סיכום פרויקט נושאים בחזית המחקר

מנחה: ד"ר אורן פרייפלד

נושא הפרויקט למידה חצי מונחית

רקע-

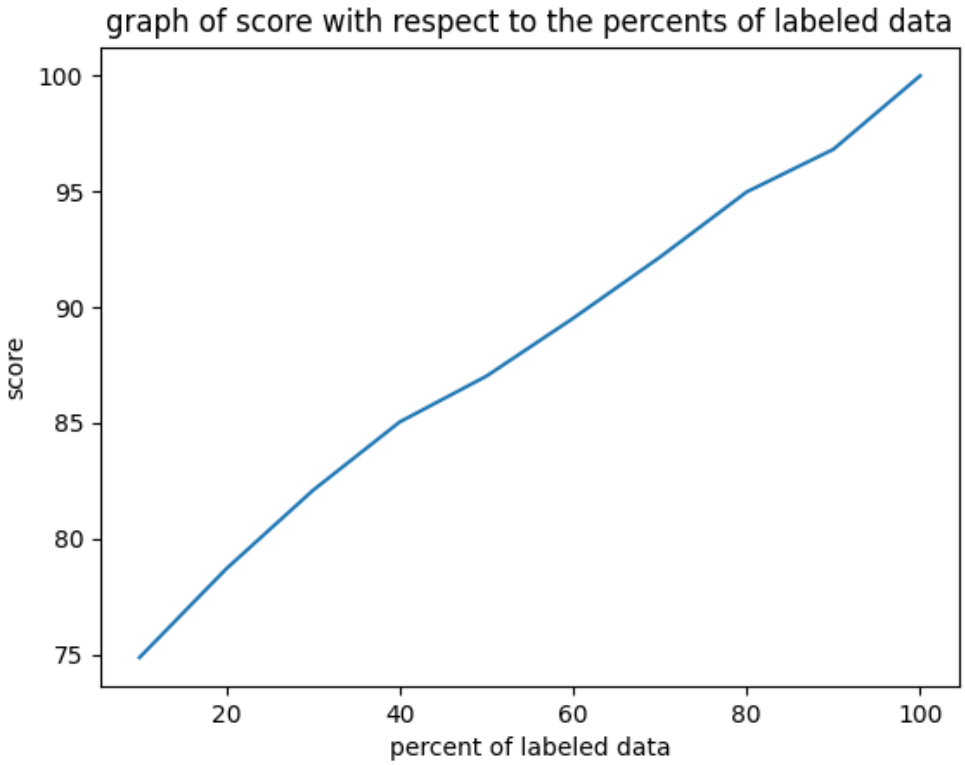
למידה חצי מונחית היא גישה בלמידת מכונה המנסה לסווג מידע בעזרת מידע שחלקו לא מתויג וחלקו מתויג. ההנחה היא שאפשר לשפר את המודל בעזרת שימוש במידע המקדים.

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 בכ75% מהתיוגים וכמובן שעבור 100% האלגוריתם צדק בכל המידע.

ההשארה של הפרויקט

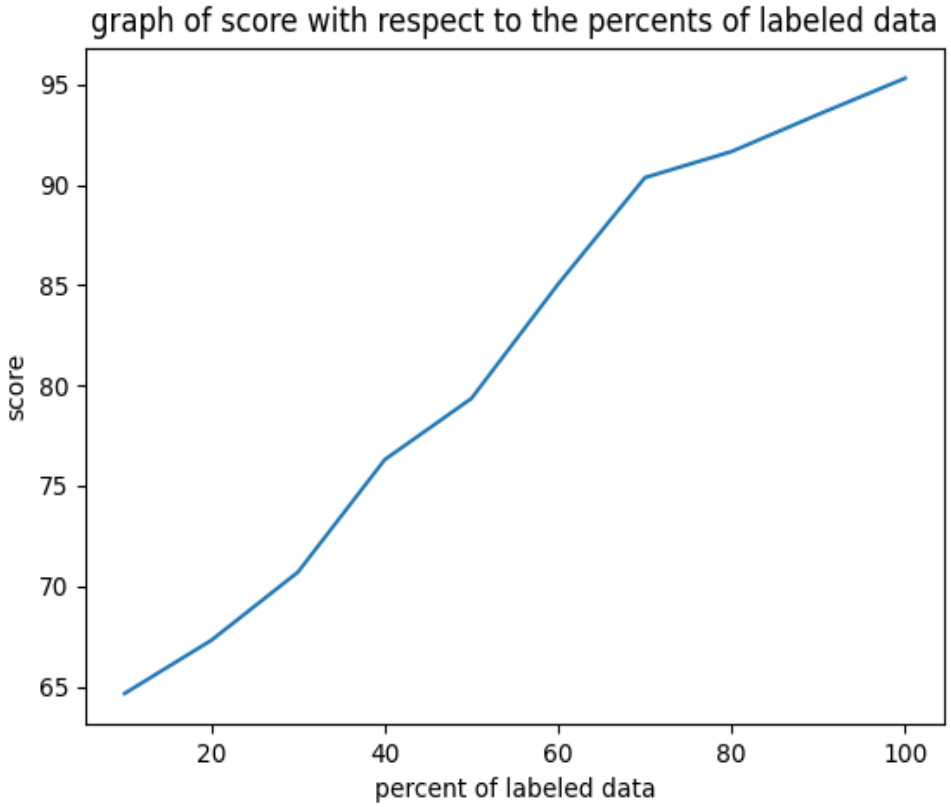
אנו רוצים לברר אם שימוש בטכניקות של למידה לא מונחית יכולה להגדיל את הtraining set ובכך לשפר את הביצועים של מודל.

