

1. ראשית עבור כלל הנתונים ביצענו:
 - א. קבענו טיפוסים חדשים לכל עמודה להוציא את 'Vote':
 1. אם הטיפוס המקורי הוא object אז הטיפוס החדש הפך להיות category.
 2. אחרת אם היו בעמודה פחות מ-1000 ערכים ייחודיים הנחנו שמדובר בעמודה מטיפוס נומינאלי והגדרנו לה את הטיפוס להיות int32.
 3. אחרת שמרנו על הטיפוס המקורי(ברוב המקרים מדובר ב-float64).
 - ב. השלמנו מידע חסר בשיטת ClosestFit: עבור כל ערך חסר חיפשנו את 1000 התצפיות הקרובות ביותר לאותה תצפית והשלמנו את המידע לפיה. ה-1000 שמתוכם חיפשנו נבחרו בכל פעם באקראי.
2. כעת חילקנו את הנתונים ל-train, test ו-validate בגדלים 70%, 20% ו-10% בהתאמה. שמרנו שני עותקים של כל סט.
3. עבור כל סט ביצענו:
 - א. הסרנו רעשים: אחרי שבחנו את הערכים בכל עמודה בעזרת plotting ומצאנו עמודות שלא יכולות להכיל ערכים שליליים הסרנו את כל התצפיות עם ערכים שליליים בעמודות אלו.
 - ב. הסרנו outliers: בעזרת שיטת ESD חיפשנו עד 50 outlier-ים בכל עמודה והסרנו את התצפיות שהכילו אותם.
 - ג. נירמול: ראשית בחנו שוב את הנתונים בכל עמודה וזיהינו את התפלגות של אחת. חילקנו ל-3 קבוצות – אחידה, נורמלית ואחרת. עבור עמודות מהתפלגות אחידה נירמלנו בשיטת ה-MinMax. את אלו עם התפלגות נורמלית נירמלנו בשיטת Z-Scale. ואת השאר בעזרת Decimal Scalling.
 - ד. בחירת עמודות – feature selection: השתמשנו בשתי שיטות:
 1. Filter Method: חיפשנו את רמת התאמה בין כל שתי עמודות ואם מצאנו התאמה גבוהה מ-0.5 הסרנו את אחת העמודות.
 2. Wrapper Method: השתמשנו במסווג KNN על-מנת לסנן עמודות.