



הפקולטה להנדסה

פרויקט גמר - תחזית גשם

במסגרת הקורס "מבוא למדידת מכונה"

המרצה : מר' דור בנק

מתרגל : מר' יעד טובלי

מגיש : אביתר וייס 308481324

Evyatarweiss1@gmail.com

תאריך : 16.7.20

הצהרת הבעיה

גשימים הם חלק חיוני מחיינו. השירות המטאורולוגי הישראלי ריא יחידת סמך במשרד בתפקידו האחראי לחיזוי מגזין האוויר. תפקיד זה הינו חשוב ביותר, לאחר ולחיזוי מגזין האוויר ישנן השלכות כלכליות רבות בנושאים מגוונים מעבר לאיך עליינו להתלבש. מכאן שבפרויקט זה, אני אישם תאוריות ומודלים שלמדנו במהלך הקורס באמצעות Python-Pandas ו- Scikit-Learn, ובננה מסוג בצדלו לחיזוי אם יורד גשם מחר או לא לפי נתונים מסוימים.

מבנה העבודה כולל ארבעה מרכיבים עיקריים: חקר וניתוח נתונים, עיבוד נתונים, הטעמת מודלים והערכת מודלים.

תקציר

בפרויקט זה השתמשתי במודלים רבים ללמידה מכונה על מנת לחיזות האם ירד גשם או לא זאת לצד סקירה וחקירה מעמיקה של הנתונים ובחירה המודלים הנכונים לנواتים בשביל תחיתות בעלת AUC מקסימלי.

חקר וניתוח נתונים

חשיבות לציין כי את חקירת הנתונים התחלנו עם 26 פיצרים, חלקם עם שם ידוע וחלקם אונוניים. אם כן הפיצרים מתחולקים לקטגוריאליים וחקלם משתנים רציפים. מבדיקה מעמיקה נראית כי (Feature_14) הינו רציף אך מתנהג כקטגוריאלי וכן אליו מוצמדתו המדידות MM. למרות היותו פיצר בעל שם אונוני החלטתי לשנות אותו להוריד את mm מכל איבר ולהפוך אותו לרציף בغال הווה קשור למידה של כמה מים. באופן זה גם החלטה שירוטית להפוך את פיצר מס' 9 למספר שכך הוא בבניי מספרים אך מתויג כקטגוריאלי. בהמשך במחקר עמוק יותר על הדאטא גליתוי שি�ינו קשור לא לינארי בין פיצר 13 ל-14. ככלומר יצרתי עמודה חדשה ולה עשית פונקציה למבדה (נספח 3). הפונקציה הנ"ל מראה כי האיבר בפיצר 13 הוא 1 אם ורק אם האיבר באותה שורה בעמודה של פיצר 14 הוא גדול מ-1, ואם קטן מ-1 הוא 0. מה שמאושש את זה הוא שלשיהם יש את אותם מספר ערכיים חסרים. לכן ככל הנראה פיצר 13 הוא מניפולציה מתמטית של 14 מפה שהחלטתי **להסיר** את פיצר מס' 14.

ביצעתי בדיקה של כמה האיברים החסרים בכל עמודה, נראה כי לעמודה sunshine ישן 1871 איברים חסרים שמהווים 8.4% מכלל העמודה, מכאן המספר לא גבוה נבחר בכך זה לא להסיר אף עמודה שכן היא מביאה יותר תועלות מאשר חוסר נתונים.

משתנים רציפים וקטגוריאליים

מעתה והלאה בוצעה הפרדה מלאה בין משתנים רציפים לקטגוריאליים על מנת לבצע ניתוחים שונים בהן. בראיה למרחוק לרגע בו נדרש להשלים את הנתונים החסרים למשתנים הקטגוריאליים ביצעתו ויזואлизציה לכל פיצר שכזה שמרתא את הה帖פלגות שלו וכמה מתוכו מתויג בלייבל 1 וכמה 0. ככלומר מתויר זה ניתן להסיק את החישובות שלו על החיזוי, דוגמה על פיצר 19 בנספח ב' שנראה שתוצאות הגשם מתפלגת אצל בהתאם לתדרות האיבר. באופן זה עשית על המשתנים הרציפים. שם ניתן לראות בבירור שלפיכר 13 (נספח ה'), ולפיכר sunshin'e (נספח ד') ישן עמודות שבחן יננה נציגות גדולה של לייבל 1, ומפה שגם מעיד על טבות על חשיבות הפיצר לחיזוי.

