מבוא לראייה חישובית ממ"ן 12

מאת: יאיר חריט 207282955

התוכניות תחילה טוענות את הקובץ config.py המכיל את כל הנדרש על מנת להפעיל את האלגוריתמים ולסכם את הנתונים שייצרו- ייבוא מסדי הנתונים, סידורם, אתחול משתנים, יתר על כן, קובץ זה שומר על אחידות ומונע כפל קוד.

לאחר אתחול המשתנים, נפעיל את האלגוריתם הנבחר על 80% מהנתונים, האלגוריתם יחשב ויסווג את הנתונים. את התוצאות שהתקבלו נשווה מול המידע הקיים במסד הנתונים על מנת לקבוע את טיב האלגוריתם.

עבור אלגוריתם kNN נאסוף את התוצאות עבור ערכי ה-k השונים עד max\_k ולבסוף נציג את המידע הנאסף בגרף לשם השוואה.

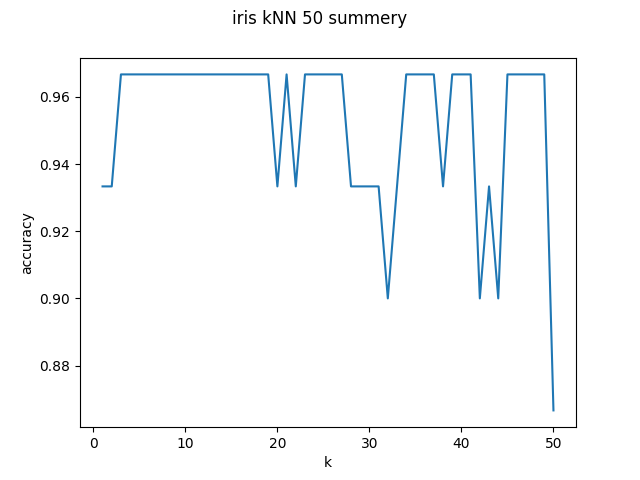
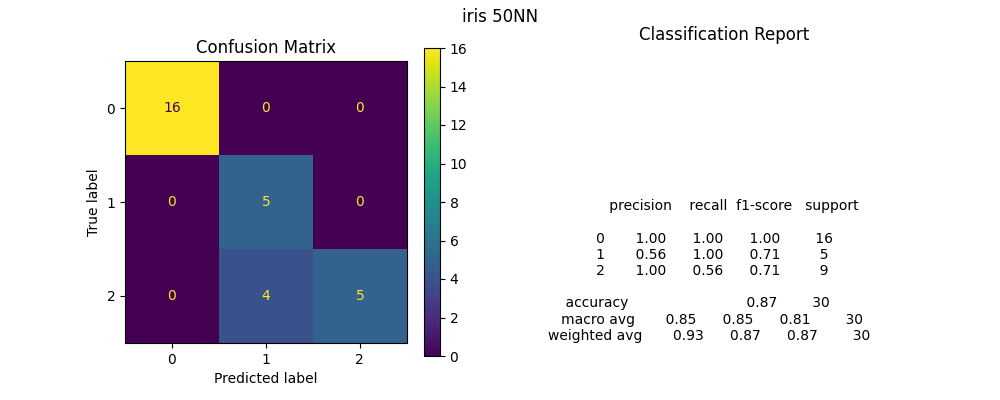
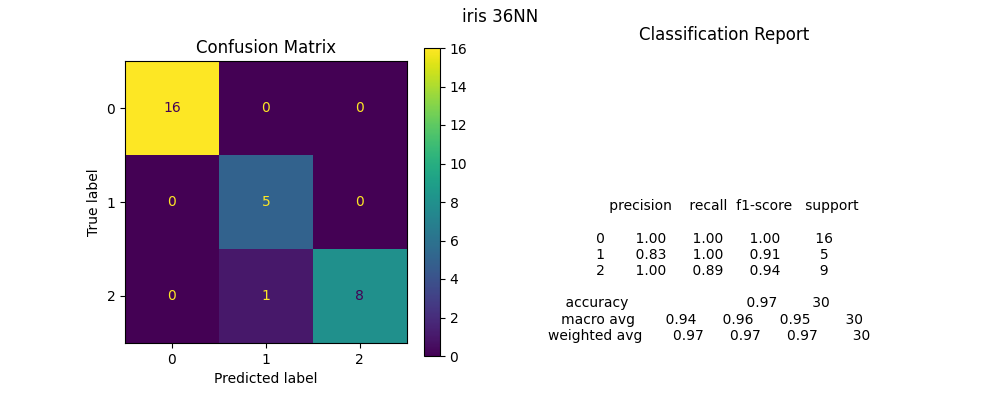
