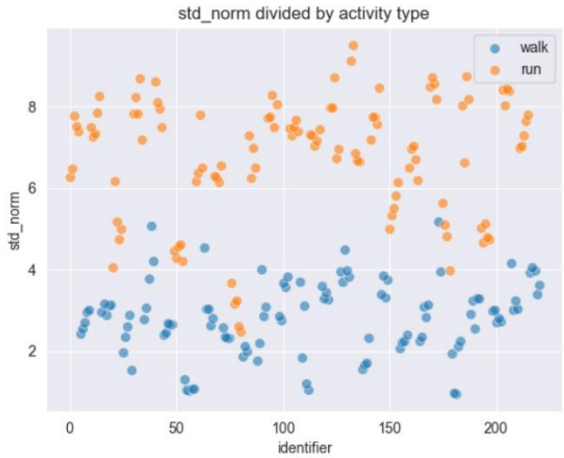
ב. <u>עבור המודל שחוזה את כמות הצעדים</u>: שאר הפיצ'רים שהוזכרו לעיל.

- **<u>הערה</u>:** לא ביצענו החלקה של הנורמות באמצעות ממוצע נע כפי שלמדנו בקורס מפני שזה השפיע לרעה על הביצועים של המודל.

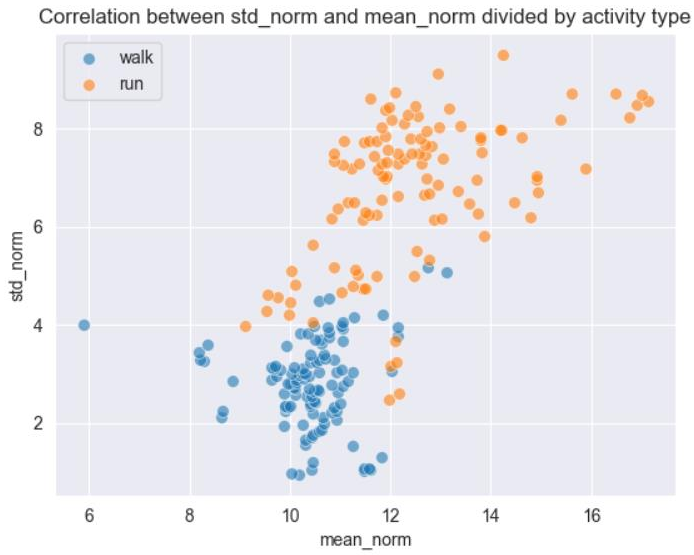
3. <u>ניתוחים סטטיסטיים</u>

א. <u>עבור המודל המסווג הליכה/ריצה</u> –

ניתחנו תחילה את ההבדלים בממוצע ובסטיית התקן של הנורמות בזמן הליכה לעומת זמן ריצה, וקיבלנו את הממצאים הבאים:



ניתן לראות באופן דיי מובהק כי מרבית הערכים שהתקבלו מריצה גדולים יותר מהערכים שהתקבלו מהליכה, הן בממוצע והן בסטיית התקן. בהמשך ישיר לכך, ועל מנת לקבל תוצאות נוספות שיחזקו את הממצאים שקיבלנו לעיל, יצרנו תרשים scatter נוסף שבו כל נקודה מייצגת (Mean, STD) וציפינו לקבל מעין שני צנטרואידים נפרדים – אחד עבור הליכה ואחר עבור ריצה. אכן –

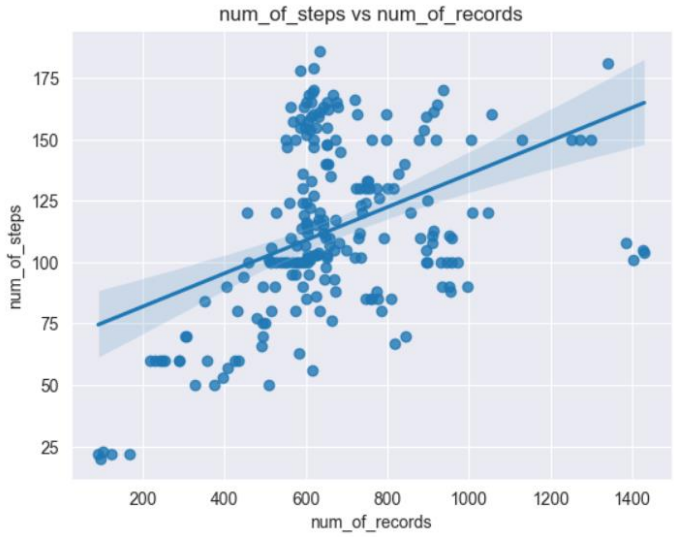


ניתן לראות התקבצות של תצפיות כתלות בסוג פעילות, כאשר התצפיות של ריצה פחות מלוכדות. על סמך כל הממצאים לעיל, הגענו לשתי מסקנות מרכזיות (ודיי צפויות) –

1. השונות יותר גבוהה בריצה (פיזור התצפיות מהווה ראייה לכך).
2. קיימת קורלציה בין ממוצע וסטיית תקן של הנורמה לבין סוג הפעילות.