שימוש נתונים

לאחר חקר הנתונים נוכחנו לדעת כי ישנו לא מעט נתונים חסרים בהרבה עמודות. אם כן עלינו להשלים אותם בצורה מסוימת. לפי התפלגות הליבל ברוב הפיצרים הנוכחיים אין חשיבות לבחירת ערך ספציפי כמו הממוצע או החציון שכן זה לא יתרום אלא רק יירות את המאזן. מפה שבחורתי שלהשלים את הנתונים החסרים ע"יvr ששם קיבלו ערך רנדומלי מתוך הערכים המקוריים. כМОון של משתנה רציף יש טוח של ערכים אינסופי ולמשתנה קטגוריאלי יש מספר סופי של אופציות.

מטריצת קורלציה

כרגע אנחנו עומדים על 25 פיצרים שכן בנספח וניתן לראות את מטריצת הקורלציה שמסוגת לפי צבעים שכן צבע כהה מעיד על קורלציה שלילית וככה על חיובית. מטריצה זו ניתן לראות בבירור כי קיימים קשרים ליניאריים שבוגדים 0.80 למשל עם feature_0 1 and evaporation.

בשbill לראות את הקשרים הליניאריים בצורה ויזואלית ביצעת pairplot (נספח 2). כדי שהמודל יהיה יציב מספיק, השונות צריכה להיות נמוכה. וכן קורלציה של שני פיצרים מגדילה את השונות שמספר בחרתי להסיר מסpter פיצרים (Year,0,1,16,17,11) (Year,0,1,16,17,11).

איברים חריגים

במבט מהיר, BOXPLOT מספקת אינדיקציות לסימטריה בתוך הנתונים. ככלומר כל יהות חריגים באמצעות השיטה זו שכן חריגים שכאלו יכולים לפגוע קשות בתוצאות הניבוי. (נספח ח') לפי הנספח ניתן לראות כי ישנה כמות גדולה של חריגים במסpter פיצרים שכן גנich והתפלגות הינה נורמלית, וכן נסיר אותם באמצעות שיטת Z-SCORE השרה של החריגים משמעותה, מסירים את השורות שבהם קיימים החריגים. אםvr בנספח ט' ניתן לראות את ה *boxplot* לאחר ההסרה.

Dummy variable

משתנה זה הוא משתנה מלאכותי שנוצר לייצג פיצר קטגוריאלי באמצעות מספרים. ככלומר זה תלוי במסpter האיברים הייחודיים שיש לכל פיצר קטגוריאלי. על מנת שהמחשב יוכל לנבא באמצעות פיצרים קטגוריאליים ולהסיק מהם דברים, עלי לפצל אותם לפיצרים נפרדים באמצעות פונקציה זו.

Data scaling

כפי שראינו הדטה שלנו מכיל פיצרים רבים מסוגים שונים וכן הכליל חריגים רבים (עכשו פחות). מאחר ואלגוריתמים רבים בלמידת מכונה משתמשים במרקח אוקלידי זו בעיה מאחר והם מזניחים את כל עניין היחידות כלומר פיצרים עם טוח גדול יותר יחס משמעותי יותר לפי המרחק האוקלידי. ומפה שבעמצעות Min_max_scaling הפכתי את כל הנתונים להיות בטוח של 0-1 והפרופורציות נשמרות.