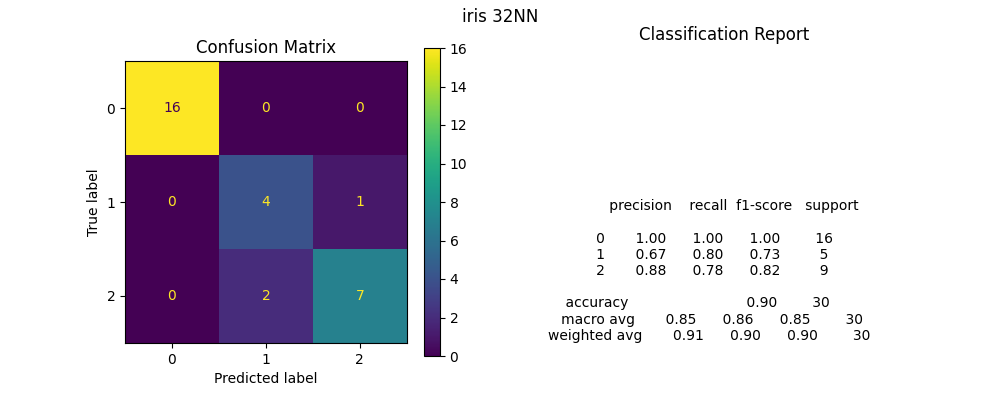
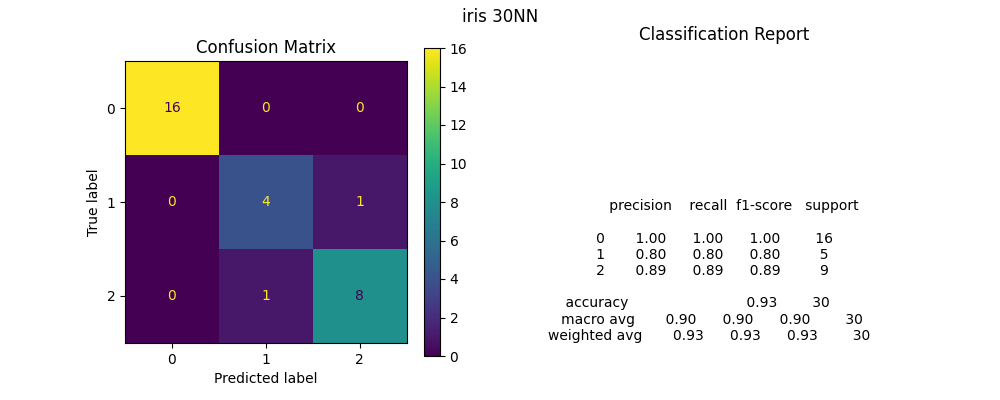
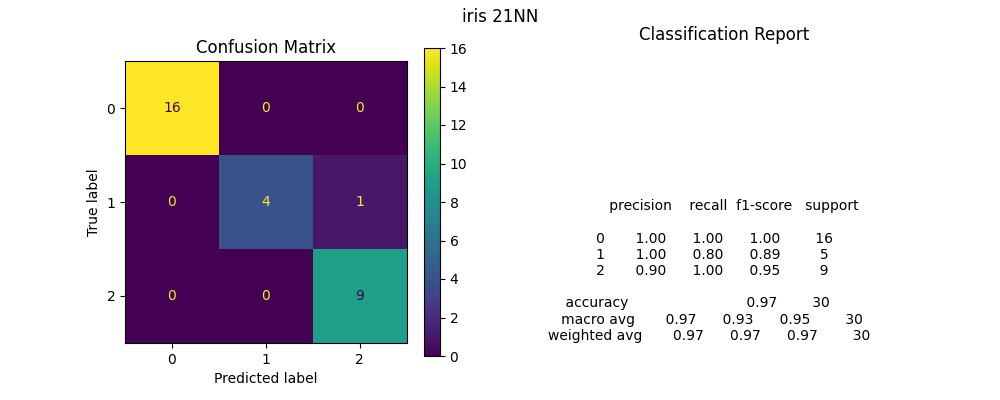
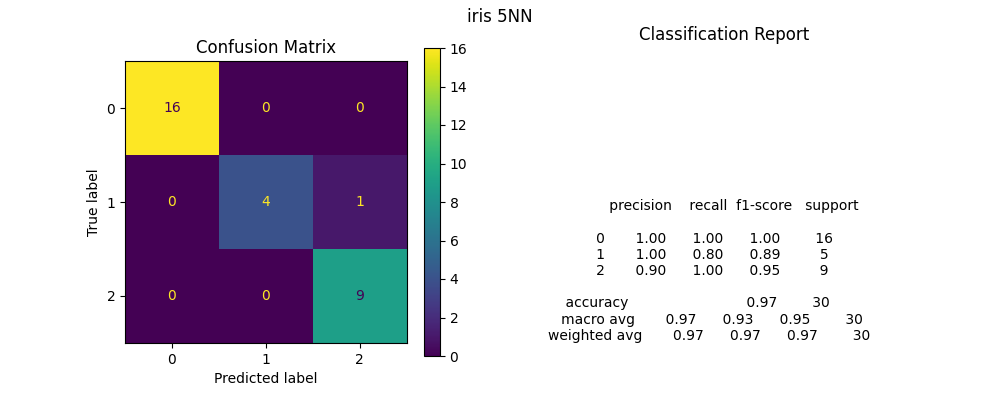
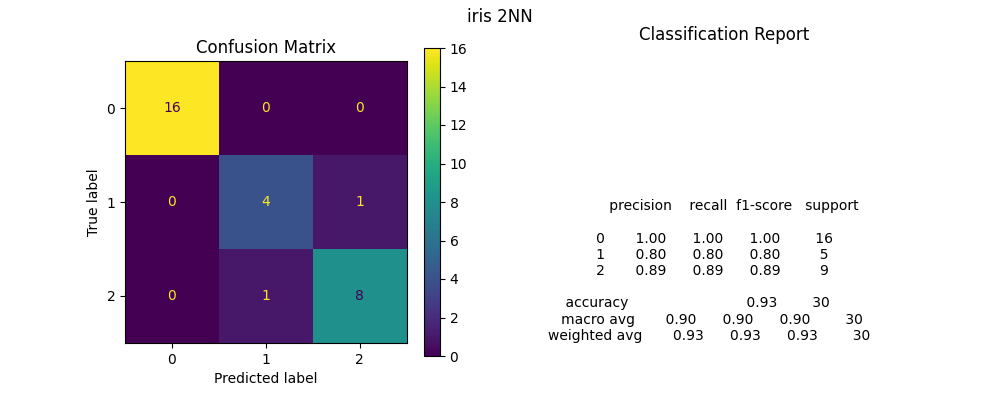
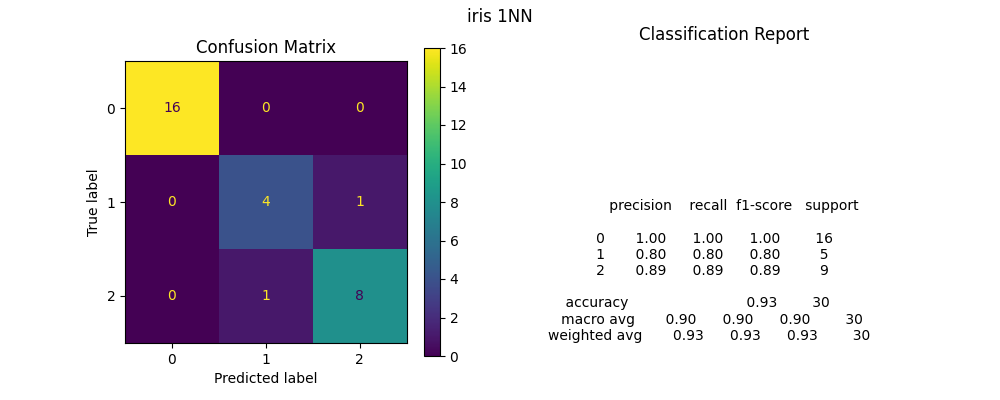
עבור אלגוריתם(י) SVM נאסוף את התוצאות עבור כל מסווג בערכי C שונים ולבסוף נשווה את הנתונים הללו בעזרת גרפים.

