את תהליך בחירת התכונות ביצענו תחילה ע"י מעבר נוסף על כלל התכונות הקיימות לנו כעת ביחד עם טבלת קורלציה,כדי לנסות לנתח ולמצוא יחסים מסוימים בין תכונות שבסט האימון.

בשלב זה וידאנו שכלל תכונות המטרה הן כעת תכונות עם ערכים מספריים של 1 ו 0 במקום מחרוזת,וכן ביצענו התאמה של תכונות אחרות לקבל ערכים מספריים במקום מחרוזות כמו: מין,zipAdd,pcr_date,state

ביצענו המרה של ערכי התכונות zipAdd,state,pcr_date לערכים מספריים בעזרת פונקציות המרה מן הטיפוס נוכחי שלהם לערך נומרי,ההמרה נעשתה באופן חד חד ערכי לכל ערך ייחודי.

לאחר מכן החלטנו לזרוק את התכונות הבאות:

Address - המידע הקריטי מתכונה זו כבר חולץ לתכונות חדשות ולכן אין צורך בתכונה זו יותר.

symptoms - מאותה סיבה בדיוק כמו address.

patient_id - הגענו למסקנה שאנו לא זקוקים לתכונה משום שתכונה זו אינה מוסיפה שום מידע נוסף לסט האימון,ניתן לראות זאת לפי ואנו מעוניינים שהמודל שלנו יהיה גנרי ולא תלוי בזהות הנדגם.

PCR_06 - קיימת קורלציה של 0.91 בין תכונה זו לבין **PCR_05** ולכן אין צורך בשתיהן (לא מוסיפה מידע חדש לסט אימון).

street - נמצא קושי להמיר את התכונה מן הטיפוס הנוכחי (מחרוזת), לטיפוס מספרי שבעזרתו נוכל להשיג מידע איכותי. מהטיפוס הנוכחי (מחרוזת).קושי זה נובע בעיקר מן העובדה שלתכונה זו ערכים יחודיים רבים מה שמקשה על ניתוח ערכיהם המספריים והעמדתם על סקאלה של גרף לניתוח.

zipcode - גם לתכונה זו מעל ל 2000 ערכים שונים, עוד בסעיפים הקודמים לא הצלחנו להעמיד את כלל הערכים בגרף כך שניתן לנתחו בצורה איכותית,סיבה זו גרמה לנו להחליט לזרוק תכונה זו ולנסות למפות איזורים בעיקר בעזרת התכונה zipAdd ,זאת משום שתכונה זו מכילה את מדינת המגורים ו 3 ספרות המיקוד של הנדגם, זהו מידע אינפורמטיבי בהרבה ממיקוד לבדו.

state - לתכונה זו אמנם מספר קטן של ערכים,אך לא הצלחנו לקבל מתכונה זו מידע אינפורמטיבי שיוכל לעזור לנו בקלסיפיקציה.

job - לא הצלחנו להפיק מידע מתכונה זו, ככל הנראה משום שבדומה לzipcode,street,patient_id,גם לתכונה זו מספר ערכים ייחודיים גבוה מאוד,דבר ההופך את המשימה לנתח גרפים שתכונה זו מעורבת בהם לבלתי אפשרית.

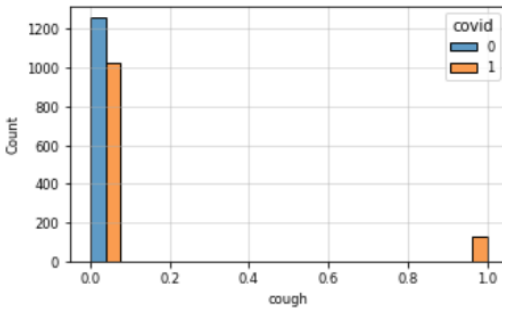
Current location - ניסינו להתייחס לקווי האורך והרוחב כקואורדינטות x ו y ולנסות לבדוק תלות בין תכונות אלה לשאר התכונות שברשותנו - תכונות אלה אינן בתלות עם אף תכונה אחרת ולא סיפקו מידע מעניין אחר. כמו כן התכונה הנ"ל לשעצמה ובצורתה (מחרוזת) אינה מספקת גם היא מידע מעניין ולכן בחרנו לזרוק אותה,ואת תת התכונות שניסינו לחלץ ממנה.