בחירה פיצרים – קלילת הממדים

לאחר פונקציית dummy variable יש בידנו 96 פיצרים מה שambilו אותנו לדבר על קלילת הממדים. ריבוי הנתונים שמתרכז כאשר עברנו למיד (מספר פיצרים) גבוהה יותר, משפיע על נפח המרחב המוצע. ככלומר עם עלייה בממדים הנתונים ממלאים פחות ופחות את מרחב. וזה מחזיר אותו לעיה שציינתי ב data scaling, שאלגוריתמים משתמשים במרקח אוקלידי, ומפה שיחסוב של מרחב למספר עולה ומעלה של פיצרים

הופל להיות לא יעיל גם אל מבחינת כוח חישוב וגם מבחינות תוצאות בפועל זה גורם **over fitting**. ניתן לו הות מדיות גבוהה בהבדל AUC בין-train וה-test, אם ההבדל ביןיהם גדול ככל more הרבה יותר אז יש **over fitting**.

אני שילבתי בין 3 שיטות להקטנת הממדיות: PCA, Sequential Forward Selection ו-KBEST.

בעצם עבדתי על 2 DATA SETS במקביל לראות מה הוא השילוב הטוב ביותר.

1. בהתחלה ביצעת את select_kbest, בבחירה ההיפר פרמטרים בחרתי להוריד את הממדיות מ95 ל20 פיצרים בדיעה שאני הולך לבצע אתPCA.

ניסיתי מספר score-function ביןיהם f_regression וchi2, הטוב מביניהם היה f_regression ואת על כך שהביא ליותר שונות מסוימת בביצוע PCA לאחריו וגם תוצאות טובות בהרבה במודלים שאח"כ.

כפי שאמרתי לאחר ביצירת ה20 פיצרים ביצעת PCA עם 4 קומפוננטות שכן זאת גם מסיבה שבדקתי מהו מספר הקומפוננטות הטוב ביותר בהתאם לתוצאות המודלים.

2. DATASET נוסף הוא באמצעות "Sequential Forward Selection," שכן בהיפר-פרמטרים בחרתי גם להוריד את הממדיות ל20 פיצרים.

הטמעת מודלים

*בכל אחד מהמודלים אציג את ההיפר-פרמטרים הנבחרים עבור כל אחד מהsets-Data. SFFS.1 PCA+KBEST.2

לאחר מכן אציג את ה-AUC במודל הספציפי עם הפרמטרים.

מודלים ראשוניים:

Naïve Bayes Classifier – זה מודל ללא בחירת היפר-פרמטרים.

```
Model AUC score - sffs with: 0.8253430
Model AUC score - pca with: 0.8236270
```

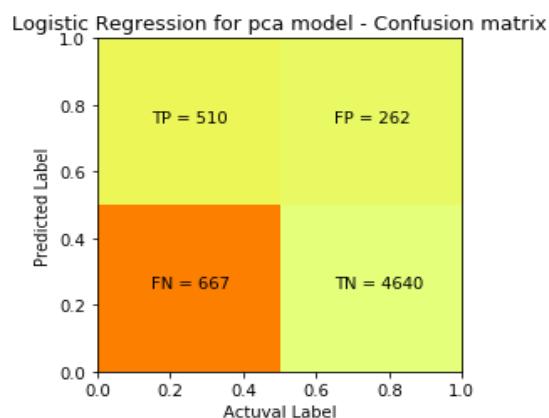
K-Nearest-Neighbors – בבחירה ההיפר פרמטרים ביצעת חיפוש חמדני מקיים על הפרמטרים הטוביים ביותר עבור כל אחד מהdataset-סט מהסעיף הקודם. שמנתי דגש עבור $k=1$ שאומר שמדובר במרחב מנהטן ולא אוקלידי. עבור וקטורים בממדים גבוהים מנהטן לרוב עובד טוב יותר מהמרחב האוקלידי.

```
Model AUC score of feature selection model pca: 0.845513
leafsize = 9,n_neighbors:70,p = 1
Model AUC score of feature selection model sffs: 0.843634
leafsize 50,n_neighbors:50,p = 1
```

Logistic Regression- מאחר מדובר בממדיות גבוהה יחסית גם אחרי regularization, בחרתי לעשות גריד-סרג' עם העדפה ל l_1 penalty בזזה לטיב בסעיף הקודם וגם C נמוך שיהווה רגולרציה חזקה כדי למנוע overfitting אבל לא מספיק בשבי למנוע למידה.