ההנחה היא שמכיוון שנתונים זהים נוטים להשתכן קרוב במרחב אז אלגוריתם הclustering יצליח לתייג אותם באותו האשכול ובעזרת המידע המתויג שלנו נצליח לתייג נכון את רוב המידע. בנוסף אנו מניחים שככל שהtraining set גדול יותר כך המודל ישתפר.

בדיקת ההשארה-

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

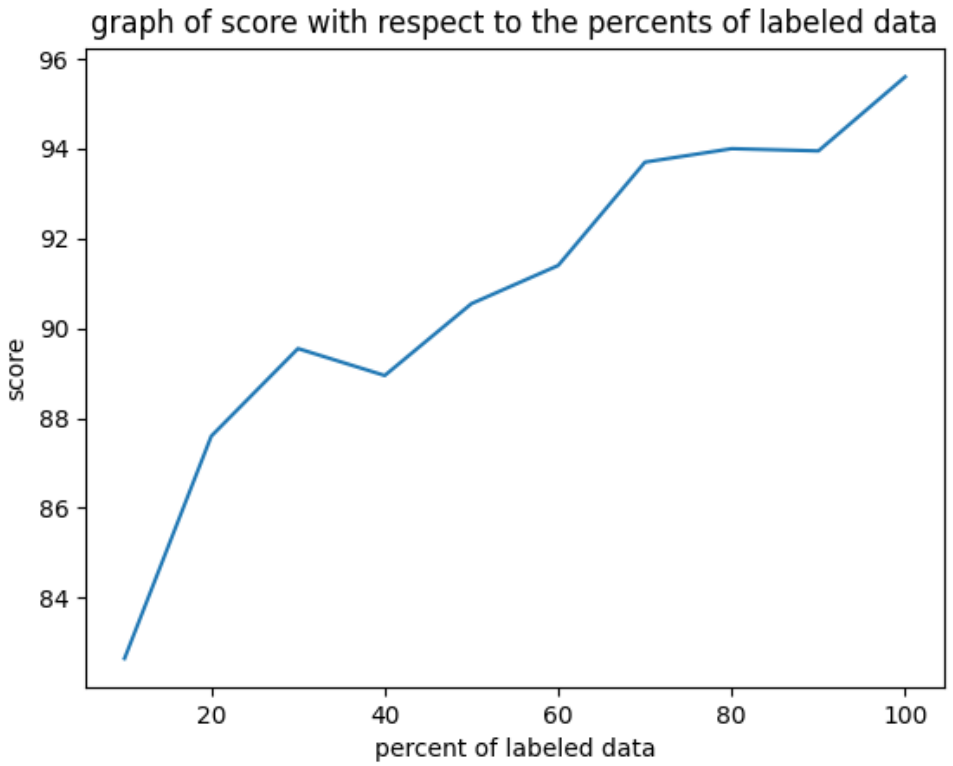
1) ביצענו PCA על מנת להוריד את הממד של המידע מ28*28 ל128.
2) הרצנו אלגוריתם למידה לא מונחית וקיבלנו חלוקה של המידע לאשכולות.
3) בחרנו בצורה שרירותית עבור אחוזים שונים מידע שיהיה מתויג, לאחר מכן נתנו תוויות חדשות לכל אשכול שקיבלנו נתנו תיוג למידע הלא מתויג בשיטת "הרוב קובע" כלומר התווית המתאימה לרוב הנקודות 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.