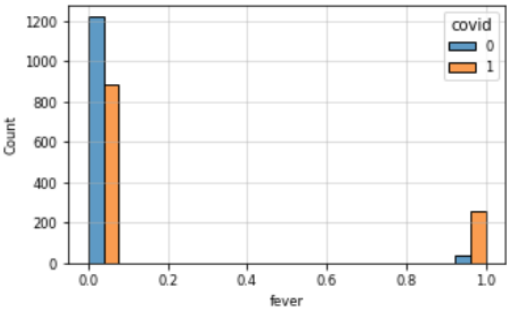
אחרי שלב ניפוי התכונות הראשוני הראשוני,נעזרנו בהמלצות הסעיף וביצענו הרצה של histplot על התכונות שנותרו לפי ערכי hue שונים (ולא בהכרח רק ערכי המטרה).

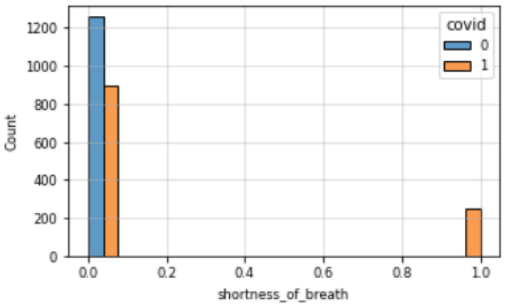
ניתן לראות בעזרת גרפים אלה כי תופעות הלוואי: שיעול,קוצר נשימה,ושפעת מגדילות משמעותית את הסיכוי שהדגימה אכן חולה בקורונה (על חלק מהתסמינים ההסתברות לחלות בקורונה בהינתן אחד התסמינים הוא 1). עבור תופעות הלוואי כאב ראש וחוסר תיאבון נמצא קשר חלש מאוד אם בכלל (לפי הגרפים ניתן לראות שתסמינים כאלה אינם מגדילים את הסיכוי לחלות בקורונה).

להלן גרפים שמאששים טענות אלה(בוצע שימוש בdodge):

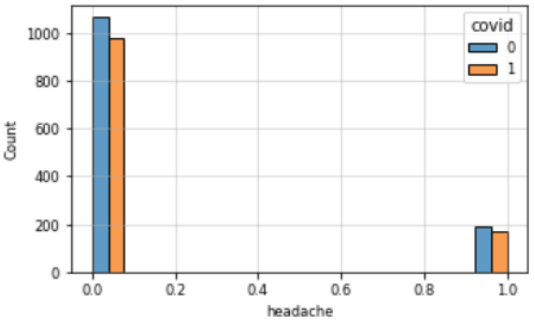
מימין לשמאל, ההיסטוגרמות של קוצר נשימה, שפעת,שיעול.

