**דו"ח ניסוי ב-unbalanced datasets**

**מטרה**

בחינת דרכי התמודדות שונות עם unbalanced datasets.

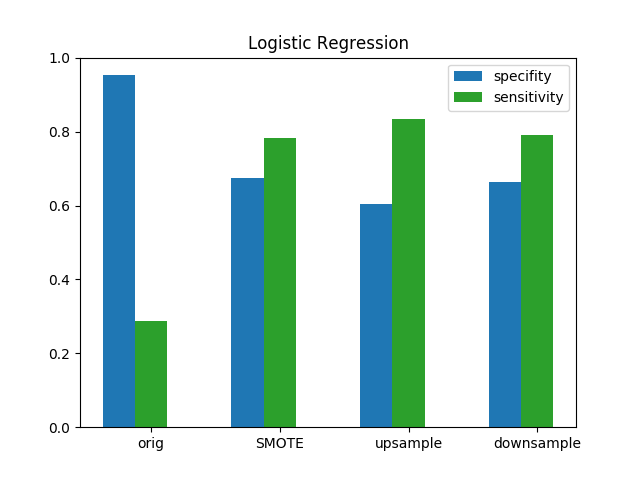
**תקציר**

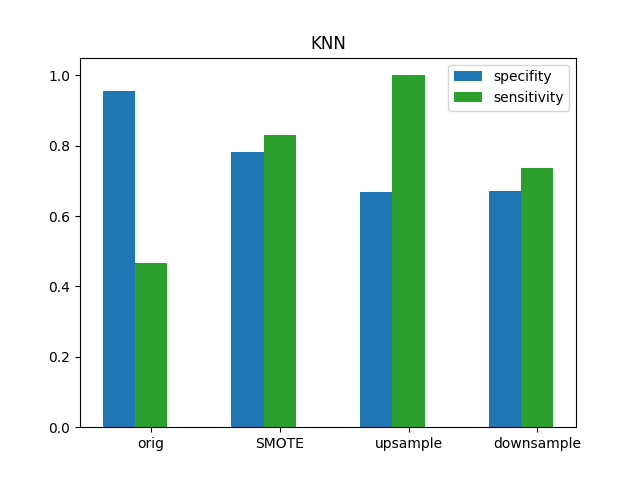
בניסוי נבחן Adults dataset שהינו unbalanced. יש תיוגים רבים יותר של 0 ב-dataset מאשר של 1. (יחס 3:1 בערך). נבחנו דרכי פעולה שונות ביניהן SMITE,upsamplling,downsampling על מס' מודלי קלסיפיקציה שונים והם Logistic Regression,KNN,Decision Tree. אפשר להבחין בשיפור משמעותי ב-sensitivity של שני המודלים הראשונים בשימוש בשיטות הנ'’ל. המודל השלישי הצליח לסווג את המידע ללא בעיה.

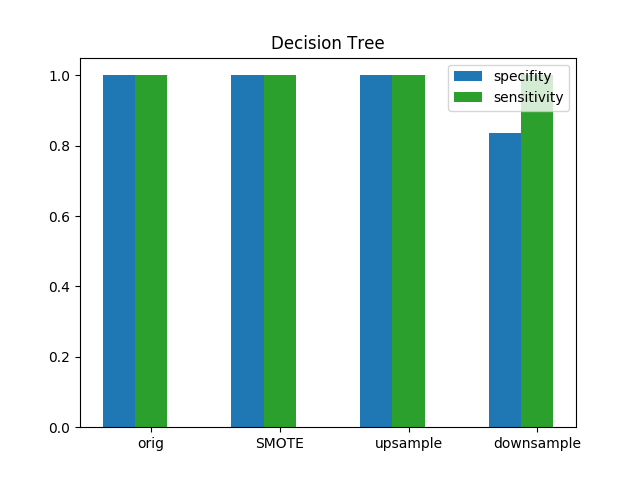
**מהלך**

הדאטה עובד תחילה בשלב preprocess. לאחר מכן הותאמו אליו שלושה מודלי ML והם Logistic Regression,KNN,Decision Tree. עבור שלושת המודלים נבחנו ה-accuracy,sensitivity,specifity.

הופעלו 3 שיטות והן SMOTE,Upsampling,downsampling לפיתרון הבעיה.גרף המתאר את השפעת השיטות עבור כל מודל:







\*היו פחות דוגמאות ממחלקה 1.

תוצאות

עבור שני המודלים הראשונים של LR,KNN התקבלה סטייה חזקה בסיווג נכון של ה-majority class. אפשר לראות כי הפעלת שיטות ה-sampling השונות מעלות את ה-sensitivity של המודלים רבות, כפי שציפינו. עבור העץ התקבל סיווג כמעט מושלם של 99.9% כבר ללא השיטות, ולכן הן לא תרמו לו ואף downsamplling מעט הרס אותו.

דיון ומסקנות

בדו'’ח זה נבחנו מס' דרכים להתמודדות עם unbalanced datasets, ע'’י בחינה של מידע טבלאי ושימוש בשיטות samplling שונות. בחרתי להתמקד בשיטות אלו מכיוון שהן המוכרות ביותר לטיפול בבעיה, והן הפשוטות ביותר.

כפי שראינו, עצים עובדים יחסית טוב למרות ה-unbalanced, ואכן כפי שאפשר למצוא בספרות הם אכן פתרון אפשרי לבעיה.

בחינה של שיטות נוספות מומלצת ואף הכרחית לפתרון מלא יותר של הבעיה, ושילוב שיטות שכאלו הינו רעיון שכדאי לבחון בהמשך.