להוסיף הנחות יסוד

הכנה לדו"ח מסכם

1. נתון
   * 1. דוח נתנוים ממדדי רכב שונים, דוח הכולל 7 מדדים:

טמפרטורה

טמפרטורה חיצונית

לחות

זרם

מתח

רעידות

סיבובי מנוע

* + 1. דוח מתוייג לפי תקלות 0.1.2.3, כאשר 0 מוגדר פעולה תקינה.
    2. נתונים בדוח בעלי 2-6 סגמנטים המופרדים באמצעות נקודה. הנתונים לפי אמורים להיות מופרדים, סגמנט ראשון נתון עיקרי, סגמנט שני נתון משני.

Xxx.yyy.zzz.www.hhh.aaa

* + 1. XXX – סגמנט ראשון – נתון עיקרי, YYY – סגמנט שני – נתון משני.

1. \*\* הוגדרה משימה \*\*
   * 1. מציאת סף קריסה מערכת (לא הוגדר מה יהווה קריסה והושאר לשיקול הדעת)
     2. ניבוי התנהגות אשר תוביל לקריסת המערכת קודם לקריסתה.
     3. להתבסס רק על המשתשנים:

זרם Current

מתח Voltage

טמפרטורה Temperature

*הוגדר לתהעלם משאר המשתנים.*

1. פעולות שבוצעו:
   1. מבט על הנתונים:
      1. התפלגות נתוני מקור לפי תיוג התקלה

|  |  |  |  |
| --- | --- | --- | --- |
|  | Count | | |
| Fault Label | Voltage (V) | Current (A) | Temperature (Â°C) |
| 0 | 5000 | 5000 | 5000 |
| 1 | 2000 | 2000 | 2000 |
| 2 | 2000 | 2000 | 2000 |
| 3 | 2000 | 2000 | 2000 |
| Total | 11000 | 11000 | 11000 |
|  |  |  |  |

* 1. ניתוח נתונים ראשוני, מציאת ערכים אשר אינם נומריים
     1. בניתוח ראשוני של נתוני המקור אפשר לזהות שתי טעיות מהותיות אשר מחייבות מחיקת התצפיות המכילות טעות

|  |  |  |  |
| --- | --- | --- | --- |
| Fault Label | Voltage (V) | Current (A) | Temperature (Â°C) |
| 2 | 5.542.896.479.769.260 | 8,03E+10 | 24.911.848.629.758.800 |
| 2 | 5.045.887.171.680.880 | 6,22E+10 | 28.416.670.806.151.000 |

* + 1. הסיבות המצדיקה את ביטול כל התצפית ולא העברה ל N/A את הערכים הלא נומריים בזרם.
       1. משקל זניח של 2 תצפיות (עשירית האחוז) מכלל התצפיות בתיוג תקלה 2.
       2. הנתון אינו נומרי כבר בשלב הראשוני של עיבוד הנתנוים עלול להקשות על חלוקה ל main, secondary ולכן יש לוותר על שתי התצפיות.

1. יצירת קובץ על בסיס הנתונים המקוריים מתוך 10998 תצפיות שנותרו לאחר ביצוע סעיף: אשר מכיל את העמודות

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fault\_Label | voltage\_main | voltage\_secondary | current\_main | current\_secondary | temperature\_main | temperature\_secondary |

* 1. לפי:

Xxx.yyy.zzz.www.hhh.aaa

XXX – סגמנט ראשון – נתון עיקרי, YYY – סגמנט שני – נתון משני.