ב. <u>עבור המודל שחוזה את כמות הצעדים</u> –

לאור הבעייתיות שהייתה בחלק מהקבצים עם תיעוד הזמנים בצורה תקינה, ובהתחשב בכך שבאגרגציה של הרשומות השמטנו את הזמנים, הדבר היחידי שהתאפשר לנו לבדוק בהקשר של צעדים בהינתן הנתונים היה קורלציה בין מספר הרשומות לבין מספר הצעדים המדווח. לשם כך, עשינו תרשים scatter והעברנו קו רגרסיה, וקיבלנו את התרשים הבא –



התרשים תואם מעט את הציפיות שהיו לנו, שכן מצד אחד ציפינו שייקח יותר זמן לתעד כמות גדולה יותר של צעדים, ומצד שני לקחנו בחשבון שמספרים אלה תלויים גם בקצב הדגימה ובדפוס ההליכה/ריצה האינדיבידואלי של כל אדם. בהתאם לכך, ניתן לראות שמספר רב של נקודות נפלו רחוק יחסית מקו הרגרסיה – מה שמעיד על מגמתיות וקורלציה נמוכה בין מספר הרשומות למספר הצעדים.

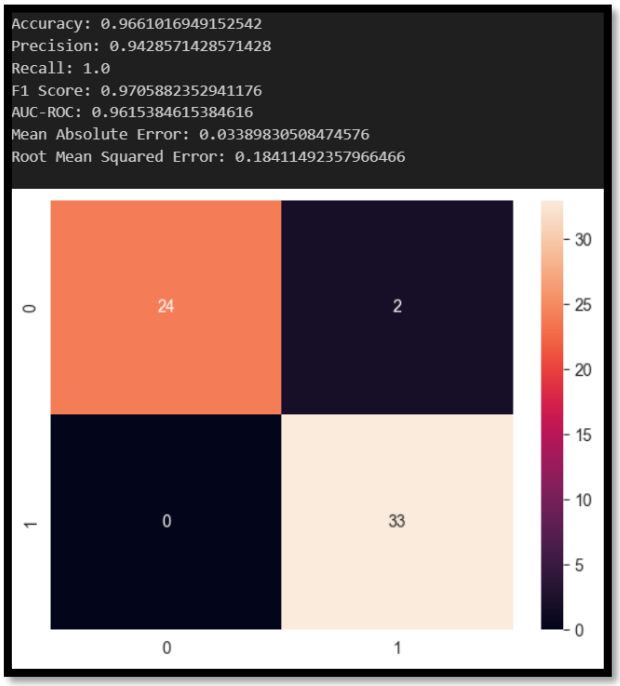
על מנת לחזק את הממצאים עוד יותר, בדקנו מה הערך של מדד הקורלציה פירסון, ואכן קיבלנו קורלציה נמוכה – $Pearson \approx 0.4364$.

4. המודלים ואימונם

חילקנו תחילה את הדאטה לסט אימון וסט מבחן. כפי שציינו קודם לכן, סביר להניח כי לכל אדם קיים דפוס הליכה ודפוס ריצה קבוע, אשר בא לידי ביטוי בין היתר בערכי התאוצות בשלושת הצירים, וכפועל יוצא מכך, גם בטרנספורמציה שלהם לנורמה. אי-לכך, הנחנו כי הימצאות של מקטעים **שונים** שנוצרו ע"י **אותו האדם** גם בסט האימון וגם בסט המבחן מהווה מעין "זליגת נתונים" שעלולה לפגוע במהימנות התוצאות שהמודל יניב. על סמך הנחה זו, יצרנו מזהה ייחודי לפי המספר של יוצר המקטע והפעילות שהוא תיעד – מה שאפשר לנו לאכוף שמקטעים של אותו יוצר שתועדה בהם אותה הפעילות יהיו רק באחד משני הסטים – ופיצלנו את הקבצים כך ש-70% נמצא בסט האימון ו-30% בסט הטסט. לאחר ביצוע החלוקה הנ"ל של סט הנתונים, חיפשנו את מודלי הלמידה שיניבו את הביצועים הטובים ביותר:

א. <u>עבור סיווג הליכה/ריצה</u> – בחנו את הביצועים של מספר מודלי סיווג, כגון Random Forest Classifier, Logistic Regression ו-XGBoost Classifier, והמודל שהניב את הביצועים הטובים ביותר היה XGBoost Classifier. נציג את ערכי ההיפר-פרמטרים שבחרנו ואת תוצאות המודל:

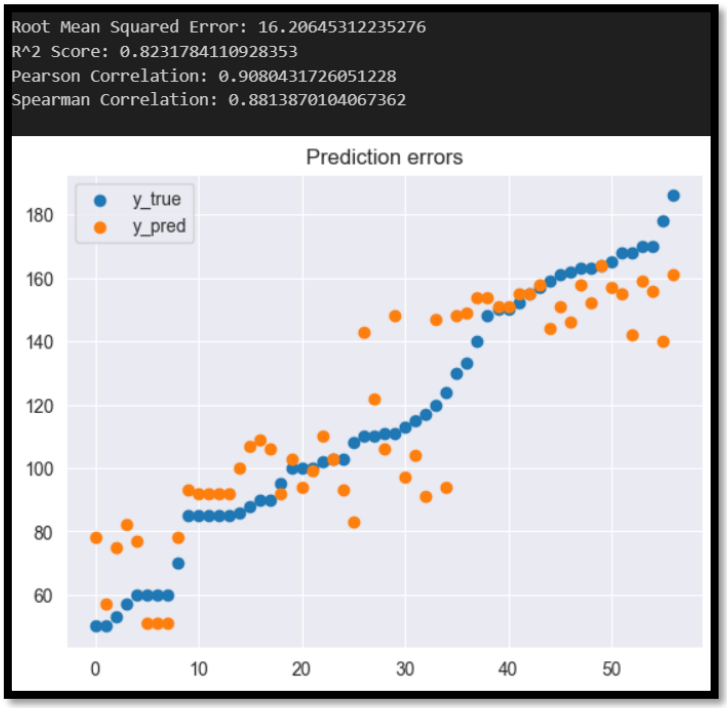
$$n\_estimators = 100, \qquad \max\_depth = 10, \qquad eta = 0.5$$
$$gamma = 0.1, \qquad reg\_lambda = 0.8$$
$$\min\_child\_weight = 2$$



ב. <u>עבור חיזוי מספר הצעדים</u> –

בחנו את הביצועים של שני מודלי רגרסיה – Lasso (מודל של רגרסיה ליניארית) ו- XGBoost Regressor. באופן כללי, לא הצלחנו למזער את ה-RMSE ולקרבו מספיק לאפס, מכיוון שלצד חיזויים מדויקים מאוד שהמודלים עשו, היו חיזויים רחוקים משמעותית מהמספר המקורי. אנו חושדים שהבעייתיות בנתונים שהצגנו בתחילת הדו"ח מנעה מהמודל ללמוד ולהתכנס בצורה אופטימלית.

המודל שהניב את הביצועים הטובים יותר היה XGBoost Regressor עם אותם היפר-פרמטרים שהצגנו לעיל. להלן התוצאות:



ניתן לראות כי מרבית החיזויים (התצפיות הכתומות) קרובים מאוד למספר הצעדים המקוריים (התצפיות הכחולות) – מה שמעיד על דיוק גבוה בחיזויים עבור תצפיות אלו.

חלק ב' – פיתוח מערכת ניטור פעילות גופנית

1. **סיפור מוצר**

1.1. רקע

באופן כללי בעולם, ובישראל בפרט, קיימת תופעה של שתייה מועטה מדי של מים ביום; ע"פ סקר עדכני שנערך השנה ע"י שטראוס (תמי4), הישראלים שותים בממוצע 7 כוסות מים ביום בלבד, זאת על אף שידוע כי מדובר במצרך חיוני לגוף. לראייה, 78% מהמשתתפים בסקר של שטראוס העידו כי הם הרגישו טוב יותר בימים שבהם שתו יותר מים. על כן, סביר להניח כי ישנו צורך במוצר שיזכיר לאנשים לשתות מים על בסיס קבוע, אך התזכורות צריכות להיות בהתבסס על מספר שיקולים, כגון פיזיולוגיים, ולא רק על סמך קבועי זמן. המוצר שלנו, "AquaStep", נותן מענה לצרכים אלו. המוצר עצמו הוא אפליקציה בפלאפון שמתחברת באמצעות Bluetooth לרכיב חומרתי, אשר מקובע על החגורה בגובה קו המותן ומספק מדדי תאוצה וטמפרטורה.

