# Logistic regression

הפרויקט עוסק בזיהוי המגדר לפי קטעי קול שהוקלטו והומרו לפרמטרים (פירוט במטלה 1). זיהוי המגדר נעשה ע"י logistic regression שממומש בTensorflow.

המודל חולק לארבעה חלקים ובין החלק השני לשלישי הוגדרו משתנים לשימוש הפונקציות ב- tensor flow.

**חלק ראשון(read_data_set):**

החלק הראשון עוסק בקריאת ה data set של המודל (במודל שלנו הוא מוצג כקובץ csv) על ידי שימוש בספרייה pandas. בגלל שהפרמטר label במודל מוצג כערכים male, female אז הוחלפו הערכים שיהיו 0,1 בהתאמה ואז עורבבו השורות על מנת שלא יהיה סדר (בקובץ csv הופיעו קודם כל הדגימות של הגברים ורק אז הדגימות של הנשים) ואז המידע הוצב במשתנה voice.

**חלק שני(split_data_set):**

החלק השני עוסק בלקיחת המידע שהושג בחלק הראשון וחילוקו ל train ו- test. החילוק נעשה על ידי הכנסת הנתונים שנמצאים במשתנה voice לתוך שני מערכים מהספרייה numpy. מערך ראשון data_y מכיל את כל הערכים שנמצאים בפרמטר label ומערך שני data_x מכיל את שאר הערכים שלא הוכנסו מקודם. אחרי החילוק של שני המערכים כל מערך חולק לשני מערכים ביחס של 30-70. ארבעת המערכים הוצבו לתוך המשתנים train_data_x , train_data_y , test_data_x , test_data_y.

בין החלק השני לשלישי הוגדרו המשתנים של הספרייה tensor flow בצורה שהמודל יוכל לחשב את פונקציית ה- logstic $h(x) = \frac{1}{1+e^{-(xW+b)}}$ בעזרת הכלים שהוגדרו ב- tensor flow.

**חלק שלישי(model_training):**

בחלק השלישי נלקחו הנתונים שנמצאים ב- train_data_x , train_data_y , ועל סמך הנתונים האלו המודל התחיל לחשב את פונקצית ה- logstic על מנת למצוא את המשקלים הטובים ביותר. בתוכנית נעשו 10000 חזרות על הנתונים עד שהמודל סיים את שלב האימון.

**חלק רביעי(model_testing):**

בחלק הרביעי אחרי מציאת המשקלים בחלק השלישי יבדק עד כמה המודל מדויק על ידי שימוש במידע הנותר שנשאר ב- test_data_x , test_data_y על ידי הרצה של דגימה והשוואה בין התוצאה של המודל לבין התוצאה האמיתית. על יד השוואת התוצאות יחושב ה- , recall , accuracy, F-measure precision של המודל.

**מסקנות:**

אחרי ה- train(2217 דוגמאות) וה- test(951 דוגמאות) שנעשו על המודל התוצאות שהתקבלו הם:

Accuracy – 0.807570977917981

Recall – 0.6673346693386774

Precision – 0.9514285714285714

F-measure – 0.7844522968197879

התוצאה שהמודל הגיע היא תוצאה טובה בגלל שהמודל מנבא בצורה יותר טובה יותר מאשר מודל שמבוסס על רנדומליות ששם הסיכוי לבחור נכון הוא 50%.