**תהליך פיתוח המודל ובחירת היפר-פרמטרים**

פיתוח המודל כלל צינור עיבוד נתונים רב-שלבי:

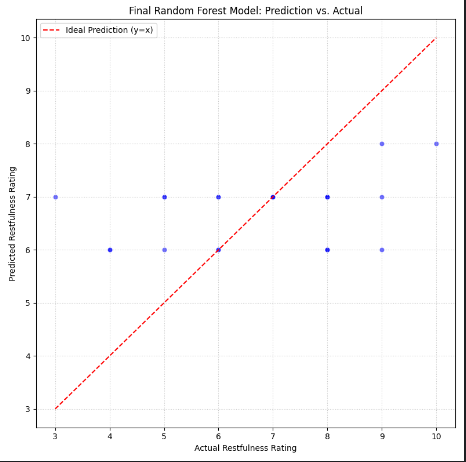
* **איסוף ועיבוד נתונים:** נטענו נתוני החיישנים ונתוני השאלונים מ-9 משתתפים משלושת הסשנים בוצע ניקוי יסודי של ערכים חסרים (NaNs) תוך שימוש במילוי ספציפי לחיישן, מילוי קדימה/אחורה, ובסוף KNNImputer ו.SimpleImputer
* **הנדסת תכונות:** חולצו 54 תכונות יומיות רלוונטיות לשינה מכל חיישן ומאפיינים בין-חיישניים.
* **פיצול נתונים ונורמליזציה:** הנתונים פוצלו למערכי אימון (80%) ובדיקה (20%) באופן פרטני לכל משתתף. בוצע נורמליזציה מסוג StandardScaler לכל משתתף בנפרד .
* **סינון תכונות (Vetting)** תהליך זה כלל טיפול בחריגים (Capping), בחירת מאגר תכונות ראשוני על בסיס Mutual Information (MI) הסרת יתירות באמצעות MI ומתאם ספירמן (Spearman correlation), ובחירה סופית של 15 תכונות באמצעות ReliefF.
* **בחירת תכונות בשיטת העטיפה (Wrapper Method)** מודל RandomForestRegressor שימש בתוך  **Exhaustive Feature Selection (EFS)** אשר בחרה באופן ממצה את 10 התכונות האופטימליות ביותר (מתוך ה-15 שעברו סינון). מדד הניקוד היה 'neg\_mean\_squared\_error' עם 5 קיפולי קרוס-ולידציה (cv=5).
* **כיוונון היפר-פרמטרים** הפרמטרים המיטביים שנמצאו: n\_estimators=100, min\_samples\_split=10, min\_samples\_leaf=2, max\_features=0.6, max\_depth=30, bootstrap=True. ציון ה-R2 הטוב ביותר בקרוס-ולידציה היה -1.0963
* .**השוואה למודלי בסיס:** השוואת ביצועי המודל הסופי למודלי בסיס חושפת את יעילותו

|  |  |  |  |
| --- | --- | --- | --- |
|  | המודל שלנו | מודל ממוצע | מודל לפי שעות שינה |
| MSE | 2.5071 | 2.9407 | 3.2645 |
| RMSE | 1.5834 | 1.7148 | 1.8068 |
| R2 | 0.1255 | -0.0257 | -0.1387 |

בהשוואה זו, מודל RANDOM FOREST הסופי מציג שיפור ברור ועקבי בכל המדדים. מקדם ה-R2 החיובי שלו מעיד על יכולתו להסביר כ-9.18% מהשונות בדירוג מנוחת השינה, וזהו יתרון משמעותי לעומת מודלי הבסיס, שערכי ה R2 השליליים שלהם מצביעים על כך שהם גרועים יותר אפילו מניבוי קבוע של הממוצע. ה MSE והRMSE הנמוכים יותר של המודל הסופי מצביעים גם הם על דיוק גבוה יותר בניבויים.

**גרפים:**

** גרף 1: ניבוי מול ערך בפועל (Prediction vs. Actual)**



* גרף פיזור המראה את הקשר בין דירוג המנוחה בפועל לבין הניבויים של המודל. הקו המקווקו האדום מייצג ניבוי אידיאלי (הערך החזוי שווה לערך בפועל). התרשים מראה שבעוד שיש התאמה כללית, קיימת נטייה של המודל לנבות ערכים נמוכים כערכים גבוהים יותר וערכים גבוהים כערכים נמוכים יותר ויש המון נקודות (רובן) לא נמצאות על הקו והנקודות שבקצוות (3 ו10) מאוד רחוקות מהקו האדום.