1.2. הפיצ'ר המרכזי של המוצר

פלטפורמה שמזכירה למשתמש לצרוך מים במהלך היום כתלות בטמפרטורת הסביבה, המאמץ הפיזי שהמשתמש עושה – הליכה/ריצה/מנוחה – ומספר הצעדים שהמשתמש ביצע במסגרת אותו מאמץ פיזי.

כמו כן, הפלטפורמה מתעדת את כמות כוסות המים שנשתו, וקובעת למשתמש יעד יומי של מספר כוסות לשתות על סמך הגורמים שצוינו לעיל ועל נתוני המשתמש "היבשים": משקל גוף ותדירות אימונים שבועית. יעד יומי זה הינו דינמי ועשוי להשתנות לאורך היום כתלות בפעילויות של המשתמש ובטמפרטורה.

1.3. <u>פיצ'רים נוספים באפליקציה</u>

א. התרעה <u>קולית</u> למשתמש על סכנת התייבשות אם תזוהה קיום פעילות ספורטיבית בטמפרטורה גבוהה מדי.

ב. פינת "הידעת?" – הצגת עובדות כלליות בנושאים של מים, ובפרט שתיית מים, כאשר בכל כניסה לאפליקציה תוצג עובדה אחרת.

ג. דאשבורד המכיל סטטיסטיקות יומיות למשתמש – כמות הכוסות ששתה עד כה, שריפת קלוריות, טמפרטורה.

ד. דאשבורד המכיל סטטיסטיקות וגרפים שבועיים למשתמש בנוגע לכמות כוסות המים שהוא שתה לאורך השבוע והקלוריות שהוא שרף.

ה. מצב שינה – טווח שעות שבו האפליקציה מפסיקה לשלוח התראות ולא מפריעה למשתמש בזמן שינה.

ו. קבלת חיווי מהמשתמש שהוא אכן שתה – בעת שליחת התראה על שתיית מים, יהיה לחצן שהמשתמש ילחץ עליו לאשר שאכן שתה.

ז. התראה <u>קולית</u> כשהמשתמש מגיע ליעד היומי של כוסות מים לשתות.

1.4. <u>לוגיקת שליחת התראות</u>

בכל פעם שהמשתמש יצבור לפחות כוס אחת של מים, הוא יקבל התראה לשתות.

2. **<u>בחירת חומרה ספציפית</u>**

א. ESP32 – משמש אותנו להתחברות באמצעות Bluetooth ולקבל נתוני תאוצה וטמפרטורה לאפליקציה בפלאפון.

ב. IMU – חיישן תאוצה שמשמש אותנו לאיסוף מידע רלוונטי למודלי ספירת הצעדים ולזיהוי סוג פעילות (ריצה/הליכה/מנוחה).

ג. Temp36 – חיישן טמפרטורה לשם מדידת טמפרטורת הסביבה.

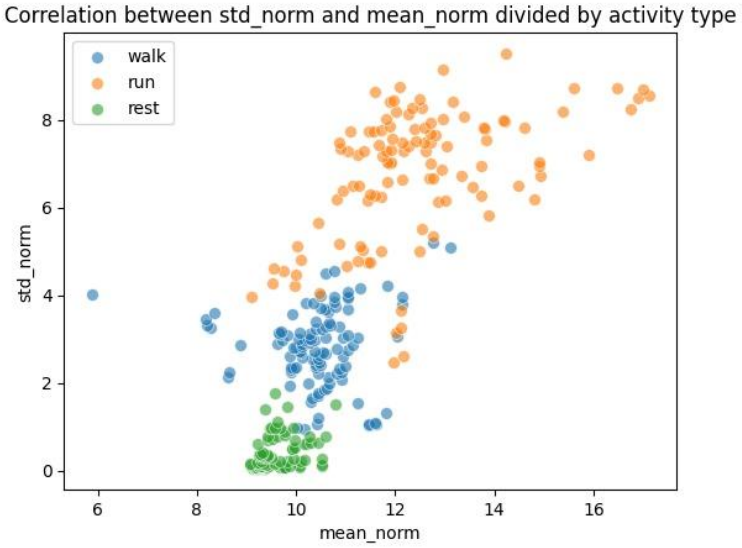
- שני החיישנים יותקנו על גבי החגורה שהמשתמש יחגור.