ניסيون נוסף היה לנשות thresholds LR בغالן חסר האיזון של הליבלים אבל זה הערך הדיפולטיבי 0.5 היה הטוב ביותר.

```
Model AUC score - sffs with hyperparameter C 0.1 : 0.86765930
Model AUC score - pca with hyperparameter C0.1 : 0.84563615
```



-Confusion matrix

במקרה הנ"ל מדובר על CM שמתיחס למודל רגסיה לוגיסטי בתיחס לdataset שuber מניפולציה של PCA. True positive (TP): אלה מקרים שניבאנו שכן ירד גשם ואכן בפועל ירד גשם.

True Negative (TN): ניבאנו שלא ירד גשם ובפועל לא ירד גשם.

False positive (FP): חיזנו שירד גשם אבל לא ירד גשם.

False negative (FN): חיזנו שלא ירד גשם אבל בפועל ירד גשם.

ישנם חישובים רבים וחשיבות רבה לCM, למשל חקלאי מסוים מכשה את הצמחה שלו אם יירד גשם, ואם קרה FN קלומר חשב שלא ירד גשם אבל ירד זה פגע לו בצמחה. קלומר אפשר לכונן מודל שייעדיף לחזות שירד גשם ויטעה לפעמים קלומר יותר FP מאשר FN.

מודלים מתקדמים:

Multi-Layer Perceptron (ANN) – הדבר החשוב ביותר היה למנוע overfitting ולכך ביצעתו גריד-סרג עם ערכי אלף שונים, וכך גם טווח ערכים לכמות hidden layers.

```
hidden_layer_size = 20 alpha = 0.1 solver = "sgd",max_iter=200
ההיפר-פרמטרים יצאו זהים לשני הדאטה-סטס.
```

```
Model AUC pca test score: 0.8441970697029665
Model AUC sffs test score: 0.86520907492893
```

Support Vectors Machine – חיפוש מקוף של היפר – פרמטרים. החל מ C הולראזיה חלשה או חזקה, סוגים שונים של "קופסאות כסם" – קרנלים וAGMA שבעצם סותר את הרגולרייזציה.

```
Model pca with kernel sigmoid and gamma=0.1 and C=1 AUC: 0.84500
Model sffs with kernel rbf and gamma 0.01 and C 100 AUC: 0.86840
Random Forest –igan עשית לולאות ובחרתי את ההיפר-פרמטרים שנוטנים את ה AUC הגבוה ביותר.
```

```
Model AUC score of feature selection model sffs: 0.8721382
n_estimators = 30, max_features = auto, max_depth = 20 ,min_samples_split = 10 ,min_samples_leaf = 4 ,bootstrap = False
```

הערכת המודלים

עבור כל dataset, הרצתי את כל ששת המודלים שהטמעתי מוקדם, כל אחד עם ההיפר-פרמטרים המתאימים לו. וכן לכל מודל ישנו גרפ' שבתוכו מוצגת עקומת ROC עבור כל fold (10), אלו נמצאים במחברת.

לאחר שלIORק כל הפרויקט רצתי עם שני DS במקביל בכך לדעת מי מבוא הכי טוב, בחרתי ב-set של SFFS זאת כי הוא הראה יציבות AUC גבוהה יותר מהדעתה סט של PCA. כתע עלי לבחור מודל שאיתו אזהה את נתוני מבחן, לאור תוצאות המודלים גם בשלב ההטמעה וגם בשלב הערכה בחרתי במודל Random Forest.

