**למידת מכונה – פרויקט מסכם**

**קבוצה 2: דורון שפירא** 316208032  **בן מלי** 204148225  **אריאל רגב** 321776627

**תקציר מנהלים:**

**חלק ראשון – אקספלורציה:**

בשלב זה בוצעו חיתוכים וניתוחים ויזואליים של הדאטה, וזאת על מנת לאפיין את הנתונים ולהבין כיצד להתמודד איתם.

ראשית, טענו את הדאטה לאובייקט data frame, ונוכחנו לדעת כי יש בידינו 21 פיצ'רים ו-21,931 דגימות. כל דגימה קיבלה label של 0 או 1. על פניו, קיימים לנו X פיצ'רים מספריים ו- X פיצ'רים קטגוריאליים, כולל אלו הבינאריים. עבור כל פיצ'ר בדקנו כמה ערכים חסרים יש.

מטריצת קוראלציה

הסתכלות ויזואלית על ההתפלגות של כל פיצ'ר

להבין את אופי הפיצ'רים (איזה פיצ'רים נומריים, קטגוראליים, string וכו)

להוסיף fitted parameters from the distribu

הורדנו:

17 בגלל קורלציה

15 שרוב הדגימות היו 0

0 רוב הדגימות היו 0

**חלק שני – עיבוד מקדים:**

ראשית, נרצה להתייחס לסוגיית **הנתונים החריגים (Outliers)**. כפי שעולה מהאקספלורציה שבוצעה בחלק הראשון, ניתן להבחין כי אכן קיימים לנו נתונים חריגים בסט הנתונים, כאלו אשר עלולים ליצור הטיות משמעותיות במדדים הנבדקים, ועל כן נרצה לקבל החלטה כיצד לנהוג עמם. מאחר ובידינו קיים מספר רב למדי של תצפיות, נוכל לנהוג בגישת הסרה אגרסיבית של תצפיות חריגות עבור הפיצ'רים הנומריים. המתודולוגיה לפיה תתבצע ההסרה תתבסס על אחת מהשתיים המקובלות – שימוש ב-IQR או שימוש בסטיית תקן. בהתאם למה שלמדנו בתרגול, השימוש בשיטת IQR, אשר משמיטה תצפיות הגדולות מ-2.7 סטיות תקן (הן לצד החיובי והן לצד השלילי), גרם לכך שהרבה תצפיות נפלו מחוץ לתחום הבטוח ולכן נפסלו, מה שהוביל אותנו להעדיף להשתמש בשיטת סטיית התקן, זו שמרחיבה במעט את תחום הקטימה ומשאירה למעשה יותר תצפיות (מרווח של 3 סטיות תקן). בנוסף, הבחנו כי בפיצ'ר 19 יש כ-80% ערכים ריקים, ומאחר ופיצ'ר זה קטגוריאלי, לא ניתן היה להשלימו בצורה נכונה, ולכן החלטנו להוריד אותו במלואו.

כעת, נרצה לתת את הדעת אודות נוגעת **הפיצ'רים הקטגוריאליים** בסט הנתונים שלנו. לטובת השלבים הבאים בניתוח, אנו נדרשים להחליט אודות אחד משני תהליכי המרה אפשריים עבור פיצ'רים אלו – One Hot Encoding או Label Encoding. שימוש בשיטת One Hot Encoding טומן בחובו יצירת מימדים רבים אך נטולי הטייה, ואילו עבור Label Encoding בדיוק ההיפך – מימד בעיה נמוך יותר אך מוטה. אנו בחרנו להשתמש בשיטת One Hot Encoding כדי להימנע מהטעיות כאלה ואחרות לסט הנתונים שלנו.

יישמנו את שיטת One Hot Encoding בשני אופנים שונים. לקחנו את סט הנתונים שהתקבל לאחר הסרת החריגים, ופיצלנו לשתי אפשרויות. הראשונה הייתה הפעלת One Hot Encoding על סט הנתונים כפי שהתקבל – as is. השנייה הייתה לבצע Feature Selection ולאחר מכן ליישם את ה- One Hot Encoding. בשלב ה- Feature Selectionהבחנו במספר פיצ'רים אשר היו זקוקים ל"תיקון". בפיצ'ר 2 כל דגימה הייתה נומרית אך הכילה את האות "d" בסופה, ואילו בפיצ'ר 18 גם כן היו דגימות נומריות אך עם האות "a" בתחילתן. בעבור שני הפיצ'רים הללו, ביצענו תיקון והורדנו את האותיות המיותרות, לקבלת דגימות נומריות טהורות בלבד. כמו כן, בפיצ'ר 12 היו שני ערכים קטגוריאליים של "y" ו-"n", אשר בחרנו להמיר אותם ל-1 ול-0, בהתאמה. בנוסף, החלטנו להוריד את פיצ'ר 17 אשר היה קורלקטיבי עם פיצ'ר 5 (קורלציה של 90%). פיצ'רים נוספים אשר הורדנו היו 15 ו-0, וזאת מאחר והכילו כמות גבוהה של אפסים, כלומר לא תרמו מידע נוסף. לאחר סינון הפיצ'רים הללו ביצענו כאמור את ה-One Hot Encoding וקיבלנו את סט הנתונים של האפשרות השנייה.

בתום הסרת החריגים נותרו בידינו 19,855 דגימות לעומת 21,931 שהיו בהתחלה.

