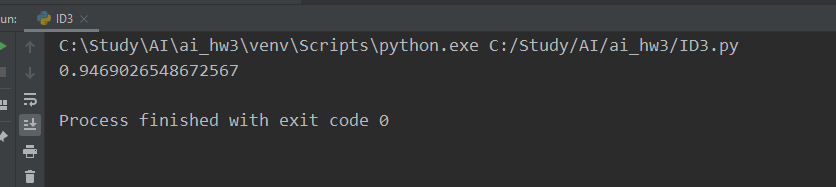
מבוא לבינה מלאכותית תרגיל בית 3

יובל גושן 205810179

# שאלה 1

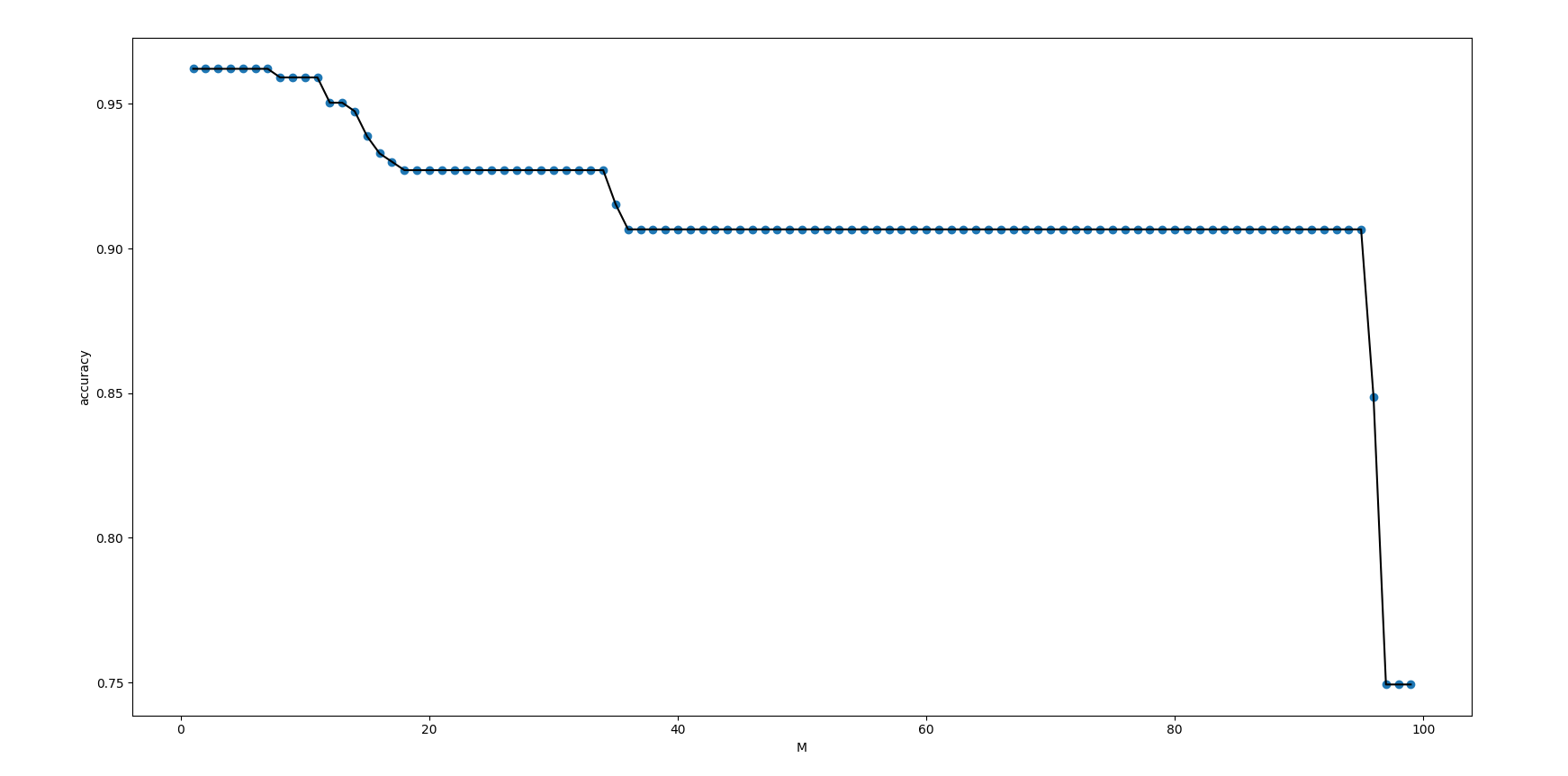
1.2 קיבלנו תוצאת דיוק של על קבוצת המבחן.  


# שאלה 2

TBD

# שאלה 3

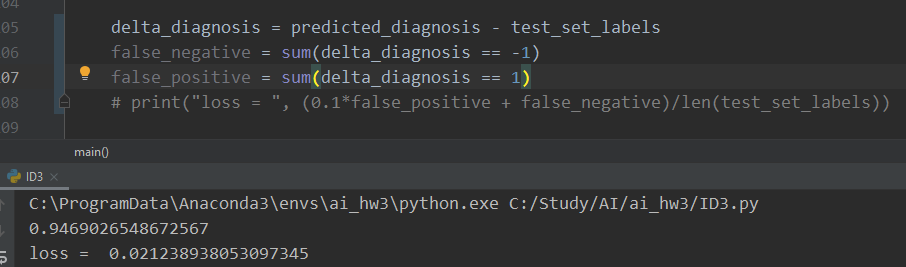
3.1. הגיזום מנסה למנוע את תופעת התאמת היתר (overfitting) בה האלגוריתם מאומן על קבוצת אימון אך לא יודע להכליל ולקבל תוצאות מספיק טובות על קבוצות אחרות. החשיבות שלו הוא שהוא מוריד עלים שהתקבלו ככל הנראה מדוגמאות שהן "רעש" ולא מייצגות את העולם באופן כללי.

3.3 מצורף הגרף, אמנם נתבקשנו לבחור לפחות 5 ערכי M שונים אבל בגלל נדיבות יתר שלי תקבלו אפילו 100 :  


קיבלנו למעשה שהתוצאה הטובה ביותר היא עבור M=1, והיא יורדת ככל שמעלים את M (בחלק מהנקודות היא לא משתנה כי אין עלה שמתפצל לפי אותה כמות דגימות).  
 כלומר ללא גיזום מוקדם כלל קיבלנו את התוצאה הטובה ביותר. התוצאה הממוצעת בין הfolds על הvalidation set היא 96 אחוז.

3.4 מכיוון שקיבלנו שהתוצאה הטובה ביותר היא ללא גיזום בכלל בסעיף זה קיבלנו שוב תוצאה של ואין שיפור בעץ.

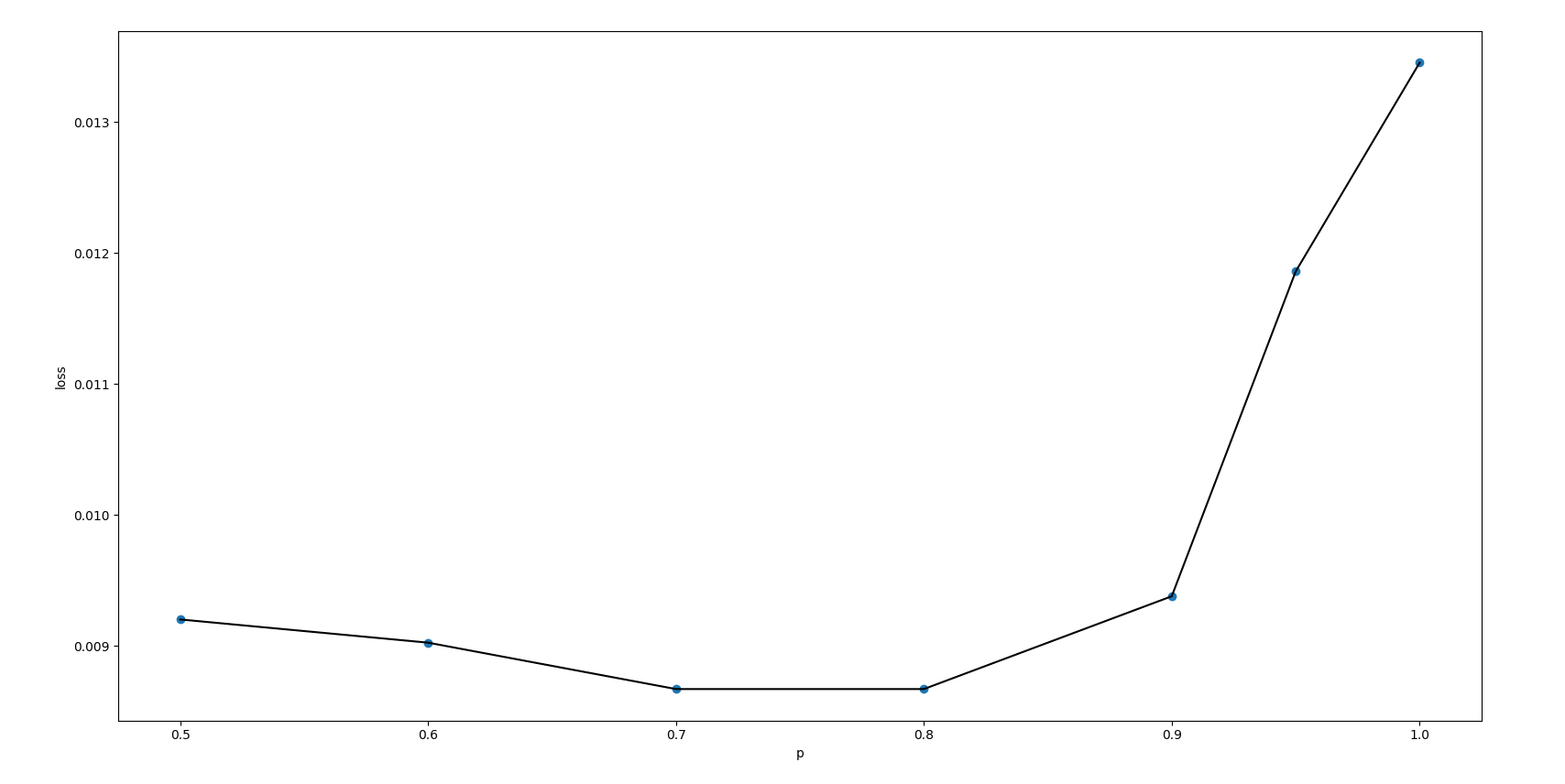
# שאלה 4

4.1 קיבלנו את ערך הloss הבא:   


0.021238938053097345

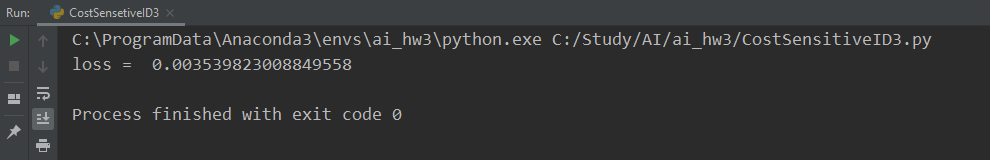
הוא נכון גם עבור "הגיזום הכי טוב" כי הגיזום הכי טוב הוא עבור M=1

4.2 נבצע החמרה בכל פיצול במקרים בהם רוב הדוגמאות הם של חולים. נגדיר פרמטר ובכל פעם שאנחנו מפצלים צומת, נבדוק אם החלק של הדגימות שערכן 1 (חולה) מכל הדגימות גדול מp, אם כן הצומת הרלוונטי יהפוך לעלה שנותן חיזוי של חולה. כך בעצם האלגוריתם מחמיר ונותן חיזוי של חולה לכל קבוצה בה יש רוב משמעותי של חולים ולא ממשיך לפצל אותה מתוך הנחה שיש הסתברות שהפיצול יתבסס על דגימות בודדות שנובעות מרעש.

4.3 ביצענו תהליך cross validation עם k-fold=5 בדומה לשאלה הקודמת על הערך p שהגדרנו. תוצאות הניסוי :

קיבלנו שהp הטוב ביותר הוא 0.7 או 0.8, בחרנו להשתמש ב0.8 ושיפרנו משמעותית את הloss.

קיבלנו:

**

*loss = 0.003539823008849558*