3. **<u>מודלי למידה</u>**

3.1. <u>סיווג פעילות משתמש</u>

לקחנו את אותו המודל שהשתמשנו בו עבור חלק א' והרחבנו אותו כך שיצליח לסווג גם מנוחה ולא רק הליכה/ריצה. לשם כך, יצרנו כמות של קבצי נתונים של מנוחה שתבטיח סט נתונים מאוזן (112 קבצים של ריצה, 109 קבצים של הליכה ו-91 קבצים של מנוחה), על מנת לאפשר למודל ללמוד בצורה מיטבית. לאחר מכן, נתנו למודל להתאמן על כל הקבצים ביחד – הן מחלק א' והן של המנוחה שיצרנו – כאשר גם כאן עשינו חלוקה של 30% לטסט ו-70% לאימון. מחקירה קטנה של הנתונים, גילינו מהר מאוד שגם המנוחה

מאופיינת ע"י טווח ערכים של ממוצע וסטיית תקן של הנורמה שונה ביחס להליכה וריצה. על מנת להמחיש זאת, כמו בחלק א', להלן תמונה של הצנטרואידים לאחר הוספת הנתונים של המנוחה –



הפוטנציאל להבחין דיי בקלות בין הפעולות בעזרת הממוצע וסטיית התקן של הנורמה בא לידי ביטוי בביצועים של המודל, אשר עמד על **97.87%** דיוק.

3.2. <u>חיזוי מספר צעדים</u>

לקחנו את המודל מחלק א' כפי שהוא, וההתאמה למנוחה נעשתה על ידנו באופן שכזה: אם המודל לסיווג פעילות קובע שהמשתמש נמצא במנוחה, אז באופן ידני וללא הפעלת מודל אנו קובעים (בצורה מאוד הגיונית ומתבקשת) שמספר הצעדים שנעשה במקטע הוא אפס.

- **<u>הערה:</u>** שני המודלים והמשקולות שלהם נשמרו לאחר אימונם בקבצים מסוג pickle. קבצים אלה אוחסנו לאחר מכן באחסון הפנימי של הפלאפון על מנת לטעון אותם לאפליקציה עצמה. קבצי ה-pickle נמצאים בתיקייה בשם **part_b_models_212984801_316111442_206713612**. בתיקייה זו נמצאת גם המחברת שבעזרתה אימנו את מודל סיווג הפעילות.

4. **<u>חישוב כמות כוסות המים לשתייה</u>**

נסמן את הפרטים היבשים שהמשתמש מזין לאפליקציה:

משקל גוף – $bodyWeight$, **שעות שינה בלילה בממוצע** – $sleepHours$, **תדירות פעילות גופנית שבועית** – $workoutWeeklyHours$. כמו כן, נסמן את **תדירות הדגימה שלנו** – $sampleRate$ ("אינטרוול דגימה").

בהינתן הפרטים היבשים של המשתמש, נחשב עבורו מדד Baseline שייתן לנו הערכה ראשונית של כמות השתייה היומית הבסיסית המומלצת של המשתמש:

$$WorkoutMinutesAvgPerDay = workoutWeeklyHours \cdot \frac{60}{7}$$

$$dailyBaseline(ounces) = bodyWeight(Ibs.) \cdot \frac{1}{2} + \frac{WorkoutMinutesAvgPerDay}{30} \cdot 12$$

- **הערה:** הנוסחה נלקחה **מאתר [1]**.

כעת, נוכל לחשב מדד Baseline של כמות השתייה המומלצת עבור המשתמש בכל אינטרוול דגימה:

$$numIntervals = \frac{24 - sleepHours}{sampleRate}$$

$$\Rightarrow baselineInterval = \frac{dailyBaseline}{numIntervals}\ (ounces) \underbrace{=}_{\boldsymbol{cup = 8\ ounces}} \frac{dailyBaseline}{8 * \ numIntervals}\ (235ml.cup)$$

לאחר מכן, בהתאם לנתונים הדינמיים באותו אינטרוול, נגדיר משקולת לכל סוג נתון בזה:

משקולת עבור סוג הפעילות – w_1, **משקולת עבור הטמפרטורה –** w_2, **משקולת עבור מספר הצעדים –** w_3.