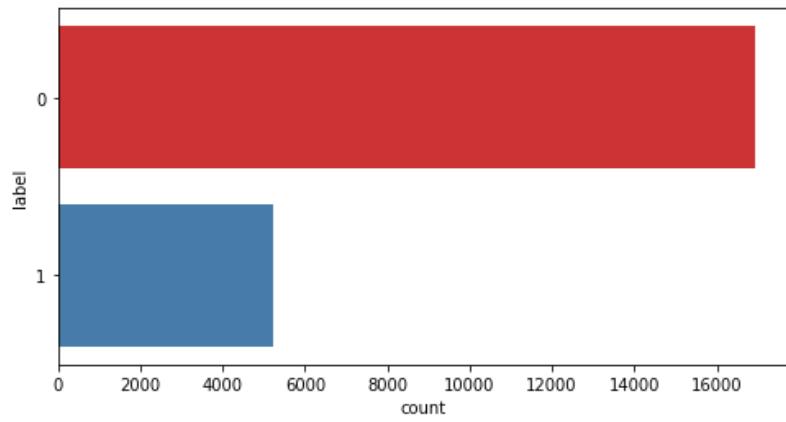
בכדי לוודא שאין overfitting ביצעתי למדיה על דатаה מפוצל ובדקתי את הישג המודל עליו וגם על הטסט. ההפרשים לא היו גדולים ומפה שאין overfitting בלבד העובדה шибועי המודל על test זהים להטמעת המודלים. על מנת להגדיל את יכולת ההכללה בחרתי היפר-פרמטרים בשלב ההכללה שימנו מצב זהה.

סיכום

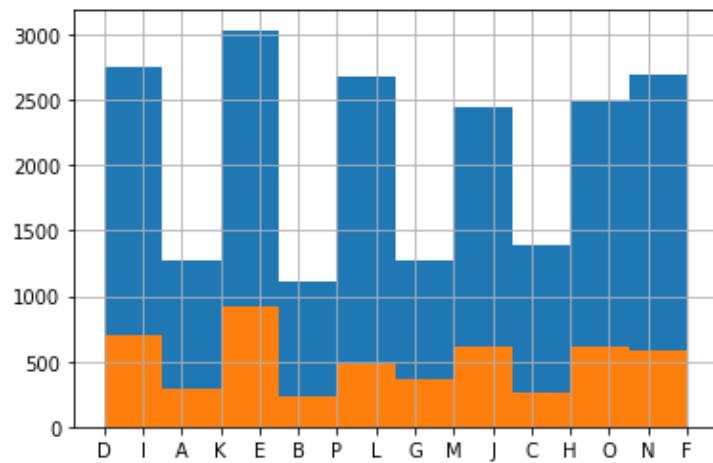
אם כך לסיום אסקור את כל מה שערכנו עלי, בהתחלה קיבלנו דатаה עם המונע רעים, בצענו מניפולציות על הפיצרים בכך להתאים אותם לקריאת מחשב, הסכנו מסקנות על הדטה ועל חשיבותן כאליה ואחרות של פיצרים.

িחרתי שני Data Sets באחד דילתי את המימדיות באמצעות PCA+ KBEST והשני באמצעות SFFS . לאחר מכן בchneriy שישה מודלים שונים על כל אחד מDS ובכל אחד מהם חיפשתי את ההיפר-פרמטרים האידיאליים לו בהתאם ל-DS וה- AUC הגבוה ביותר. לאחר מכן בצעתי 10-fold-cross-validation, ולמדתי את המודלים על כלל הדטה בפולדים השונים. ראייתי לנכון שהמודל הטוב ביותר לביצוע הפרדיקציה הוא Random Forest ובאמצעותו חיזיתי את סט test.