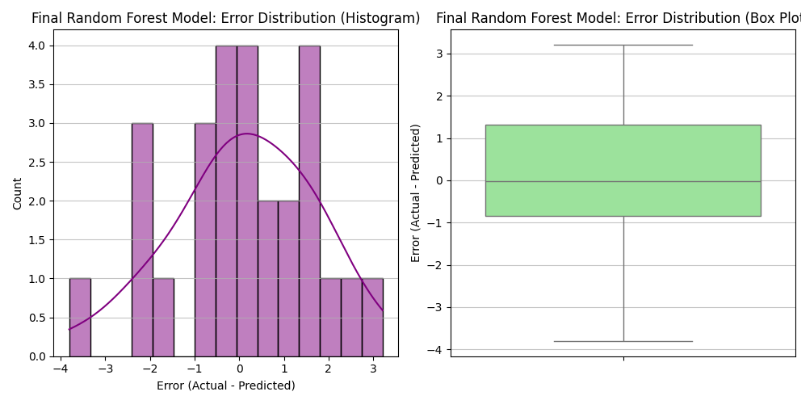
** גרף 2: חשיבות תכונות (Feature Importance)**

תמונה שמכילה טקסט, צילום מסך, קו, תרשים

תוכן בינה מלאכותית גנרטיבית עשוי להיות שגוי.

* גרף עמודות המציג את חשיבותן היחסית של 10 התכונות הסופיות שנבחרו עבור המודל. התכונה last\_call\_to\_sleep בולטת כחשובה ביותר, ואחריה תכונות הקשורות ל-Wi-Fi ופעילות שיחות.

** גרף 3: התפלגות שגיאות (Error Distribution)**



מורכב מהיסטוגרמה (משמאל) ותרשים קופסה (מימין) של שגיאות הניבוי (ערך בפועל - ערך חזוי). ההיסטוגרמה מציגה התפלגות שגיאות הממורכזת סביב אפס, אך עם זנבות המעידים על שגיאות ניבוי גדולות יחסית. תרשים הקופסה מאשש את הפיזור, עם חריגים המעידים על שגיאות מהותיות.

**ניתוח חוסן ומגבלות המודל**

* **ניתוח חוסן:**
  + **ימי חול מול סופי שבוע:** המודל פחות חזק יותר בימי חול (MSE=2.62, R2= 0.0687) מאשר בסופי שבוע (MSE=2.29, R2= 0.1906). הדבר מצביע על שגרות עקביות יותר בסופי שבוע שקל יותר למודל ללמוד.
  + **עונות שונות:** הנתונים מוגבלים לחודשים אפריל-מאי, ולכן לא ניתן להסיק לגבי חוסן המודל לשינויים עונתיים.
  + **רמות פעילות משתנות:** המודל כולל תכונות הקשורות לפעילות, אך לא נבדק באופן ספציפי על פני קטגוריות רמות פעילות מוגדרות.
* **דיון במגבלות:**
  + **אילוצי גודל מדגם** 108: נקודות נתונים בלבד (81 אימון, 27 בדיקה) הן מדגם קטן מאוד. זה מגביל את יכולת המודל ללמוד דפוסים כלליים, מעלה סיכון להתאמת יתר ומקשה על הערכה סטטיסטית מהימנה.
  + **מגבלות תכונות:** המודל מתבסס רק על נתוני סמארטפון. גורמים חשובים חסרים כוללים:
    - **נתונים פיזיולוגיים:** שלבי שינה, קצב לב, טמפרטורת גוף .
    - **מדדים סובייקטיביים נוספים:** זמן הירדמות, יעילות שינה, רעננות בבוקר.
    - **גורמים התנהגותיים/סביבתיים:** צריכת קפאין/אלכוהול, רמות רעש בחדר, טמפרטורת החדר.
    - **גורמים פסיכולוגיים/בריאותיים:** מתח, חרדה, הפרעות שינה.
    - **הבדלים אישיים:** גיל, מין.
* **הכללה:**
  + **לתקופות זמן חדשות :** המודל מציג יכולת מסוימת להכליל לימים חדשים עבור אותו פרט (בתוך תקופת הנתונים), אך ייתכן שידרדר בתקופות ארוכות יותר או בתנאי חיים שונים (כמו פער הביצועים בין ימי חול לסופי שבוע).
  + **לפרטים אחרים:** זו המגבלה המשמעותית ביותר. המודל אומן על קבוצה קטנה של 9 משתתפים. למרות נורמליזציה , דפוסי שינה וגורמי השפעה יכולים להשתנות מאוד בין אנשים. המודל הנוכחי צפוי להתקשות מאוד בהכללה לאוכלוסייה רחבה ומגוונת יותר ללא אימון על דאטה-סט גדול ורחב יותר או שימוש בגישות התאמה אישית מתקדמות.

**שיפורים פוטנציאליים:**

* **הגדלת גודל המדגם:** איסוף נתונים מכמות גדולה יותר של משתתפים ולתקופות ארוכות יותר.
* **העשרת התכונות:** שילוב נתונים ממכשירים לבישים (פיזיולוגיים), דיווחים ידניים מפורטים יותר ונתונים סביבתיים נוספים.
* **מודלים מתקדמים:** בחינת מודלים של למידה עמוקה (RNNs, LSTMs) עבור נתוני סדרות זמן מורכבים יותר.
* **התאמה אישית:** פיתוח מודלים פרטניים או מודל בסיס שמותאם אישית (fine-tuning) לכל משתמש חדש.