מגישה מירב דיין 203317052

# דו"ח פרויקט

## תקציר:

מילאנו ערכים ריקים נומאריים ע"י שימוש בממוצע ועבור משתנים קטגוריאליים ע"י הערך השכיח. המרנו משתנים קטגוריאליים בבינאריים ע"י שימוש ב- dummy features. זיהינו ערכי outliers בשיטת IQR (טווח בין רבעוני) והורדנו בשורות המכילות מעל 5 ערכי outliers. ע"י שימוש בפונקציית קורולציה וויזואליה איתרנו משתנים קורלטיביים, ולפי הקורולציה של כל משתנה עם הלייבל, נפרדנו מהמשתנים המיותרים. בנוסף זיהינו את הפיצ'רים שבקורלציה הכי גבוהה עם הלייבל וייצרנו מהם פיצ'ר חדש. ע"י שימוש בפונקציית KBest ולאחר מכן PCA נותרנו עם דאטה המכיל 15 פיצ'רים. אימנו מודלים של ANN, KNN, Linear Regression ו- Random Forest. את ערכי ההיפר-פרמטרים, בחרנו בצורה חמדנית ובשימוש בKFOLD. כדי לוודא שהמודל אינו overfit בדקנו את ההפרש בין הAUC של הואלידציה לשל קבוצת האימון. המודל שהשיג את ערך הAUC הגבוה ביותר היה ANN (0.88) ולכן אימנו אותו על כל הtrain וביצענו בעזרתו פרדיקציה על הtest.

## מהלך הפרויקט:

ראשית ע"י תצפית בדאטה זיהיתי משתנים שדורשים טיפול, כגון משתנה 14 ו-6 מהם החסרנו את הסטרינגים כיוון שהיו לא נחוצים. לאחר מכן, בעזרת שימוש בכלים שונים של ויזואליזציה קיבלתי תמונה מצב לגבי התפלגות הפיצ'רים השונים, מה שאפשר לקבל החלטות לגבי טיפול בערכים חסרים. כיוון שלא היה משתנה נקודתי עבורו אחוז הערכים החסרים היה משמעותי, בחרתי להשלים את הערכים הקטגוריאלים ע"י הערך השכיח ביותר, ועבור משתנים רציפים השתמשנו בהשלמה על ידי הערך הממוצע.
לאחר מכן טיפלתי במשתנים קטגוריאליים. על מנת שניתן יהיה לעבוד איתם יש צורך להמיר אותם לבינאריים, לכן השתמשתי בפונקציית הdummy features. כיוון שההתפלגות של כל משתנה קטגוריאלי הייתה יחסית אחידה לא יכולנו לשלב ערכים שונים ובכך לצמצם את כמות הפיצ'רים. בשלב הבא, בעזרת שימוש בויזואליזציה של מטריצת הקורלציה, יכולתי לאתר קבוצות של פיצ'רים מתואמים, כאשר ערך הסף למשתנים קורלטיביים שהגדרתי היה 0.85. הקבוצות בהם הבחנתי מפונקציה זו היו:

group1 = ['MaxTemp', 'Feature_17', 'Feature_16', 'Feature_8']

group2 = ['Feature_11','Feature_12']

group3 = ['Evaporation', 'Feature_1', 'Feature_0']

בנוסף הבחנתי בקשר בין משתנה 13 למשתנה 14, כאשר עיינתי בערכים. נראה כי כאשר הערך במשתנה 14 גבוה מ-1 הערך במשתנה 13 הוא 1 אחרת 0.

על מנת לבחור, איזה משתנה מכל קבוצה להשאיר בדקתי למי יש את הקורלציה הכי גבוהה עם הלייבל ואת השאר הסרתי.

על מנת שהדאטה יהיה בסקאלה אחידה, ביצעתי סטנדרטיזציה של הערכים, צעד החשוב בעיקר לשלבים הבאים כמו הPCA כשנרצה לאמוד מרחקים.

בעזרת שימוש בויזואליזציה של boxplot וע"י שימוש בIQR זיהיתי ערכים שנמצאים במרחק IQR * 1.5, כלומר ערכי קיצון. בהתחלה ניסיתי להוריד את כל ערכי הקיצון הקיימים אך צעד כזה פגע בביצועי המודל והשאיר כמות יחסית מצומצמת של ערכים לכן החלטתי במקום לבדוק כמה ערכי קיצון מכילה כל תצפית, כאשר את התצפיות שהכילו מעל ל5 ערכים קיצוניים הורדנו.

בשלב זה, ניסיתי לייצר פיצ'רים חדשים ע"י מניפולציה של הקיימים. לשם כך מדדתי את הקורלציה של המשתנים עם הלייבל כאשר הערכים הגבוהים ביותר היו Sunshine ו- Feature_3 הפיצ'ר החדש היה המכפלה של שניהם.

בשלב הבא ניסיתי לצמצם את כמות הפיצ'רים, כיוון שכמות גבוהה יכולה להוביל לoverfit של המודל ובנוסף בעלת ערך חישובי כבד.

תחילה השתמשתי בפונקציה SelectKBest בעזרתי שמרתי על 25 הפיצ'רים המובילים ועל אותם פיצ'רים השתמשתי בPCA עם 15 קומפוננטות ששימרו 78% מהשונות.

