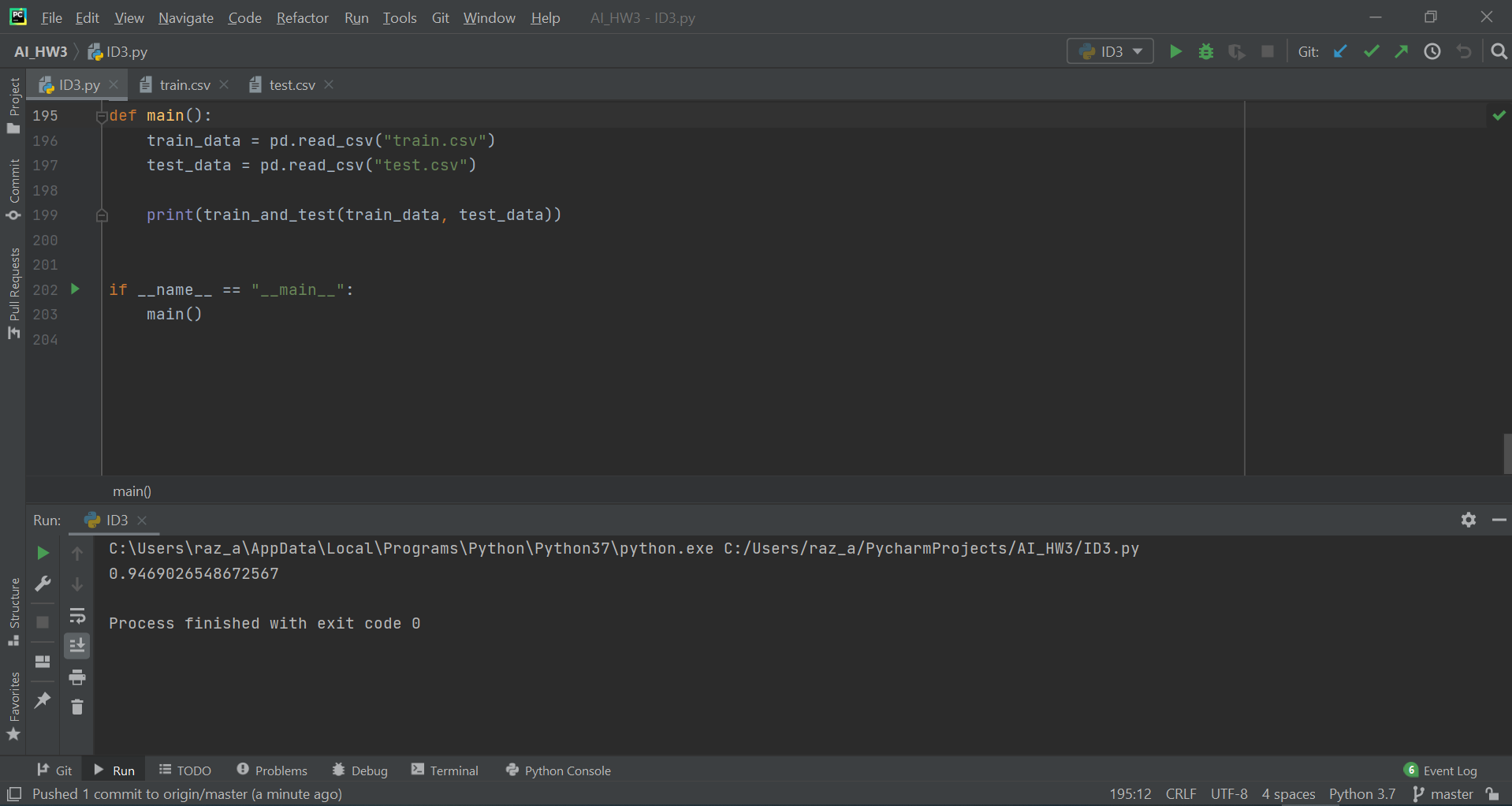
**תרגיל בית 3 – חלק יבש**

**שאלה 1**

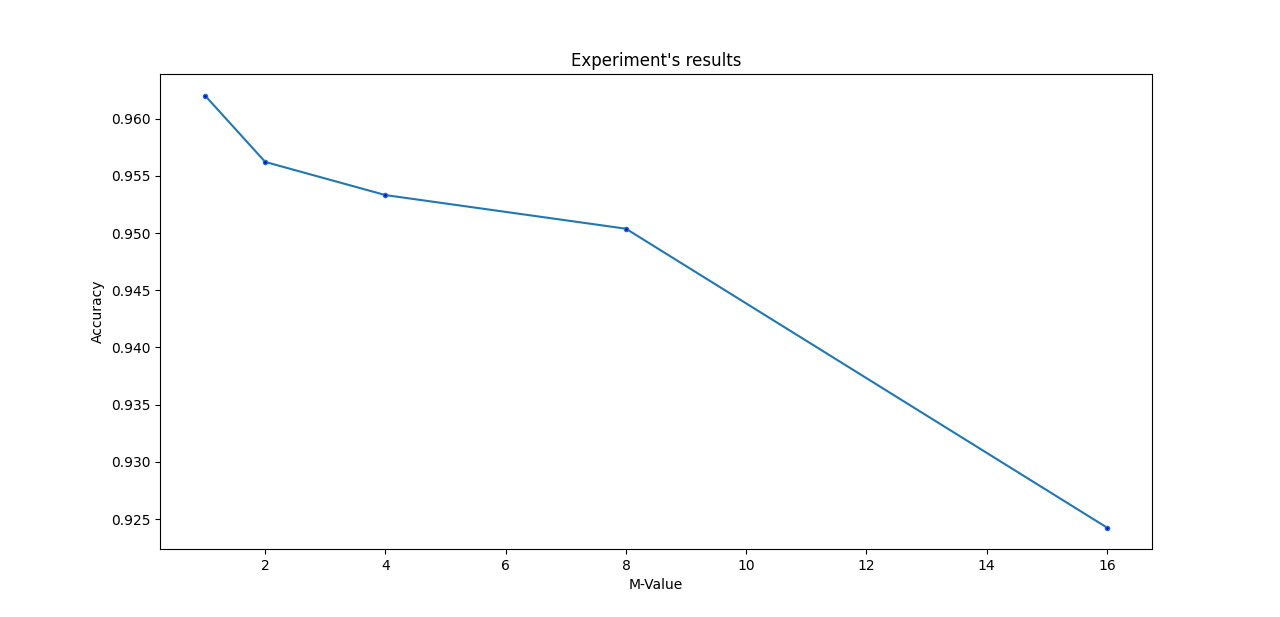
להלן צילום מסך של תוצאת הדיוק (התוצאה היא 0.9469~):



**שאלה 2**

(להשלים)

**שאלה 3**

1. הגיזום באופן כללי נועד כדי להקטין את העצים, בין אם זה לפני או אחרי בנייתם, במטרה להחליש את אפקט התאמת היתר (overfitting), שזו תופעה שאנו מנסים למנוע באופן זה שנשאף להגדיל את שגיאת האימון בשביל להקטין את שגיאת המבחן.
2. להלן הגרף המציג את השפעת הפרמטר M על הדיוק עבור M=1,2,4,8,16:  
   

כפי שניתן לראות, הגרף נמצא במגמת ירידה, מה שמצביע על כך שככל שנגדיל את הפרמטר M כך נקבל דיוק נמוך יותר. זה הגיוני מכיוון שעבור M גדול יותר האלגוריתם אמור לגזום יותר תתי עצים ובעקבות זה יגדיל את שגיאת האימון. כפי שניתן לראות, הM האופטימלי הוא 1 עם דיוק של 0.962~, שזה הדיוק הגבוה ביותר עם שגיאת האימון הנמוכה ביותר. עבור M=1 אין גיזום כי עבור דוגמא אחת תמיד נקבל צומת שלפי האלגוריתם יחזיר מיד את הסיווג שלו.

1. עבור הערך האופטימלי M=1 קיבלנו בדיוק את אותה התוצאה משאלה 1.