על מנת לבדוק את הקוד ובכדי לשמור על זמן עיבוד הגיוני צמצמתי את MNIST ל10,000 דוגמיות, המידע הנתון מטה חושב עבור חצי ממסד הנתונים- 30,000 דוגמיות.

להלן חלק מהנתונים שנאספו, שאר הנתונים נמצאים בתיקיית figures.

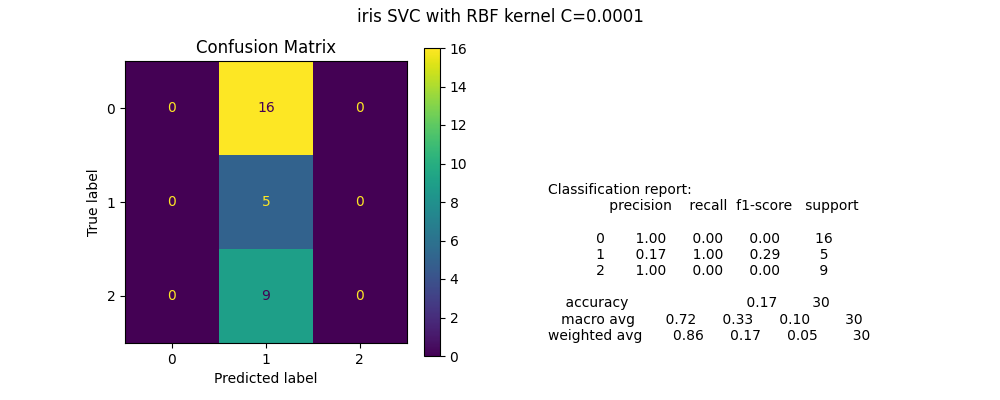
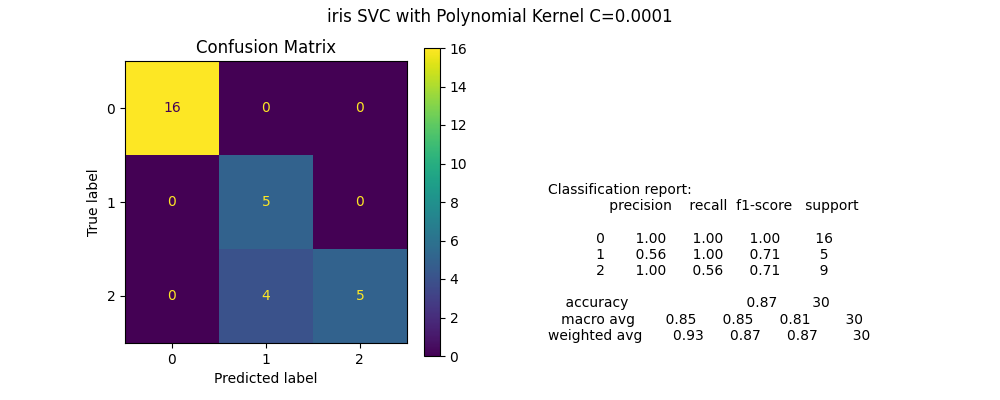
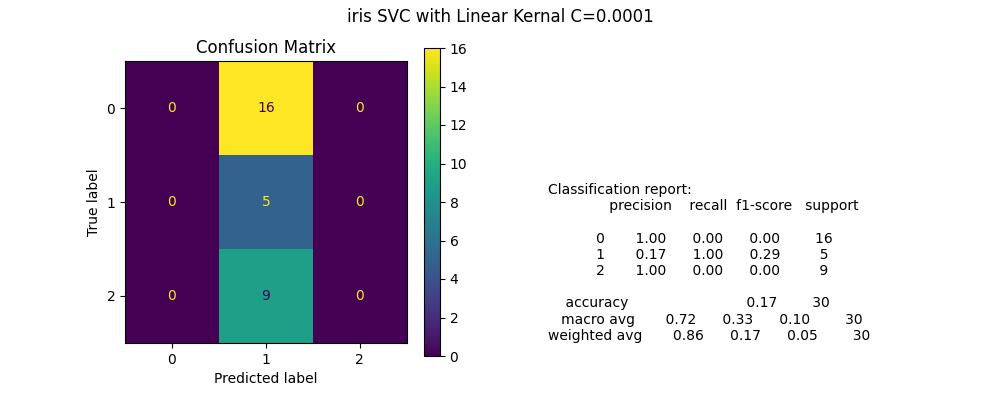
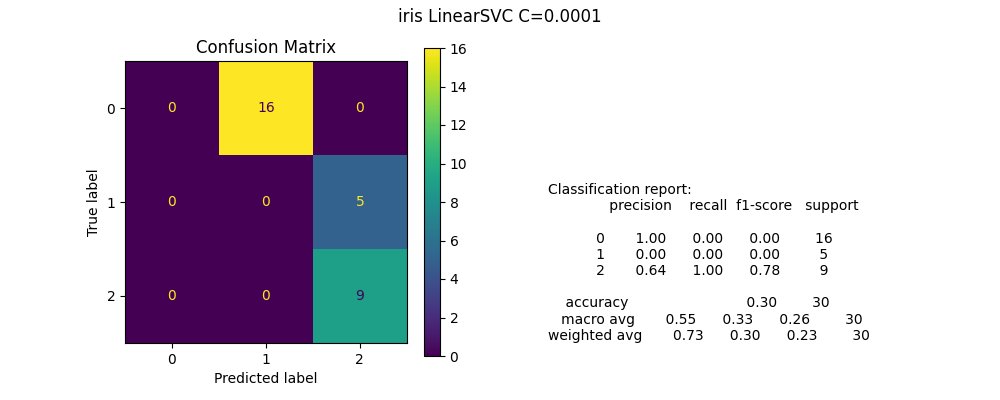
