**בחירת פיצ'רים:**

הקוד מכיל מספר אופציות לסינון פיצ'רים וניקוי של הדאטה.

**def** what\_type(X)

הפונקציה בודקת מה הסוג של הדאטה על מנת ליצור סקריפט אחיד לכל סוגי הדאטה.(תומך בdataframe ובnparray)

def norm\_data(X,method='zscore')

הפונקציה מנרמלת את הדאטה לפי השיטה שנבחרה (כרגע zscore ו minmax ), ברירת המחדל zscore.

**def** clean\_data(X,PercentMean=25,PercentStd=20)

הפונקציה מנקה את X מפיצ'רים שנמצאים ב PercentMean וב PercentStd הנמוך ביותר, הערכים 25 ו 20 הם ערכי ברירת מחדל ויכולים להשתנות על ידי המשתמש.

**def** Fdr\_anova(X, y, alpha)

הפונקציה מחזירה את X הממוין לפי ערכי p-value (אנובה) ושכלול של **בנימיני והוכבר *p(i)<=a\*i/n.*** ה a נקבעת על ידי המשתמש.

לקריאה נוספת על אנובה <https://he.wikipedia.org/wiki/%D7%A0%D7%99%D7%AA%D7%95%D7%97_%D7%A9%D7%95%D7%A0%D7%95%D7%AA>

**def** kBest\_anova(X, y, k)

הפונקציה מחזירה את K הפיצ'רים בעלי הציון הגבוה ביותר לפי אנובה (X ממוין מהגבוה לנמוך). K נקבע על ידי המשתמש.

**def** mutual\_feature(X, y, num\_neighbors, threshold)

הפונקציה בודקת תלות בין פיצ'ר למטרה (Y) ונותנת ציון נמוך לפיצ'רים בעלי תלות נמוכה, הפונקציה מחזירה את הפיצ'רים בעלי תלות גבוהה מ threshold הנקבע על ידי המשתמש (X חוזר ממוין).

**def** Collinear\_Features(fs,corr\_threshold)

הפונקציה משתמשת בספריה של בחירת פיצ'רים <https://github.com/WillKoehrsen/feature-selector>

הפונקציה עושה קורלציות לכל הפיצ'רים ופיצ'רים בעלי קורלציה גבוהה מ corr\_threshold היא מוחקת את אחד הפיצ'רים.

**def** Zero\_importance\_Features(fs)

הפונקציה משתמשת בספריה של בחירת פיצ'רים <https://github.com/WillKoehrsen/feature-selector>

הפונקציה מוצאת (בצורה לא דטרמינסטית) אילו פיצ'רים הם ללא חשיבות כלל לניבוי Y ומורידה את הפיצ'רים הללו מ X.

**def** Low\_importance\_Features(fs,cumulative\_importance)

הפונקציה משתמשת בספריה של בחירת פיצ'רים <https://github.com/WillKoehrsen/feature-selector>

הפונקציה מוצאת (בצורה לא דטרמינסטית) את הפיצ'רים הפחות חשובים שאינם נדרשים להשגת cumulative\_importance% מכלל החשיבות לניבוי Y ומורידה את הפיצ'רים הללו מ X.

**def** Select\_Features(X, y, alpha=0.05, k=0, num\_neighbors=0, threshold=0,isnormal=False, isclean=False,corr\_threshold=1,cumulative\_importance=0)

הפונקציה הראשית שקוראת לפונקציות שהוזכרו למעלה בהתאם להוראת המשתמש.

* כאשר ניתן K>0 וגם k קטן ממספר הפיצ'רים אחרי ניקוי הדאטה נקרא לפונקציה kBest\_anova אחרת נקרא לפונקציה Fdr\_anova.
* כאשר num\_neighbors>0 ו threshold>0 נקרא לפונקציה mutual\_feature לאחר קריאה לאחת מהפונקציות בסעיף הקודם.
* כאשר isnormal=False נקרא לפונקציה norm\_data
* כאשר isclean=False נקרא לפונקציה clean\_data
* כאשר corr\_threshold !=1 הפונקציה שולחת ל Collinear\_Features
* כאשר cumulative\_importance !=0 הפונקציה שולחת ל Low\_importance\_Features ו Zero\_importance\_Features

הפונקציה תחזיר X נקי וללא הפיצ'רים שבחרנו להוריד. (שימו לב ששיניתי שהפונקציה הזאת תחזיר רק משתנה אחד)

שיניתי את כל הפונקציות כך של פונקציה שתרוץ תדפיס כמה פיצ'רים היא הורידה ומה היו המשתנים (שנקבעים על ידי המשתמש)