**שאלה 4**

1. מימשתי את הפונקציה get\_loss() אשר מחזירה את ערך ה-loss של ID3, וערך ה-loss שיצא מהרצת הפונקציה עם M=1 לפי סעיף 3.4 הינו 0.021238938~.
2. מכיוון שזה מאוד חמור שאדם חולה יסווג כבריא מבחינת ה-loss, אז חשבתי על לנסות למזער את ההפרש בין מספר הדוגמאות של אנשים בריאים לעומת אנשים חולים בכל צומת כאשר יש יותר דוגמאות של בריאים מחולים, וכך נוריד את ההסתברות לסיווג אדם חולה כבריא. ניתן לממש זאת ע"י גיזום מוקדם עם היפר-פרמטר N שמוגדר להיות מספר מינימלי של הפרש דוגמאות בין בריאים לחולים, והאלגוריתם המשופר יגזום רק כאשר יש יותר דוגמאות של בריאים מדוגמאות של חולים.
3. לאחר מימוש דרך זו, ביצעתי ניסויים לקביעת פרמטר N האופטימלי מN=1 ועד 100, ואלו התוצאות שקיבלתי:

|  |  |
| --- | --- |
| loss בקירוב | N |
| 0.021238938 | 1-2 |
| 0.022123893 | 3-7 |
| 0.004424779 | 8-9 |
| 0.002654867 | >=10 |

לכן עבור N=10 נקבל את האלגוריתם המשופר עם ערך אופטימלי loss=0.002654867, שהוא נמוך משמעותית מזה שקיבלנו בסעיף 1.