נספח א'



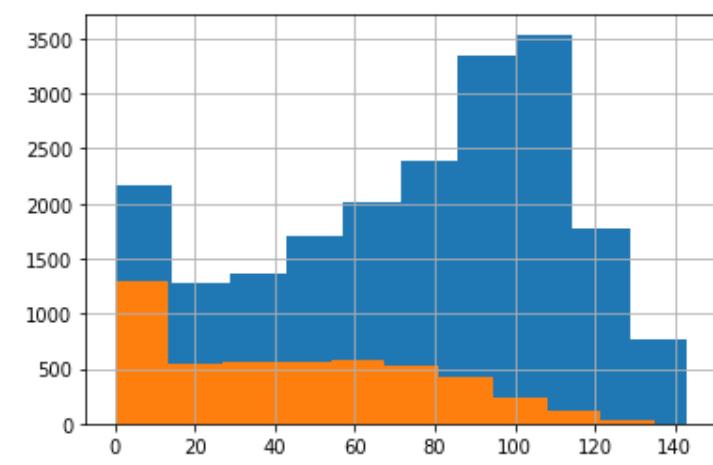
נספח ב'



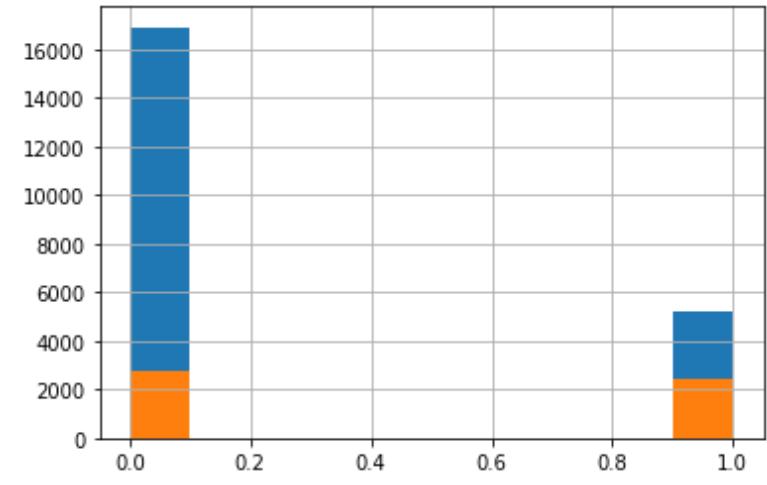
נספח ג'

```
df['check'] = df.Feature_14.apply(lambda x: "1" if x > 1 else ("0" if x <= 1 else np.nan )) df['check'].equals(df['Feature_13'])#
```

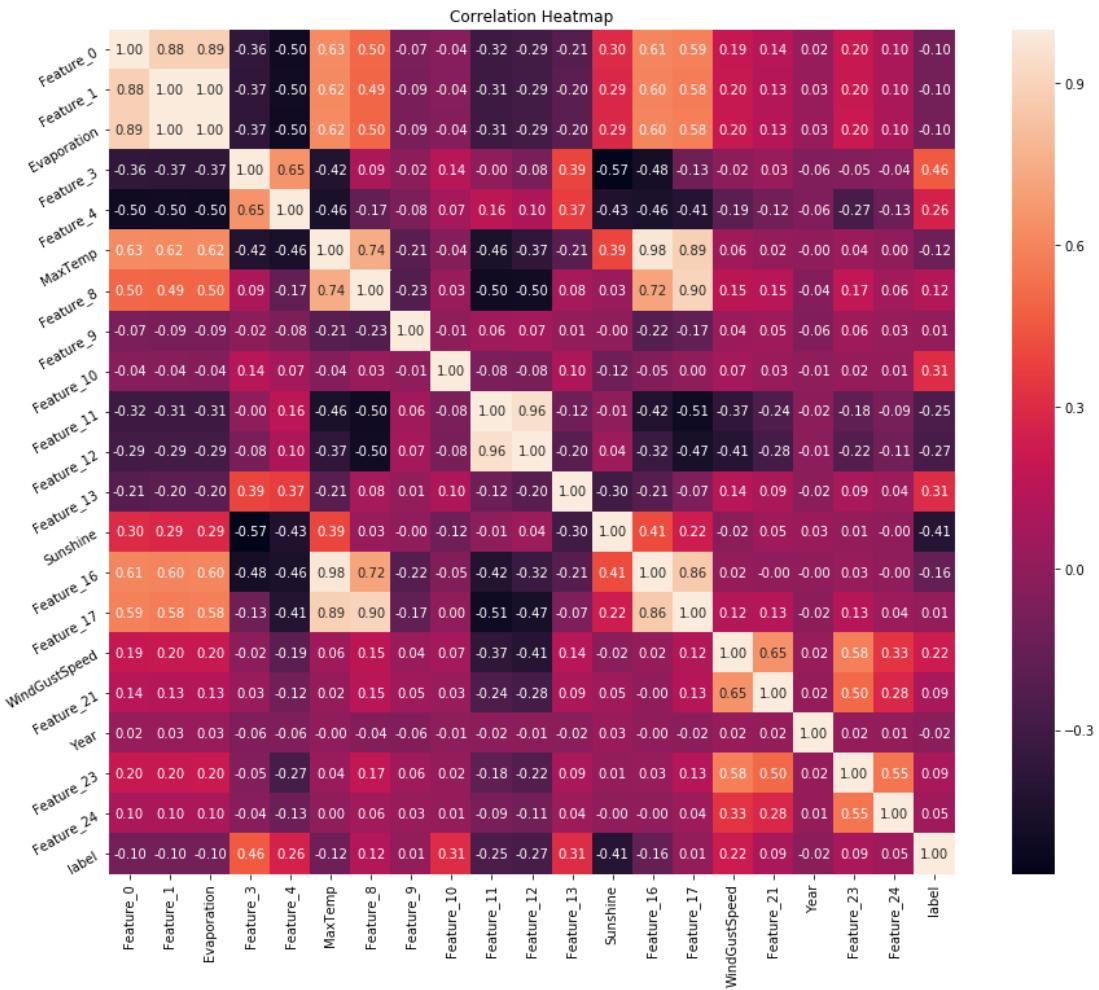
נספח ד'