עד כה היה השלב של העיבוד המקדים של הדאטה. בשלב הבא בחרתי 4 מודלים (במקור היו יותר אך נעזרתי באלו שהיו בעלי תוצאות טובות יותר וזמן חישוב קצר יותר).

המודלים שנבחרו היו: ANN, KNN, Random Forest וLinear Regression.

עבור כל מודל חיפשתי את ההיפר-פרמטרים שימקסמו את הביצועים שלו. על מנת למדוד זאת השתמשתי בפונקציה של KFOLD ועבור כל חזרה ניסיתי לאתר את ההיפר-פרמטרים, בשיטה החמדנית (ומהירה יותר), שימקסמו את ערך ה- AUC. הפרמטרים שנבחרו היו של הFold בעל ערך הAUC הגבוה ביותר.

עם אותם היפר-פרמטרים ביצענו שוב KFold (K=10), ויצרנו ROC Curve עבור כל Fold וחישבנו את הAUC הממוצע של המודל.

סקירת ההיפר-פרמטרים שנבחרו עבור כל מודל וה-AUC הממוצע:

<u>עבור Logistic regression:</u>

LogisticRegression(max_iter=50, solver='newton-cg', C=0.001, penalty='l2')

0.85 -AUC

<u>עבור KNN (K Nearest Neighbors):</u>

KNeighborsClassifier(weights='distance', n_neighbors=100, algorithm='auto', leaf_size=5, p=1)

0.86 -AUC

עבור Multi-Layer Perceptron (ANN):

MLPClassifier(max_iter=250, solver='adam', hidden_layer_sizes=50, activation='logistic', alpha=1e-07)

0.88 -AUC

עבור Random Forest:

RandomForestClassifier(criterion='entropy', n_estimators=161, max_features='log2', max_depth=13, min_samples_split=10, min_samples_leaf=10, bootstrap=False, random_state=1)

0.87 -AUC

כיוון שערך הAUC המקסימלי היה של הANN המשכנו עם מודל זה וביצענו לו גם confusion matrix. מאותם נתונים ניתן לחשב FPR ו- TPR שכאמור מהווים את הצירים לעקומת הROC אותה הראנו עבור כל מודל ומובילים לחישוב הAUC.

על מנת להעריך האם המודלים שלנו סובלים מ- overfit, על אף שעקומות הROC יחסית זהות עבור כל Fold, החלטנו גם לבדוק את המודל על סט אימון לעומת ואלידציה ולבדוק את ההפרשים ביניהם. פיצלנו את הדאטה בהתאם כאשר 30% ממנו הפך לואלידציה ואימנו את המודל על סט האימון. עבור כל מודל התקבל ההפרש הבא:

-ANN

delta=0.02851366807039102

-KNN

delta=0.15479823404479287

הפרש זה יחסית גבוה, אך יתכן ויש קשר לכך שבניגוד לclassifiers אחרים מודל זה אינו מתאמן על התוצאות אלא משווה אותם לקבוצת האימון ולכן כשנעשה פרדיקציה על קבוצת האימון עצמה הערכים שנקבל יהיו גבוהים (הם יושוו לעצמם). בכל אופן לא נמשיך עם מודל זה.

-Random Forest

delta=0.1157544505898328

-Logistic regression

delta=0.017972327055935988

כלומר ניתן להסיק שהמודלים הללו אינם overfit.

כאמור המודל שבחרנו להמשיך לעבוד עימו ולבצע בעזרתו את הפרדיקציה על הtest היה ANN. לפני שביצענו את הפרדיקציה ביצענו fit של המודל בפעם על כל סט הדאטה השלם ולאחר מכן בדקנו את הסיכוי לקבל 1 עבור כל אחת מדוגמאות הtest. אותן פרדיקציות הועתקו לקובץ אקסל המצורף.

**סיכום:**

הדאטה שקיבלנו עבר עיבוד מקדים במסגרתו טיפלנו בערכים חסרים בהתאם לסוג המשתנים, ערכי קיצון, יצרנו פיצ'רים חדשים בעזרת שימוש במשתנים שהיו במתאם הכי גבוה עם הlabels וגם השתמשנו בשיטות שונות כמו PCA על מנת להוריד את כמות המימדים של הדאטה אל סכום סופי של 15 פיצ'רים כאשר סט הtest עבר עיבוד תואם (פרט לערכי הקיצון).

על הדאטה הנקי בחנו 4 סוגי מודלים: ANN, KNN, logistic regression ו- random forest ומצאנו את קבוצת ההיפר-פרמטרים שמיקסמו את ביצועי המודלים. ביצענו תהליכים שונים להערכת כל מודל כגון K fold cross validation שבסופם בחרנו להמשיך עם מודל הANN שכן ביצועיו היו הטובים ביותר. וידאנו כי המודלים שלנו אינם overfit כאשר השוונו את הAUC של האימון לשל הואלידציה וראינו שהוא נמוך. לסיום אימנו את המודל של ANN על פני כל סט האימון שניתן לנו ובוצעו פרדיקציות על סט הtest.