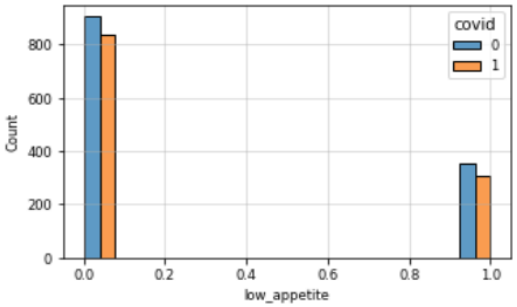






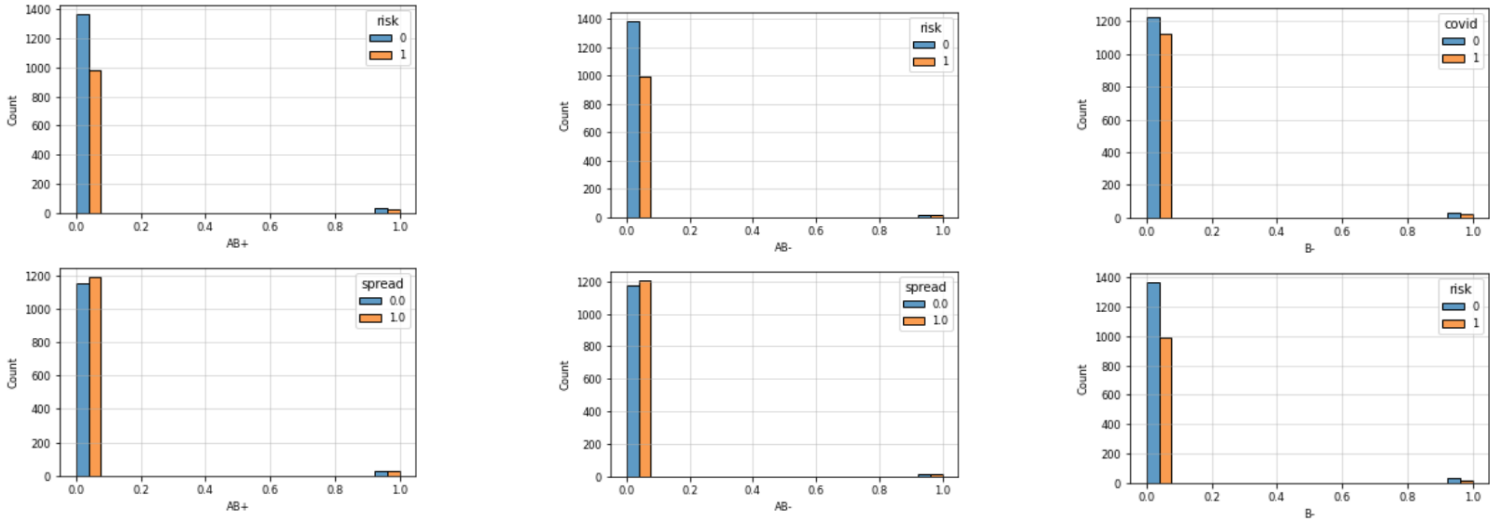
מימין לשמאל, ההיסטוגרמות של כאבי ראש וחוסר תיאבון:





ביצענו 2 בדיקות דומות עבור תכונות אלה כאשר hue = risk/spread .
ב 2 הבדיקות ,ההיסטוגרמות שקיבלנו עבור כל התכונות היו דומות מאוד לממצאים שהצגנו בעמוד הקודם עבור התכונות כאבי ראש,וחוסר תיאבון.ניתן להסיק מתוצאות אלה שאף אחת מן התכונות אינה תורמת מידע נוסף כאשר מדובר בלייבל המטרה risk or spread.
מן הממצאים של בדיקות אלה השתכנענו שהתכונות headache,low_appetite ניתנות להשמטה משום שהן לא עוזרות בסיווג של אף אחת מלייבלי המטרה,הן אינן תורמות מידע נוסף על אף שלכאורה הן אמורות להוות תסמין לקורונה, למשל בהינתן שלדוגמא מסויימת יש כאב ראש ההסתברות שהיא חולה בקורונה אינה שונה (ואף גדולה באופן שולי מאוד) מן ההסתברות שלדוגמא בלי כאב ראש יש קורונה.

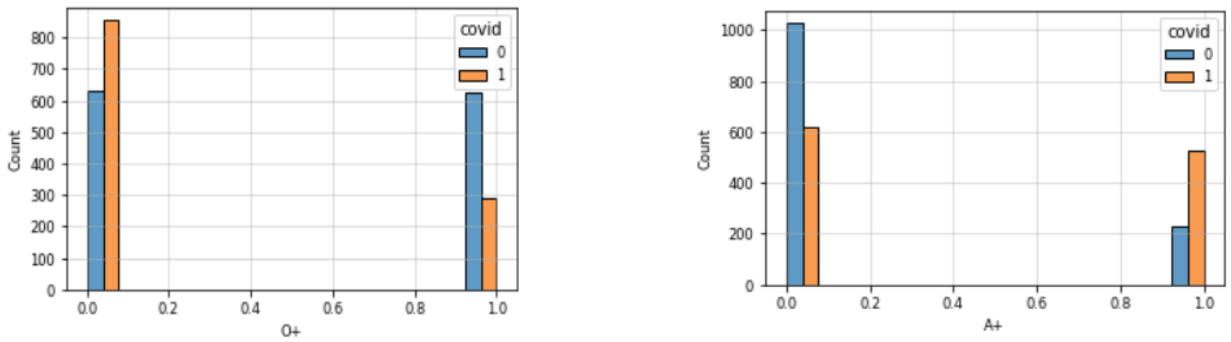
הבחנה נוספת שראינו בעת הרצת ההיסטוגרמות היא שבמערכת קיימים שלושה סוגי הדם : B- , AB+ , AB- שרק למספר מועט מאוד מכלל דוגמאות סט האימון קיים אחד משלושת סוגי דם אלה.
מן ההיסטוגרמות שקיבלנו עבור סוגי דם אלה ניתן לראות שתכונות אלה אינן עוזרות בסיווג של לייבלי המטרה.
אישוש לטענות אלה ניתן לראות לפי ההיסטוגרמות הבאות:



כמו כן, ניסינו לראות לבדוק האם ניתן למצוא קשר כלשהו בין תכונות אלה לבין תכונות אחרות שגילינו את חשיבותן הרבה כמו:שפעת או שיעול ,אך גם ניסיון זה נכשל,ולכן החלטנו לזרוק את תכונות אלה.

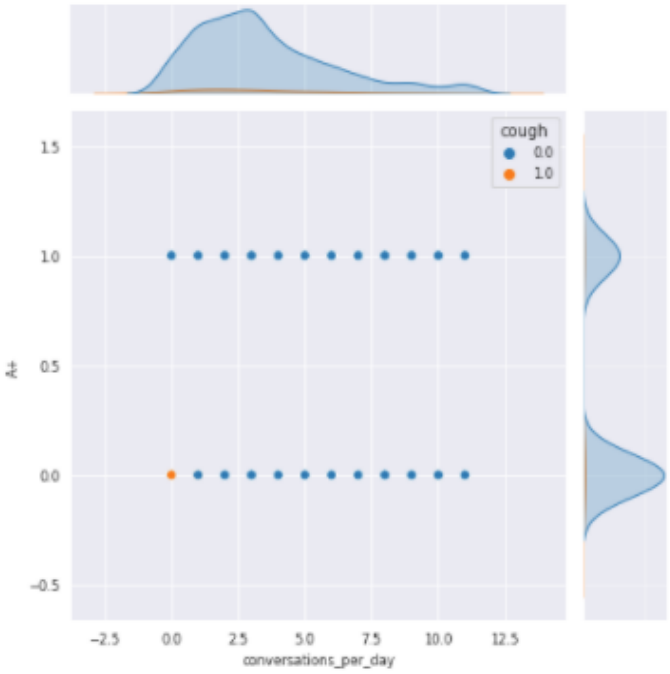
בנוסף גילינו שדוגמאות עם סוג דם O+ נוטות לחלות פחות בקורונה לעומת שאר סוגי הדם וכן דוגמאות עם סוג דם A+ נוטות לחלות יותר לעומת שאר סוגי הדם.

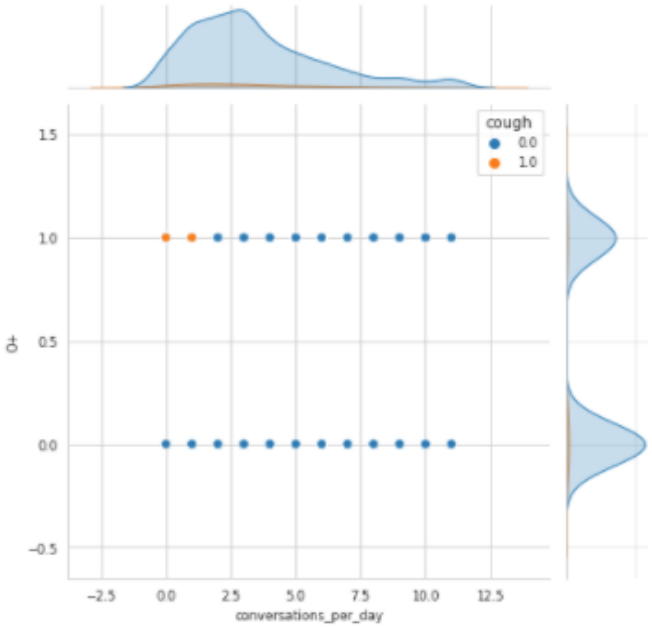
ניתן לראות זאת לפי הגרפים הבאים:



במהלך הרצת jointplots ראינו בגרפים בין מס' שיחות ליום לבין סוגי הדם O +A/-A /+ שמי שנוטה לבצע פחות שיחות ליום נוטה לפתח תסמינים של המחלה,לפי מה שראינו בעמוד הקודם קשר זה יכול לעזור לנו בתהליך החיזוי שכן בתסמינים מסוימים ל 100 אחוז מדוגמאות האימון הלוקות בתסמינים אלה יש את המחלה.

