**תרגיל בית 2 מבוא למערכות לומדות**

**דוח עבודה**

שלבי עבודה:  
1. **חלוקת הדאטה**:  
תחילה ביצענו חלוקה של הדאטה ל3 קבוצות – train(70%), val(15%), test(15%)  
ביצענו את החלוקה באמצעות הפונקציה StratifiedShuffleSplit של sklearn, כאשר החלוקה התבצעה בצורה אקראית כאשר נשמר היחס בין הלייבלים.  
לאחר פיצול שמרנו את הדאטה המקורי ללא שינוי.

2. **המרת פיצ'רים למספריים**:  
דבר ראשון שביצענו היה זיהוי כל הפיצ'רים וסוגם – כלומר זיהינו איזה פיצ'רים הינם נומינליים, בינארים, רציפים וכו'.  
לאחר זיהוי זה, ביצענו המרה של הפיצ'רים לערכים מספריים על מנת שנוכל לעבוד איתם בהמשך, כאשר עבור פיצ'רים נומינליים ביצענו כמה שינויים – עבור פיצ'רים ללא חשיבות לסדר, ביצענו המרה ל one-hot, עבור פיצ'רים עם חשיבות לסדר ו-3 ערכים אפשריים (כמו "Age Group") ביצענו המרה לערכים -1 0 1, לפי הסדר המתאים.  
עבור פיצ'רים בינאריים ביצענו המרה ל -1 1 .  
ביצענו זאת על מנת שהפיצ'רים יהיו ממורכזים סביב 0 (בהמשך נבצע scaling אשר ימרכז את כל הפיצ'רים ל0)  
המרות אלה ממומשות בפונקציות :   
convert\_to\_onehot, convert\_to\_categorical, change\_binary\_values

3. **תיקון ערכים:**כאשר בחנו את הדאטה שקיבלנו, זיהינו כי ישנם כמה פיצ'רים אשר מכילים ערכים שליליים כאשר הפיצ'ר עצמו אינו אמור להכיל ערכים שליליים.  
פיצ'רים אלה הם 'Avg\_monthly\_expense\_when\_under\_age\_21', 'Avg\_lottary\_expanses', 'Avg\_monthly\_income\_all\_years'  
לאחר שזיהינו זאת, בחנו את הפיצ'רים האלו על פי ההתפלגות שלהם וזיהינו כי הערכים השליליים נלקחים מתוך ההתפלגות של הפיצ'רים אם היו בעלי סימן חיובי.  
כלומר לדעתנו בזמן הדגימה נפלה טעות והערכים נרשמו עם סימן שלילי במקום חיובי ולכן החלטנו לשנות את הסימן שלהם לחיובי במקום למחוק אותם.  
פעולה זאת מתבצעת בפונקציה abs\_negative.  
להוסיף גרף המציג את התפלגות הפיצ'רים בלי שליליים ואת התפלגות השליליים

4. **Outlier detection:**בהתמודדות עם outliers החלטנו לטפל רק בסט האימון שלנו, זה נובע מכך שבחרנו לבצע את זיהוי ה – outlier על ידי התייחסות ללייבל עצמו ולהתפלגות הדאטה ביחס ללייבל.  
עבורי הסטים של test, val בחרנו לא לבצע אף פעולה של זיהוי outlier בגלל שבמצב אמיתי, לא נוכל לדעת את הלייבלים של הדאטה ונצטרך לבצע פעולות ללא התייחסות ללייבל.  
לדעתנו אי טיפול זה גורם לכך שהבדיקה של הדאטה שלנו על סט הבדיקה תהיה כמה שיותר קרובה למצב אמיתי.  
זיהוי outlier – על מנת לבצע זיהוי זה חישבנו z\_score עבור כל אחד מהדוגמאות בדאטה שלנו ולכל פיצ'ר.  
כעת בחרנו את הדוגמאות בעלות הפיצ'רים אשר מקיימים כי הz\_score שלהם גדול מסף מסויים (בחרנו ב4.5 כסף) ומחקנו את המקומות המתאימים (כלומר הצבנו NAN בכל דוגמא עם פיצ'ר שזיהינו כ – outlier).  
בשימוש בשיטה זאת, בעצם אפשרנו לדאטה שלנו בכל פיצ'ר להיות במרחק של 4.5 סטיות תקן מהממוצע. להוסיף גרף של היסטוגרמה עם סימון מה שזיהינו  
פעולה זאת מתבצעת בפונקציה remove\_outlier

5. **Imputation**:  
את הטיפול בערכים חסרים ביצענו בשתי שיטות, כאשר שיטה אחת הופעלה את סט האימון שלנו, ושיטה שנייה על הסטים של הבדיקה.  
נזכיר כי את הטיפול בסט האימון נוכל לבצע עם ייחוס ללייבלים של הדאטה, כאשר עבור סט הולידציה והבדיקה לא ניתן להתייחס ללייבל מפני שמידע זה לא ידוע לנו במקרה אמיתי.  
א. טיפול בסט האימון:   
עבור סט האימון בחרנו בשיטה הנקראת Bootstraping שלמדנו בכיתה.  
בשיטה זו, עבור כל ערך חסר, השלמנו על ידי כך שדגמנו מתוך הדאטה שלנו עם חשיבות ללייבל ערך חדש.  
כלומר, בהינתן כי חסר לנו ערך בפיצ'ר מס' 1 עבור דוגמא עם לייבל 5, יצרנו סט חדש המכיל רק דוגמאות מתוך סט האימון שהלייבל שלהם הוא 5 (ללא ערכים חסרים בפיצ'ר מס' 1) ולאחר מכן דגמנו ממנו סמפל דוגמא יחידה, והעתקנו את ערך פיצ'ר 1 שלה לערך החסר בסט המקורי.  
בצורה זאת, אנחנו שומרים על פילוג הדאטה שלנו ומשלימים ערכים חסרים עם חשיבות ללייבל המתאים להם. לסדר את הדוגמא שתהיה ברורה יותר   
ב. טיפול בסט הולידציה והבדיקה:  
עבור סטים אלו בחרנו לטפל בצורה שונה.  
תחילה לקחנו את סט האימון, ועבור כל פיצ'ר חישבנו את ממוצע הפיצ'ר עבור ערכים נומריים ואת הערך הנפוץ ביותר עבור ערכים נומינליים (ללא חשיבות ללייבל).  
לאחר מכן, השלמנו את סט הולידציה והבדיקה בעזרת הערכים שחושבו מסט האימון.  
בשיטה זאת, אנחנו מונעים התייחסות ללייבל בסט הולידציה והבדיקה, ועדיין משלימים ערכים חסרים עם חשיבות לדאטה שלנו ולא בצורה עיוורת.  
פונקציות הלו ממומשות במחלקה Imputation

6.**Scaling**: