**סיכום פרויקט נושאים בחזית המחקר**

**מנחה: ד"ר אורן פרייפלד**

**נושא הפרויקט למידה חצי מונחית**

**רקע-**

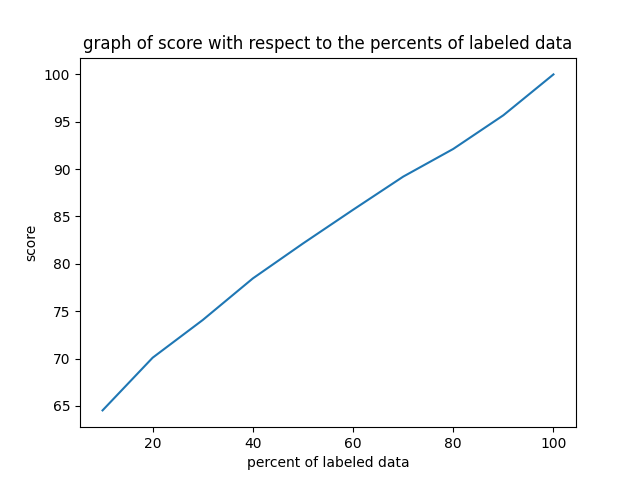
למידה חצי מונחית היא גישה בלמידת מכונה המנסה לסווג מידע בעזרת מידע שחלקו לא מתויג וחלקו מתויג. ההנחה היא שאפשר לשפר את המודל בעזרת שימוש במידע המקדים.

אנו רצינו לבדוק אם שימוש במידע לא מתויג משפרת את ההצלחה של מודל המזהה ספרות מ0-9.

**שלב מקדים-**

הורדנו Data-Set מתויג של תמונות של ספרות מ0-9. בשלב הראשון כדי להבין מהי למידה חצי מונחית השתמשנו באלגוריתם label propagation של sklearn, הרצנו את האלגוריתם עם מידע מתיוג באחוזים שונים את התוצאה השוונו למידע האמיתי שקיבלנו.

ראינו שיש קורלציה לינארית בן כמות המידע המתויג לבן ההצלחה של המודל.

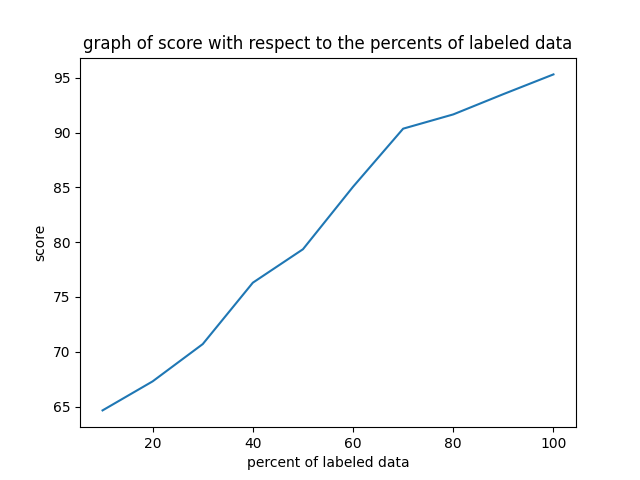


כש10% מהמידע היה מתויג האלגוריתם צדק בפחות מ65% מהתיוגים וכמובן שעבור 100% האלגוריתם צדק בכל המידע.

**בדיקת ההשארה-**

בשלב השני ניסינו להשתמש במידע הלא מתויג כדי לאמן מודל, מתוך 10,000 הדוגמאות לקחנו 8,000 תמונות שיהווה training-set למודל, ואת ה2,000 הנוספים השארנו בתור test set.

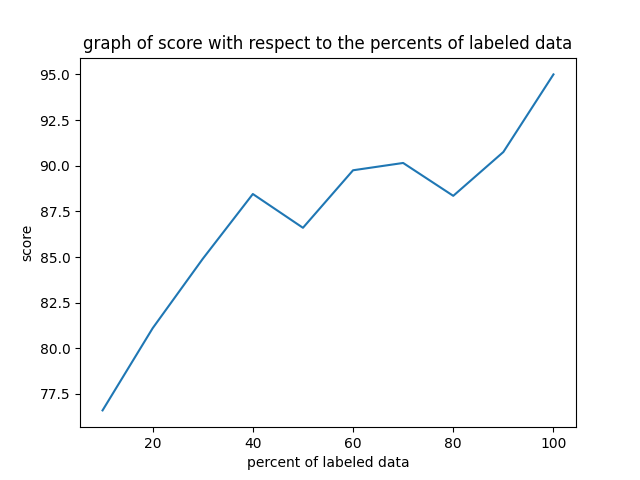
1. ביצענו PCA על מנת להוריד את הממד של המידע מ28\*28 ל128.
2. הרצנו אלגוריתם למידה לא מונחית וקיבלנו חלוקה של המידע לאשכולות.
3. בחרנו בצורה שרירותית עבור אחוזים שונים מידע שיהיה מתויג, לאחר מכן נתנו תוויות חדשות לכל אשכול שקיבלנו נתנו תיוג למידע הלא מתויג בשיטת "הרוב קובע" כלומר התווית המתאימה לרוב הנקודות המתויגות באשכול.
4. בעזרת הtraining set והמידע שתייגנו אימנו מודל שיזהה ספרות.
5. בדקנו עבור הtest-set את ההצלחה בזיהוי. ושמרנו את ההצלחה עבור כל כמות שונה של תיוג



**תוצאות-**

ניתן לראות שככל שמתייגים יותר מידע ההצלחה של המודל עולה.

אך על מנת לבדוק אם השימוש בטכניקה משפר את התוצאות בדקנו מה ההצלחה באימון המודל רק בעזרת המידע המתויג. דגמנו כמות זהה של מידע ואימנו מודל רק באמצעות המידע המתויג. לאחר מכן בדקנו על אותו הtest set את ההצלחה של המודל. וקיבלנו את הגרף הבא:



ניתן לראות שכאשר משתמשים בטכניקה הקודמת עם אחוז גבוהה של מידע מתויג, מעל 70% יש שיפור באיכות של המודל (ב100% מדובר באותו המקרה). אך בכמות נמוכה מזאת המודל לא משתפר ואף נהיה פחות טוב, ייתכן כי הסיבה לכך היא שכאשר מעט מידע מתויג יש יותר טעיות ונוספות למודל דוגמאות לא טובות, אך כאשר יש הרבה מידע מתויג אז ההצלחה עולה ונוספות למודל הרבה דוגמאות טובות שמשפרות אותו.