גרפים להמחשה:





קשרים דומים עם סוגי הדם הנ''ל מצאנו גם עם התכונה של מספר פעולות הספורט בשבוע. נציין שנמצאו קשרים מסויימים גם עם happiness_score - אך עם תכונה זו במידה מועטה,ופחות מספקת משל 2 התכונות הראשונות,ממצאים אלה חיזקו את ההערכה שלנו שיש להוריד את תכונה זו משום דלות המידע שתכונה זו מוסיפה למודל ואיכותו,ואכן בחרנו לזרוק תכונה זו..
החלטנו להוריד גם את התכונות: household_income,sex,.
הסיבה לכך היא שמן הגרפים השונים שהרצנו (jointplot,histplot) ראינו שחלוקת המידע לפי הערכים השונים של תכונות אלה אינה מוסיפה לנו מידע באף אחת מלייבלי המטרה. כמו כן, ניסינו לבדוק האם ניתן למצוא קשר כלשהו בין תכונות אלה לבין תכונות אחרות שגילינו את חשיבותן הרבה כמו:שפעת או שיעול ,אך גם ניסיון זה נכשל.
בנוגע לתכונות אורדינליות שהומרו לערכים מספריים חח''ע כמו zipAdd, pcr_date נראה שלא ניתן לנתח מידע איכותי מתכונות אלה לאחר ההמרה בעזרת גרפים,וגם בעץ ההחלטות הייתה להן חשיבות נמוכה מאוד אם בכלל ,ולכן זרקנו גם את 2 תכונות אלה.

בשלב הבא ביצענו יצירה של מספר עצי החלטה עם סיווגי מטרה שונים,ועומקים שונים.גילינו מהר מאוד שסוג הדם B+ מהווה תכונה חסרת חשיבות בכל העצים(בעזרת המתודה feature_importance של העץ),החלטנו להוריד תכונה זו על סמך מסקנה זו והגרפים של תכונה זו בשלב ההיסטוגרמות.מאותם שיקולים בדיוק הורדנו גם את O-.

במהלך הרצות העצים גילינו שלכלל בדיקות הPCR יש חשיבות משמעותית בעץ ההחלטות לפחות לפי מסווג אחד,מלבד PCR_02 שחשיבותה חלשה יחסית לפי כל קבוצות המטרה שהרצנו. חשוב לציין שגם התסמינים הראו ביצועים חלשים יחסית בעצים השונים ולכן יש לקחת בערבון מוגבל מסקנות מעצים אלה(כי ראינו תוצאות אחרות בשלב ההיסטוגרמות). בראייה גדולה ,נראה ש PCR_02 היא תכונה חיונית,למשל בסעיף 21 אנחנו רואים שלא ניתן לסווג באופן יעיל את risk על ידי ויתור עליה ,אבל כן ניתן לסווג בעזרתה ובעזרת PCR_01 את לייבל מטרה זה עם שגיאה אמפירית סבירה על ידי חישוב מרחק מהראשית של הגרף המצויין בשאלה זו. על סמך כלל המסקנות משאלה זו בחרנו להשאיר את כל בדיקות ה PCR איתן הגענו לשלב זה.

ראינו שלתכונות משקל,סוכר בדם ,וגיל יש משמעות בינונית-חלשה בשקלול כולל של העצים שבנינו,נראה שתכונות אלה בראייה כוללת מוסיפות מידע שיכול להיות יעיל לחידוי,לשתי תכונות אל קורלציה בינונית-גבוהה עם תכונת הגיל אנו סבורים שיש להשאיר את שלוש תכונות אלה שכן התלות ביניהן יכולה לסייע לחידוי ואולי להוות מאין מדד אלטרנטיבי ל BMI.

עוד ראינו בשלב זה שהתכונה של מספר השיחות ביום הראתה ביצועים חלשים מאוד (מדד חשיבות נמוך בכלל העצים),כמו כן משום שתכונה זו בתיאום הנמוך ביותר עם שאר התכונות,החלטנו לזרוק את תכונה זו.

התכונה של מספר האחים הראתה ביצועים בינוניים -חזקים באופן יחסי בשלב זה ולכן החלטנו להשאיר אותה.

