מס' פרויקט	שם הפרויקט	
2022-01-134	מערכת למציאת חולי סוכרת בטוויטר ולמידה מניסיונם האישי	
מנחה אקדמי	**מנחה שותף**	
פרופ' ישראל פרמט		
חברי הצוות		
עידן הרשקוביץ	ניב כהן	
idanhers@post.bgu.ac.il	Niv2@post.bgu.ac.il	

תקציר

בפרויקט זה נבנה מערכת למציאת חולי סוכרת בטוויטר. בפרויקט נממש מודל הזהה במהותו לזה שהוצג בעבודת הדוקטורט של מיה שטמר (A Framework for Identifying Patients on Twitter from Their Personal Experience, 2020 and Learning), ונסמך על מסקנותיו. כיום אנשים רבים ברחבי העולם משתפים מנסיונם האישי מידע אשר ללא כלים מתאימים נותר לא נגיש. אנו מאמינים כי מאגר המידע שיווצר בעזרת רצף העבודה שנציג בפרויקט, עשוי להיות שימושי בידי חוקרים וגורמים מסחריים. כפי שמיה הראתה בעבודתה, על מאגר מסוג זה ניתן לבצע ניתוחים טקסטואליים על מנת ללמוד מנסיונם האישי של חולי סוכרת.

ראשית נמשכו בעזרת רישיון המחקר של טוויטר (twitter API) כל הציוצים הפרושים על פני יממה, שבהם נכללו מילים שנמצאו כקשורות למחלת הסוכרת. המשתמשים שכתבו את הציוצים שעלו מחיפוש זה סווגו ידנית כחולים/לא חולים, כך שכל משתמש סווג פעמיים (ע"י כל אחד מכותבי הפרויקט) ונכנס למאגר רק לאחר הסכמה חד משמעית באשר לסיווגו. לכל אחד מהמשתמשים מתוך מאגר המשתמשים המתויגים שיצרנו נמשכו כלל ציוציו במשך שבוע שלם. המודל אותו שאפנו לממש הינו מודל מסוג MI (Multi Instance), מאחר ולכל משתמש מספר שונה של דגימות. בחנו שני מודלים שונים. במודל הראשון, עבור כל משתמש יצרנו וקטור metadata של פיצ'רים, הנבנה על כלל ציוציו, ואינו מקושר לאף ציוץ ספציפי. במודל השני וקטור הפיצ'רים נבנה על כל ציוץ ממאגר המידע בנפרד, כאשר משתמש סווג כחולה אם לפחות ציוץ אחד שלו קיבל סיווג חיובי. הפיצ'רים שנבחנו כללו פיצ'רים משלושה סוגים: פיצ'רים התנהגותיים המתארים את פעילותו של המשתמש ברשת החברתית, פיצ'רים מילוליים המסתמכים על ניתוח טקסט הציוצים ופיצ'רים הנדסיים שנבנו בהתאמה לבעיה הניצבת בפנינו. על מאגר המידע שהתקבל הרצנו מודלים שונים מתחום לימוד המכונה. בעיה נוספת שעמדה בפנינו הייתה המידע הלא מאוזן שהתקבל מתהליך התיוג, שכן רק כ-16% מהמשתמשים סווגו כחולים. כדי להתמודד עם בעיה זו נבחנו שיטות שמבצעות מניפולציות שונות על הנתונים, על מנת להתאים את הבעיה לפתרון ע"י אלגוריתמים סטנדרטיים של לימוד מכונה. כלל המודלים נבחנו גם בשיטת k-fold להערכת ביצועי המודל ע"י מדדי דיוק שונים.

רוב המודלים שנבחנו הציגו ערכי דיוק משביעי רצון, כאשר הממוצע המשוקלל של אחוזי הדיוק נע בטווח שבין 0.7% לבין 0.93% דיוק. אלגוריתם הסיווג שהביא לתוצאות הטובות ביותר הוא זה שבו הוספו דגימות באופן סינטטי בשיטת SMOTE. בשיטה זו, עבור אלגוריתם Random forest ולאחר ביצוע ולידציה בשיטת 5-fold, אחוזי הדיוק שהתקבלו היו 0.93122% עבור מדד Precision.

לסיכום, בפרויקט זה חיזקנו את המסקנות שעלו מעבודת המחקר עליה נסמכנו, ומימשנו את רצף העבודה כפי שהוצג. בעזרת המודלים השונים שיצרנו ניתן לבנות מאגר משתמשים שבסבירות גבוהה משתפים מנסיונם האישי על מחלת הסוכרת. אנו מאמינים כי בעזרת כלי מחקר מתקדמים, פרויקט זה עשוי לשמש כבסיס למחקר עתידי, ולניתוח יחסם של מטופלים בסוכרת באשר לאורחות חייהם והרגליהם.

מילות מפתח: לימוד מכונה, סיווג, יצירת דאטאסט, מידע לא מאוזן