1. התפלגות המתקבלת לםי תיוג התקלה
   1. Count For Fault Label 0
      * 1. voltage\_main: 5000
        2. voltage\_secondary: 5000
        3. current\_main: 5000
        4. current\_secondary: 5000
        5. temperature\_main: 5000
        6. temperature\_secondary: 5000
   2. Count For Fault Label 1
      * 1. voltage\_main: 2000
        2. voltage\_secondary: 2000
        3. current\_main: 2000
        4. current\_secondary: 2000
        5. temperature\_main: 2000
        6. temperature\_secondary: 2000
   3. Count For Fault Label 2
      * 1. voltage\_main: 1998
        2. voltage\_secondary: 1998
        3. current\_main: 1998
        4. current\_secondary: 1998
        5. temperature\_main: 1998
        6. temperature\_secondary: 1998
   4. Count For Fault Label 3
      * 1. voltage\_main: 2000
        2. voltage\_secondary: 2000
        3. current\_main: 2000
        4. current\_secondary: 2000
        5. temperature\_main: 2000
        6. temperature\_secondary: 2000
   5. זו בדיקה שמאשרת ששלמות הנתונים מתקיימת בעיבוד הנתנוים (בניכוי סעיף 3.b.i)
2. לפי המשימה, סעיף 2.i בהיעדר הגדרת הלקוח/יצרן מהי קריסה וכן הוגדר בסעיף 1.ii כי תיוג Fault\_Label, בערך 0 מהווה פעילות תקינה.

הנתונים יחולקו למצב בינארי לפי תיוג Fault\_Lable, תקינים עם ערך 0 ב Fault\_Label וערך גדול מ 0 וקטן ושווה ל 3 ב Fault Label . תצפיות המכילות מידע על תקלה (כלשהי, ללא ידיעה או הגדרה מה רמת הקריטיות שלה) ותצפיות להתנהגות תקינה.

* + 1. יצירת שני קבצי נתונים: מתוייגים כתקלה 0 ומתוייגים כ 1,2,3. כל קובץ לא מכיל את עמודת Fault label, תיוג התקלה.

Fault\_0\_main\_secondary\_Data\_Set.prod.csv

Count: 5000

Fault\_1\_to\_3\_main\_secondary\_Data\_Set.prod.csv

Count: 5998

שלמות הנתנונים נשמרה.

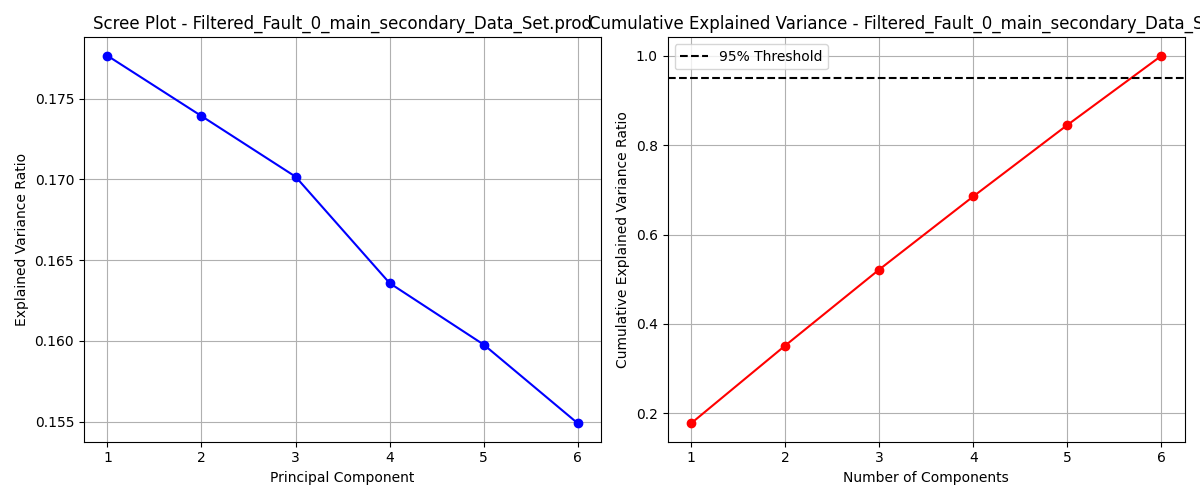
1. תיאור נתונים מתוייגים תקינים:
2. Voltage\_main:
   1. Count: 5000
   2. Mean: 94.12
   3. Std: 258.48
   4. Min: 7.00
   5. Max: 999.00
3. Voltage\_secondary:
   1. Count: 5000
   2. Mean: 542.49
   3. Std: 285.52
   4. Min: 0.00
   5. Max: 999.00
4. Current\_main:
   1. Count: 5000
   2. Mean: 62.18
   3. Std: 179.79
   4. Min: 1.00
   5. Max: 994.00
5. Current\_secondary:
   1. Count: 5000
   2. Mean: 501.16
   3. Std: 288.91
   4. Min: 0.00
   5. Max: 999.00
6. Temperature\_main:
   1. Count: 5000
   2. Mean: 39.20
   3. Std: 114.02
   4. Min: 1.00
   5. Max: 981.00
7. Temperature\_secondary:
   1. Count: 5000
   2. Mean: 489.74
   3. Std: 283.90
   4. Min: 0.00
   5. Max: 999.00
8. האנומליות שהוגדרו:
   1. Current Main > 250
   2. Temperature > 250
   3. Voltage > 850
      * 1. ניסיון למלא את הנתנוים האנומליים בנתונים מנורמלים באמצעות SVD הניב תוצאות בעיתיות בעקבות בדיקת MSE לדגומא
           1. Current measurements: Main: 34561.928010 Secondary: 0.000000 Temperature measurements: Main: 13199.103741 Secondary: 0.000000
           2. לכן התצתפיות המכילות את האנומליות 14a,14b,14c נמחקו מהנתונים המעובדים.
9. הנתונים ללא האנומליות :
   1. נתונים המתוייגים כלא תקינים:
   2. \*\*\* נתונים\*\*\*
10. ניתוח נתונים
    1. בדיקת לינאריות של הנתונים המתוייגים כשגויים