כעת נעבור להתייחס **לטיפול בחוסרים**. בנתונים החסרים בסט הנתונים החלטנו לטפל באמצעות ב-4 דרכים שונים, כלומר 4 Imputers שונים. הראשון שבהם, Simple Imputer, התבסס על השלמת הערכים החסרים באמצעות שימוש בחציון של כל פיצ'ר. השני, Distribution Imputer, התבסס על השלמת ערכים לפי הגרלת ערך רנדומי מההתפלגות המשוערכת של הפיצ'ר. כלומר, עבור כל פיצ'ר בוצע ניסיון השוואה להתפלגויות מוכרות שמזכירות בצורתן את זו של הפיצ'ר. ע"י שימוש במדד SSE, בחרנו את ההתפלגות המוכרת שמתאימה בצורה הטובה ביותר לנתוני הפיצ'ר, כלומר בעלת ערך ה-SSE הקטן ביותר (ניתן להשתמש ב-SSE מאחר וכמות הדגימות עבור כל התפלגות נבחנת הייתה זהה). ההתפלגויות המוכרות שנבחנו הינן: נורמלית, F, חי בריבוע, אקספוננציאלית. הדרך השלישית שעשינו בה שימוש הייתה Choice Imputer. בגישה זו בחרנו לבצע את ההשלמות עבור כל פיצ'ר ע"י הגרלת דגימה מסט הנתונים של אותו פיצ'ר, כאשר כל דגימה הינה בעלת הסתברות זהה, וזאת מאחר ובידינו יש מעל 21 אלף דגימות, והנחנו כי דבר זה מייצג כהלכה את כלל הדגימות האפשריות עבור אותו פיצ'ר. הגישה הרביעית והאחרונה הייתה Iterative Imputer. בגישה זו ההשלמה נעשית באמצעות חיזוי שנעשה ע"י סוגי רגרסיות שונות, וזאת על מנת למצוא את הערך המסתבר ביותר מהרגרסיות השונות, ולבסוף איתו תתבצע ההשלמה[[1]](#footnote-1).

השלב הבא בתהליך העיבוד המקדים הינו **נרמול הנתונים**. לאחר בדיקת הנתונים ניתן היה לראות כי הנתונים אינם מנורמלים, ולכן נרצה לנרמל אותם כדי שלכל הפיצ'רים יהיה אותו טווח ערכים, כלומר שכל הפיצ'רים "ידברו" באותה שפה. לטובת הנרמול השתמשנו בשיטת MinMax, וזאת משום שרצינו שכל הדגימות בכל הפיצ'רים יהיו בטווח שבין 0 ל-1, מה שיעלה בקנה אחד עם הדגימות הבינאריות של הפיצ'רים הקטגוריאליים, שבלאו הכי נמצאים בטווח זה (לאחר ביצוע One Hot Encoding).

השלב האחרון בעיבוד המקדים היה נושא **הקטנת המימדיות**. טרם ביצוע השלב הזה אנו עדיין נותרים עם כמות גבוהה של מימדים, 512 בסטי הנתונים ללא ה-Feature Selection, ו-30 בסטים שעברו Feature Selection. נבחין כי 512 המימדים עתידים ליצור עומס ומורכבות גבוהה מאוד שתצריך כוח עיבוד משמעותי בשלבים הבאים של יישום המודלים. בהמשך ישיר למה שלמדנו בהרצאה לגבי כלל האצבע ש-n מימדים מצריכים לפחות n^2 תצפיות, במקרה שלנו ישנן כאמור כ-20,000 תצפיות בלבד עבור מעל ל-500 מימדים, ולכן אנו ניעזר בשיטת PCA לטובת הורדת המימדיות לסטי הנתונים הללו. בנוסף, למרות שבסטי הנתונים האחרים ישנם 30 מימדים, אנו בחרנו להחיל את שיטת ה-PCA גם עליהם, וזאת מאחר ואם יתאפשר להוריד את המימדיות ולשמור על הסבר מספק של השונות, אז בהחלט זה יפחית, גם אם במעט, את הסיבוכיות החישובית של המודלים. בכל ניתוח PCA שכזה, אנו נבחן 3 אפשרויות של אחוז השונות המוסברת - בראשונה נדרוש כמות פיצ'רים שתסביר 90% מהשונות, בשנייה 95% ובשלישית 98%. עבור כל אפשרות כזו בחרנו להריץ, כמעין מודל מייצג, את מודל ה-ANN, ובחרנו לקחת את אפשרות ה-PCA בעלת ערך ה-AUC הגבוה ביותר. נציין כי כל אפשרות אנו בדקנו הן על סט הנתונים לאחר ה- One Hot Encoding ללא התיקונים, ועל הסט השני שכן כלל תיקונים.

נבחין כי בשלב זה יש בידינו 24 סטי נתונים (3\*4\*2). על מנת לבחור את סט האימון שעליו נריץ בהמשך את המודלים, נדרש מאיתנו לבצע הערכה כלשהי ולהכריע בסוגייה זו. לטובת ההערכה, החלטנו לבצע פרידיקציה באמצעות ANN, בחירה שרירותית. עבור כל סט אימון פוטנציאלי בוצעו 5 פרידיקציות בשיטת K-Fold (עם 5 פולדים), וכמו כן חושב מדד ה-AUC הממוצע עבור כל סט. בסופו של דבר, נבחר סט האימון בעל ערך ה-AUC הממוצע הגבוה ביותר.

לאחר שמצאנו את סט האימון הטוב ביותר, אנו נבצע על סט המבחן את כלל שלבי העיבוד המקדים שיושמו על אותו סט ספציפי, שלב אחר שלב.

1. <https://scikit-learn.org/stable/auto_examples/impute/plot_iterative_imputer_variants_comparison.html> [↑](#footnote-ref-1)