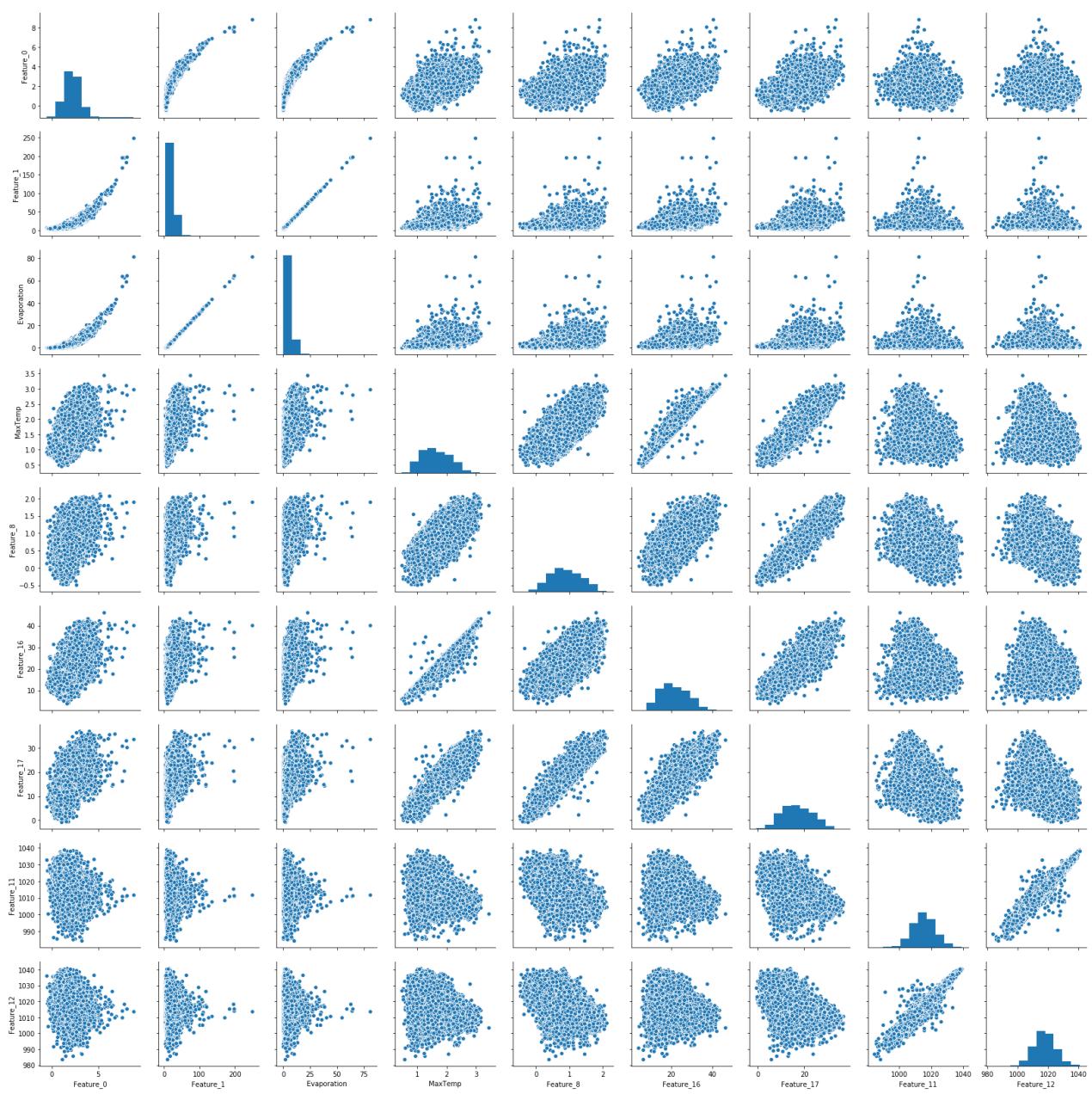
נספח ה'