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Voltage\_main | Voltage\_secondary | Current\_main | Current\_secondary | Temperature\_main | Temperature\_secondary |
| Voltage\_main | 1 | -0.019 | 0.023 | 0.015 | -0.007 | 0.004 |
| Voltage\_secondary | -0.019 | 1 | 0.019 | -0.011 | 0.007 | -0.011 |
| Current\_main | 0.023 | 0.019 | 1 | 0.005 | -0.008 | -0.004 |
| Current\_secondary | 0.015 | -0.011 | 0.005 | 1 | 0.025 | -0.003 |
| Temperature\_main | -0.007 | 0.007 | -0.008 | 0.025 | 1 | 0.005 |
| Temperature\_secondary | 0.004 | -0.011 | -0.004 | -0.003 | 0.005 | 1 |

* 1. בדיקת לינאריות של הנתונים המתוייגים כשגויים

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Voltage\_main | Voltage\_secondary | Current\_main | Current\_secondary | Temperature\_main | Temperature\_secondary |
| Voltage\_main | 1 | -0.032 | -0.011 | -0.014 | -0.002 | 0 |
| Voltage\_secondary | -0.032 | 1 | 0.04 | -0.001 | 0.011 | -0.035 |
| Current\_main | -0.011 | 0.04 | 1 | -0.003 | 0.017 | 0.004 |
| Current\_secondary | -0.014 | -0.001 | -0.003 | 1 | -0.015 | 0.028 |
| Temperature\_main | -0.002 | 0.011 | 0.017 | -0.015 | 1 | 0.035 |
| Temperature\_secondary | 0 | -0.035 | 0.004 | 0.028 | 0.035 | 1 |