המשקולות מחושבות באופן הבא –

$$\boldsymbol{w_1} = \begin{cases} 0, & rest \\ 1.1, & walk \\ 1.2, & run \end{cases} \quad ; \quad \boldsymbol{w_2} = \begin{cases} 1, & Temp \le 25^o \\ 1.1, & 25^o < Temp \le 30^o \\ 1.2, & 30^o < Temp \end{cases} \quad ; \quad \boldsymbol{w_3} = \frac{numSteps}{minPaceSteps_{Men/Women}}$$

$$\boldsymbol{s.t.}\ minPaceSteps_{Men} = \begin{cases} 720\ steps\ (\approx 6.5\ \ Km/h), & run \\ 382\ steps\ (\approx 3.5\ \ Km/h), & walk \end{cases}$$

$$minPaceSteps_{Women} = \begin{cases} 809\ steps\ (\approx 6.5\ \ Km/h), & run \\ 436\ steps\ (\approx 3.5\ \ Km/h), & walk \end{cases}$$

כמות השתייה שנמליץ למשתמש לשתות באותו אינטרוול תהיה – $(\boldsymbol{w_1} \cdot \boldsymbol{w_3} + \boldsymbol{w_2}) \cdot \boldsymbol{baselineInterval}$.

- **הערות:** קבענו את המהירות המינימלית עבור הליכה וריצה על סמך ההיגיון שלנו מחיי היום-יום. כנ"ל לגבי ערכי המשקולות. כמו כן, החישוב וההמרה ממספר צעדים למהירות התבססה על המידע **באתר [2]**. בנוגע לשילוב המשקולות בנוסחה הסופית, היא נוצרה על סמך ההיגיון שרצינו לקשר בין סוג הפעילות לבין כמות הצעדים (ולכן המשקולות מוכפלות אחת בשנייה) ולהוסיף על כך באופן בלתי תלוי את המשקל של הטמפרטורה. נשים לב כי בכל אינטרוול מתקיים – $(w_1 \cdot w_3 + w_2) \cdot baselineInterval \ge baselineInterval$.

5. **<u>חישוב כמות הקלוריות שנשרפו</u>**

להלן הנוסחה שמצאנו **באתר [3]** לחישוב כמות הקלוריות שנשרפו בזמן פעילות כלשהי –

$$calBurned = \frac{activityDuration(Mins.) \cdot MET \cdot 3.5 \cdot bodyWeight(Kg)}{200}$$

כאשר MET מהווה מדד שמשתנה כתלות בפעילות הנעשית ובאינטנסיביות שלה. במקרה שלנו, ה-MET מושפע מסוג הפעילות ומהקצב שבו היא מתבצעת. להלן הנוסחה להמרה ממספר הצעדים שנמדדו במקטע למהירות בשעה, שהתבססה גם היא על המידע **באתר [2]** –

$$speed_{Men} = \frac{numOfSteps \cdot 0.762}{1000} \cdot \frac{60}{sampleRate}\ (Km/h)$$

$$speed_{Women} = \frac{numOfSteps \cdot 0.67}{1000} \cdot \frac{60}{sampleRate}\ (Km/h)$$

לאחר קביעת המהירויות, המדד MET נגזר אוטומטית על סמך הטבלאות **שבאתרים [4], [5] ו-[6]**:

$$\boldsymbol{MET_{walk}} = \begin{cases} 3, & speed \le 4 \\ 4, & 4 < speed \le 6 \\ 5, & 6 < speed \end{cases} \quad ; \quad \boldsymbol{MET_{run}} = \begin{cases} 8.3, & speed \le 8 \\ 10.5, & 8 < speed \le 12 \\ 13, & 12 < speed \end{cases} \quad ; \quad \boldsymbol{MET_{rest}} = 1.3$$

ביבליוגרפיה

[1] How to calculate how much water you should drink (University of Missouri System)

[2] Average Stride Length Statistics: Stride Length By Height And Sex

Relevant Passage –

Many fitness pedometers and watches use a default average step length of 2.2 feet (0.67 meters) for women and 2.5 feet (0.762 meters) for men, which can be converted to 4.4 feet and 5 feet for the average stride length for women and men, respectively.

[3] Calories Burned Calculator: A Simple Way To Find Out How Many Calories You Burn Daily

Relevant Formula –

The equation for the Exercise Calories Burned Calculator is:

Duration of physical activity in minutes × (MET × 3.5 × your weight in kg) / 200 = Total calories burned.

[4] MET Table - Inactivity

[5] MET Table - Walking

[6] MET Table - Running

[7] סקר שטראוס