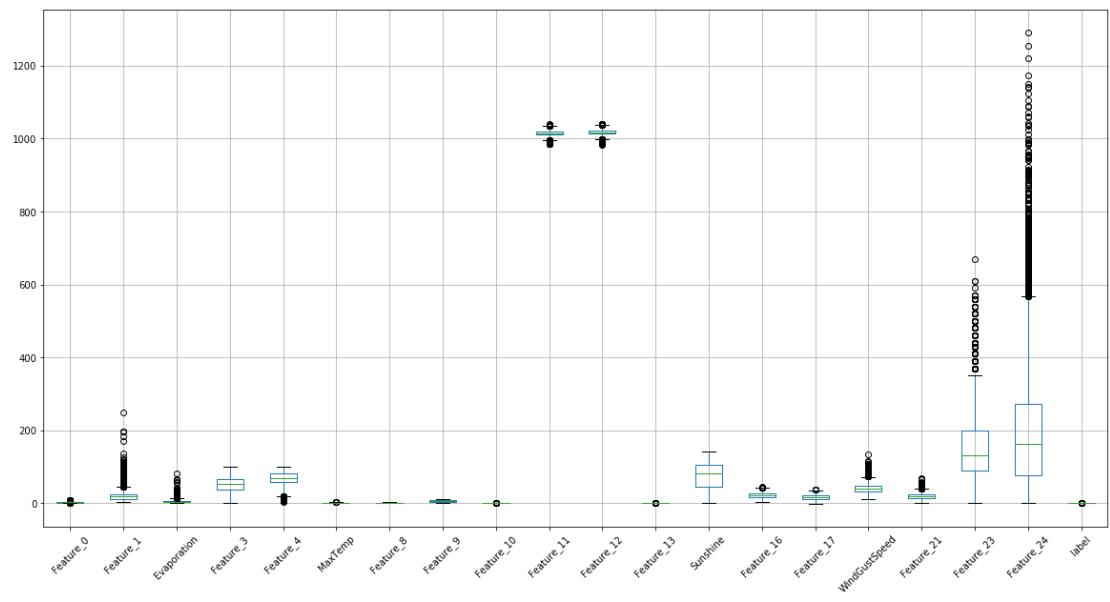
נספח ו'



נספח 1:



נספח ח'



נספח נ'