IRIS Dataset

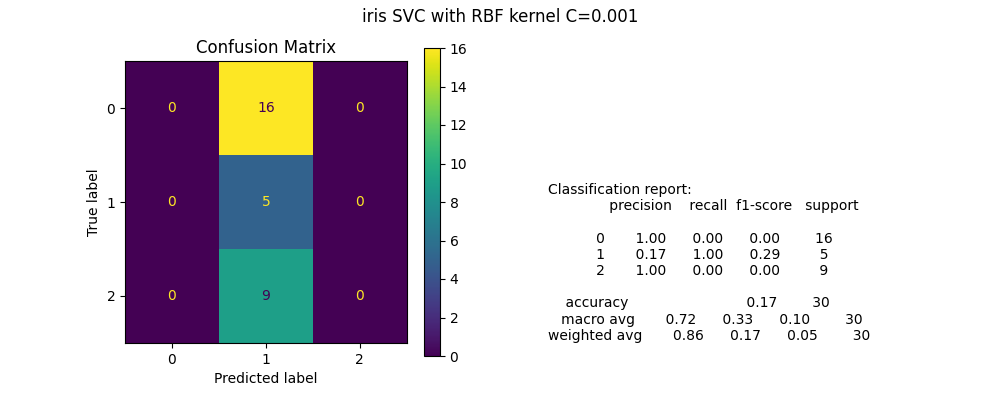
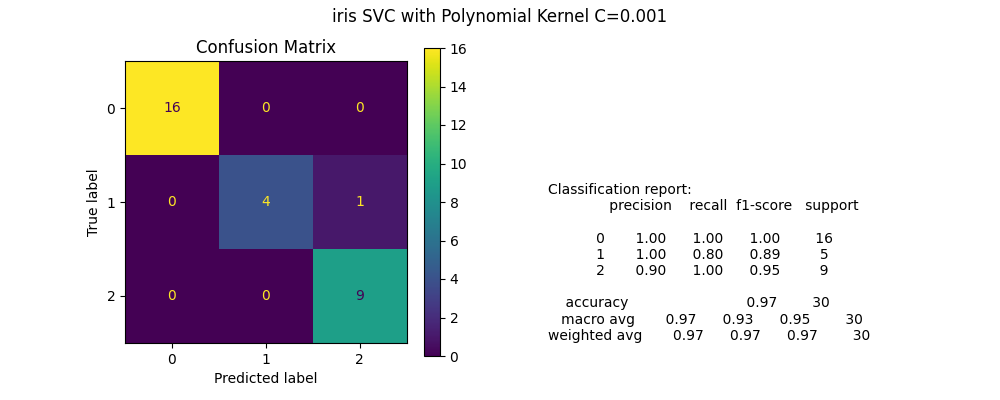
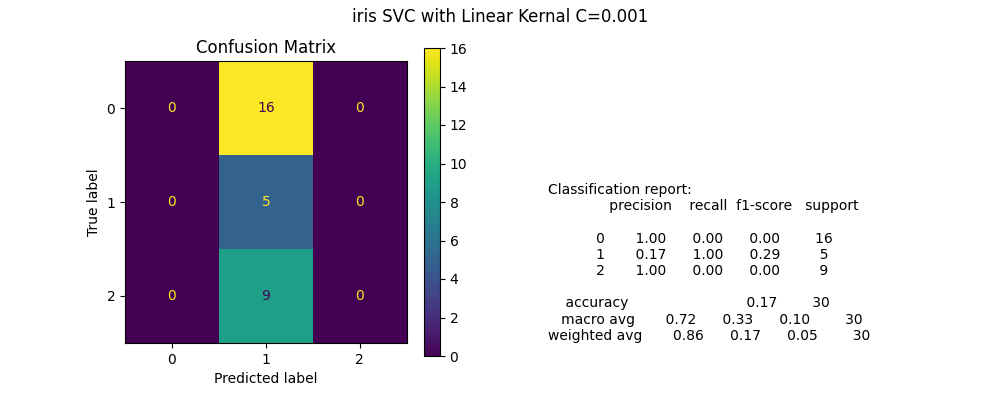
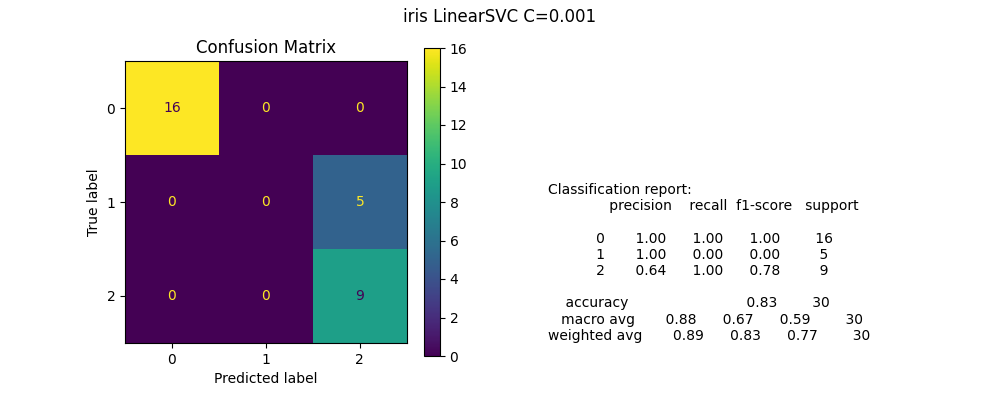