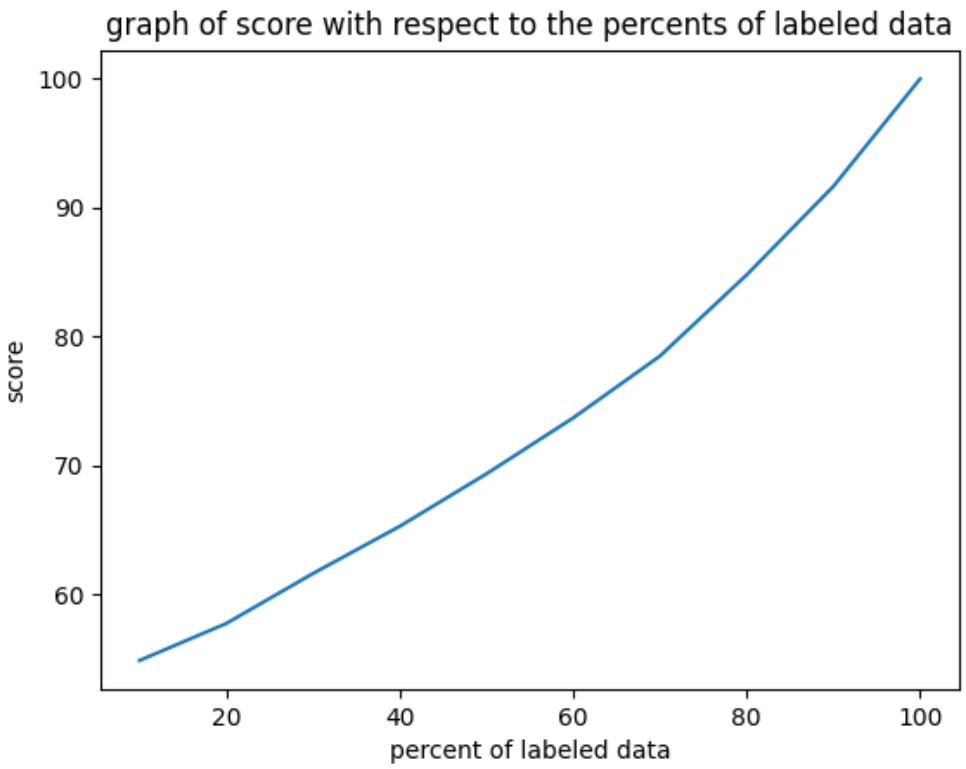
אך על מנת לבדוק אם השימוש בטכניקה משפר את התוצאות בדקנו מה ההצלחה באימון המודל רק בעזרת המידע המתויג. דגמנו כמות זהה של מידע ואימנו מודל רק באמצעות המידע המתויג. לאחר מכן בדקנו על אותו הtest set את ההצלחה של המודל. וקיבלנו את הגרף הבא:



כלומר התוצאה של אימון המודל בעזרת המידע המתויג בלבד ללא תוספת המידע שהושג בעזרת הלמידה החצי מונחית טובה יותר. הסיבה לכך היא שהתיוג של המידע החדש בעזרת ה majority rule אינה טובה.

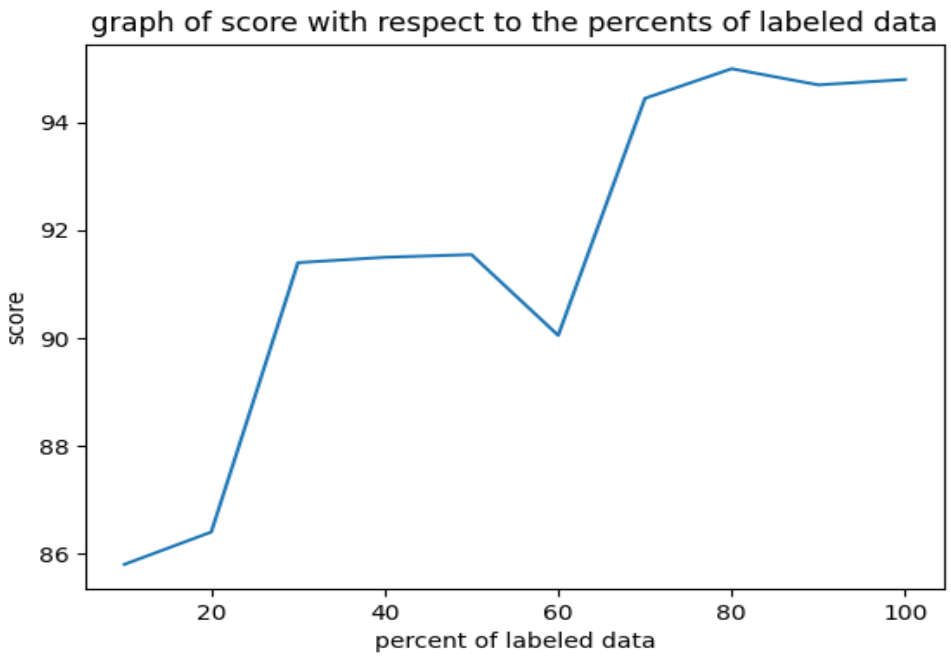
לשם כך בדקנו מה ההצלחה של הסיווג של המידע שלנו בשיטת הרוב קובע ביחס למידע האמיתי,

וקיבלנו את הגרף הבא:



קיבלנו כי השיטה של הרוב קובע צדקה בפחות מ55% מהתיוגים בלבד מה שגרם למודל להתאמן עם למעלה מ3,000 דוגמאות לא נכונות שמהוות כמעט חצי מהמודל.

לכן ניסנו להשתמש באלגוריתם הlabel propagation על מנת לבדוק אם שימוש בו ישפר את התוצאות. והתקבל הגרף הבא:



ניתן לראות שהשימוש בלמידה חצי מונחית משפר את הביצועים של הרשת ברוב המקרים ופרט כאשר יש שימוש בכמות גדולה של מידע לא מתויג.

מסקנות-

שימוש בטכניקות של למידה חצי מונחית יכולה לשפר את המודל למרות תוספת של דוגמאות לא נכונות, צריך מחקר נוסף כיצד ניתן לשפר את התיוג בעזרת המידע הלא מתויג.