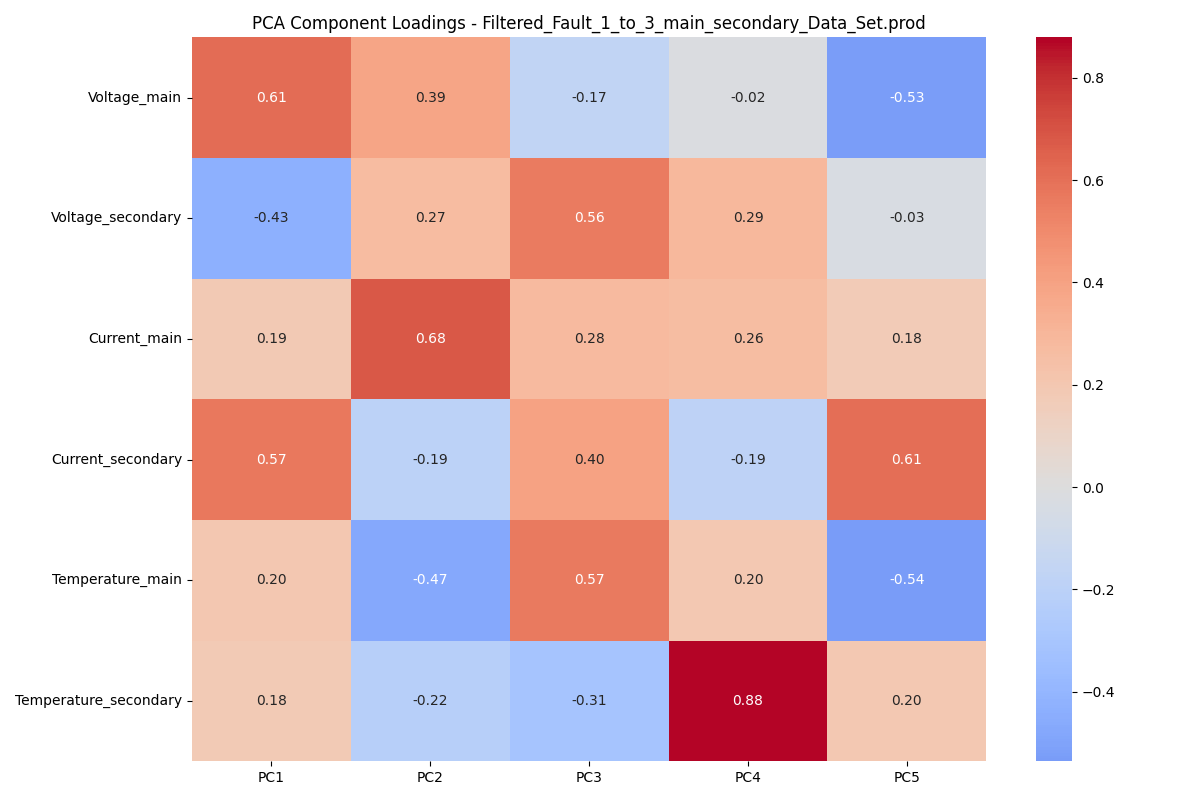
* 1. מהנתונים עולה כי אין קשרים מובהקים בין המשתנים ואם הקשרים האלה קיימים הם חבויים. ההחלטה לבדוק קשרים נועדה לזיהוי משתנים אשר באמצעותם אפשר יהיה לנבא את נקודות הקריסה.
  2. מניתוח PCA עולה כי אין קבוצת משתנים אחת שמסבירה את רוב השונות.
     1. עבור נתונים המתוייגים כתקינים