מספר פעולות הספורט בשבוע הראתה ביצועים חלשים-בינוניים בשלב זה אך יש תיאום שלילי מתון בינה לבין משקל,סוכר בדם,וגיל, ביחד עם כלל מסקנות סעיף זה בחרנו להשאיר את תכונה זו.

את כלל התכונות שהגיעו לשלב זה ולא ציינו אותן בחרנו להשאיר על סמך כלל הביצועים שלהם בכל הבדיקות שבוצעו בסעיף זה.

feature name	keep	new	explanation
patient_id	X	X	patient_id is not relevant to any other feature,and gives no additional information.
age	V	X	there's a higher probability for certain age groups to be classified as TRUE/FALSE rather than other age groups in all target labels,which may be very useful during classification.
sex	X	X	uninformative feature.
blood_type	X	X	Transformed into OHE by the different blood types.
current_location	X	X	Found irrelevant for further analysis, checked in zipAdd
happiness_score	X	X	Feature is not correlated with other features, and did not provide any information during the graphs analysis phase in question 22.
household_income	X	X	uninformative feature.
pcr_date	X	X	after converting its values to numerical we couldn't get any meaningful information out of this feature.
symptoms	X	X	transformed into 5 features , each representing a possible symptom.
conversation_per_day	X	X	unsatisfying results in both graph analysis ,and checking importance in decision tree phases. both phases occurred in (Q22).
sugar_levels	V	X	correlated with age and weight, those 3 features may be used for group classification.
weight	V	X	correlated with age and weight, those 3 features may be used for group classification.
job	X	V	Found irrelevant for further analysis
sport_activity	V	X	low - moderate correlated with age,sugar_levels,and weight. may be useful for classification,together with these features.
pcr_01	V	X	may beessential,especially for risk classification,together with pcr_02,and pcr_05.
pcr_02	V	X	may be essential,especially for risk classification,together with pcr_01,and pcr_05.
pcr_03	V	X	may be essential,especially for spread classification,together with pcr_10,pcr_08,and pcr_07.
pcr_04	V	X	From decision tree analysis it seems we may gain useful information from this feature.
pcr_05	V	X	may beessential,especially for risk classification,together with pcr_02,and pcr_01.
pcr_06	X	X	Highly correlated with pcr_05.

pcr_07	V	X	may be essential,especially for spread classification,together with pcr_10,pcr_08,and pcr_03.
pcr_08	V	X	may be essential,especially for spread classification,together with pcr_10,pcr_03,and pcr_07.
pcr_09	V	X	from decision tree analysis it seems we may gain useful information from this feature
pcr_10	V	X	may be essential,especially for spread classification,together with pcr_3,pcr_08,and pcr_07.
O+	V	V	as seen in (Q22),may be useful for covid classification.
A+	V	V	as seen in (Q22),may be useful for covid classification.
AB+	X	V	uninformative feature,with only fewer samples possessing this feature - makes it even harder to generate any based observations.
B+	X	V	uninformative feature
A-	V	V	may be essential for covid classification.
AB-	X	V	uninformative feature,with only fewer samples possessing this feature - makes it even harder to generate any based observations.
O-	X	V	uninformative feature.
B-	X	V	uninformative feature,with only fewer samples possessing this feature - makes it even harder to generate any based observations.
zipAdd	X	V	extracted from address.we were not able to gain any information from this feature after converting it's values from string to numerical,mainly because of the large amount of unique value.
fever	V	V	extracted from symptoms. provides very informative information for covid prediction.
low_appetite	X	V	extracted from symptoms. it does not seem that there's any connection between these symptoms and target labels.seems that this information may even harm in prediction.
cough	V	V	extracted from symptoms. provides very informative information for covid prediction.
headache	X	V	extracted from symptoms. it does not seem that there's any connection between these symptoms and target labels.seems that this information may even harm in prediction.
shortness_of_breath	V	V	extracted from symptoms. provides very informative information for covid prediction.
state	X	V	extracted from address.we were not able to gain any information from this feature,mainly because of the large amount of unique value.
zipcode	X	V	extracted from address.we were not able to gain any information from this feature,mainly because of the large amount of unique value.

address	X	X	Parsed into state,zipcode,zipAdd features for better analyzing and understanding.