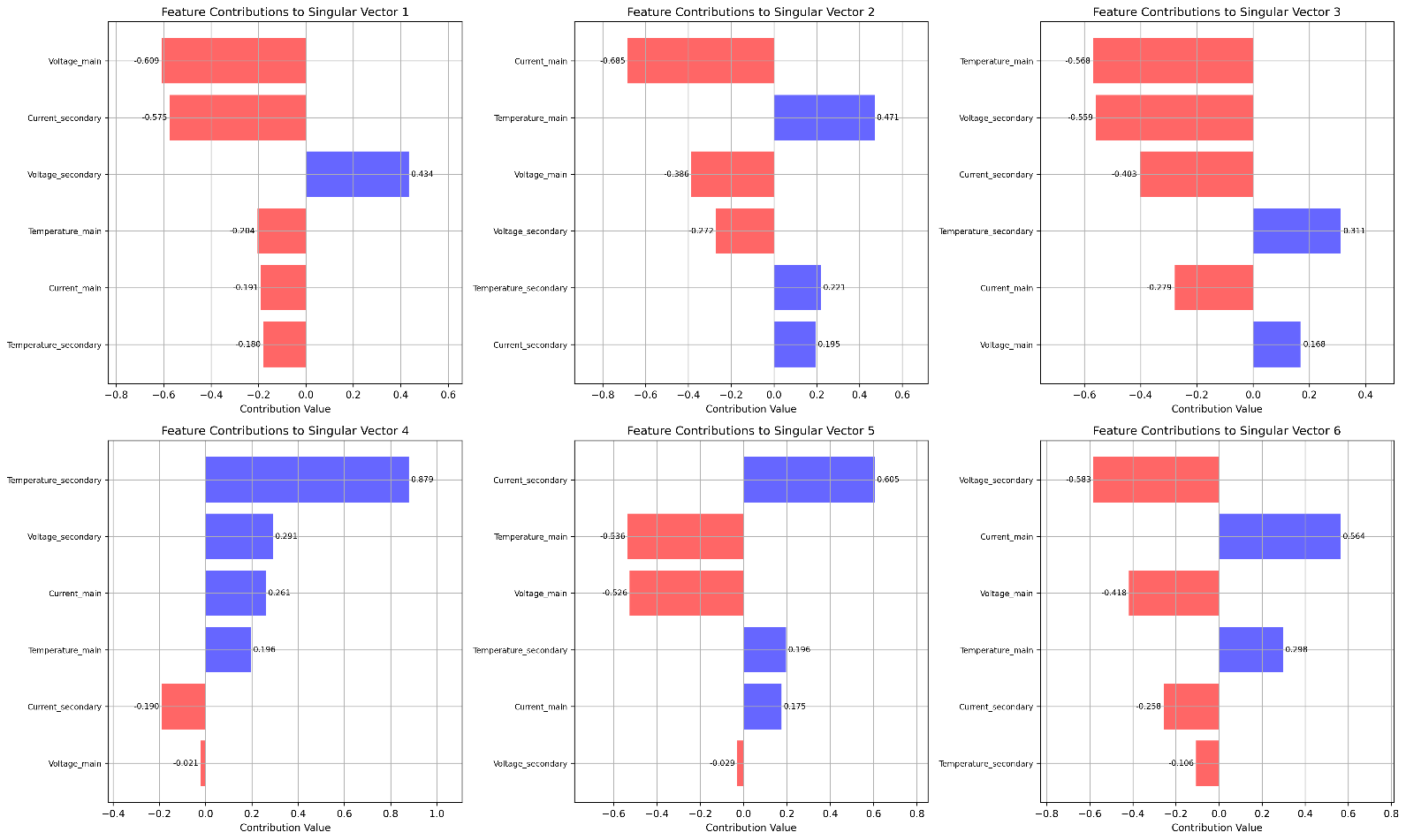
* + 1. עבור נתונים המתוייגים כלא תקינים – תקלה A graph of a line and a line

       AI-generated content may be incorrect.
    2. מהנתונים עולה כי אפשר להצביע על קשרים מורכבים בים המשתנים וצורך במספר רב של משתנים להסביר את רוב השונות

|  |  |
| --- | --- |
| Principal Component | Explained Variance (%) |
| PC1 | 17.228377937125153 |
| PC2 | 17.089712580032337 |
| PC3 | 17.01015354282037 |
| PC4 | 16.612006254832316 |
| PC5 | 16.096554482422437 |
| PC6 | 15.9632 |



* + 1. זה מתחזק עם ניתוח SVD



* + 1. בהיעדר קביעה חד ערכית באמצעות PCA ו SVD ועדיים יש לקבוע את נקודות הכשל, נוחלט לבחור בטבלת שכיחויות למציאת נקודות כשל.
    2. מניתוח השכיחויות אפשר לזהות
    3. \*\*\*\*\* נתונים \*\*\*\*
    4. נקודות הכשל הן:
       - 1. \*\*\*\*נתונים\*\*\*
         2. \*\*\*\*נתונים\*\*\*
         3. \*\*\*\*נתונים\*\*\*
         4. \*\*\*\*נתונים\*\*\*
         5. \*\*\*\*נתונים\*\*\*
         6. \*\*\*\*נתונים\*\*\*
         7. \*\*\*\*נתונים\*\*\*
    5. בהינתן הנחת היסוד שמדובר בתצפיות מבוססות סמן (כל תצפית היא נקודת זמן בודדת)
    6. בהינתן שהנתונים אינם לינאריים והנתונים הינם סדרת זמן, אזי מודל prophet הוא המתאים לחיזוי הגעה לנקודת הקריסה. הסיבות: יכולת עבודה עם נתנוים לא לניאריים, מהירות.
       1. לכל תצפית בכל סט נתנוים (תקיניים ולא תקיניים) יתווסף משתנה timestamp אשר ישתנה בערך של שניה אחת מתצפית לתצפית.