kNN

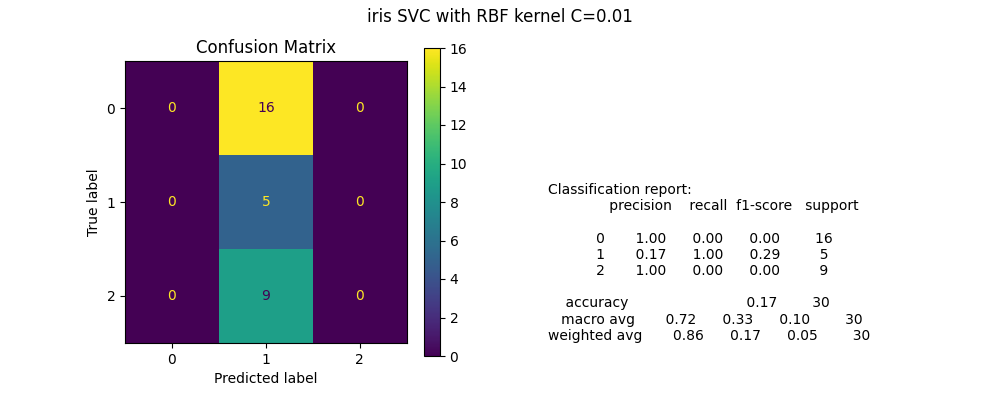
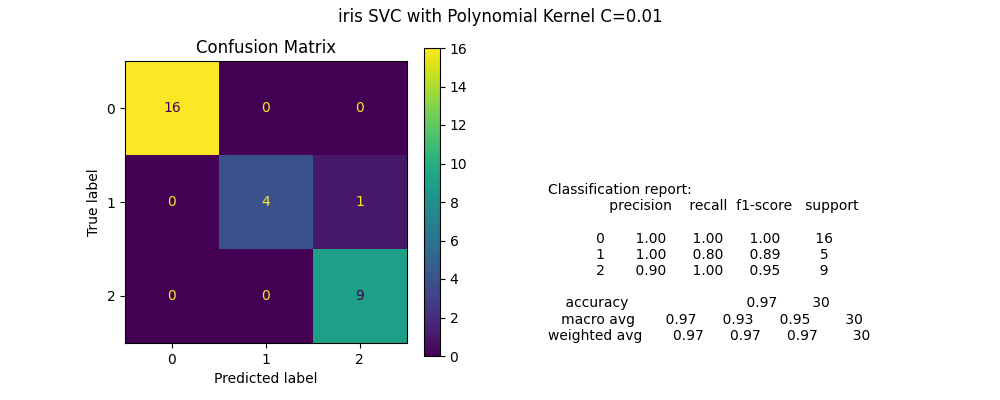
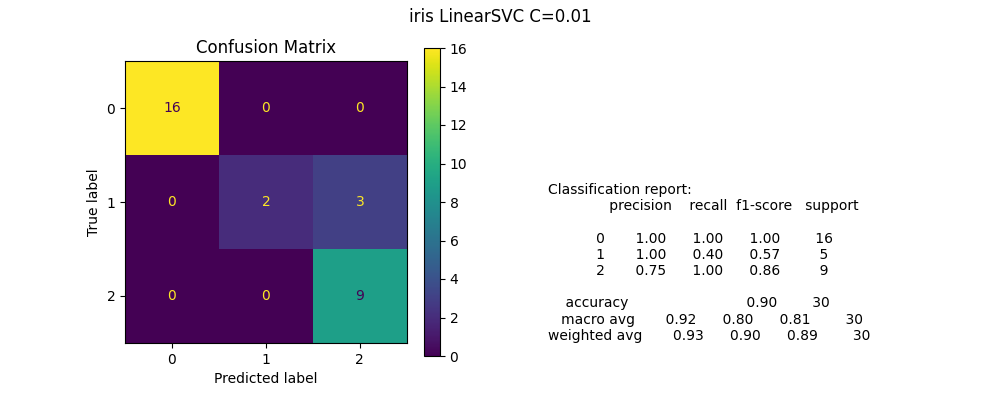


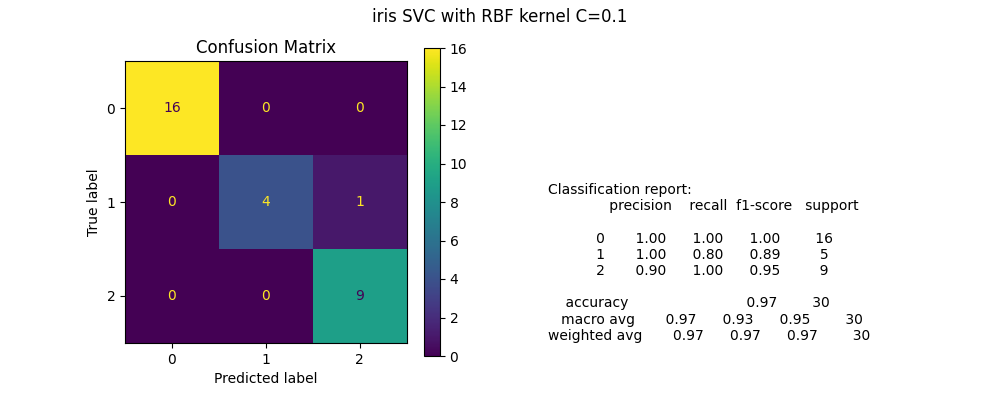
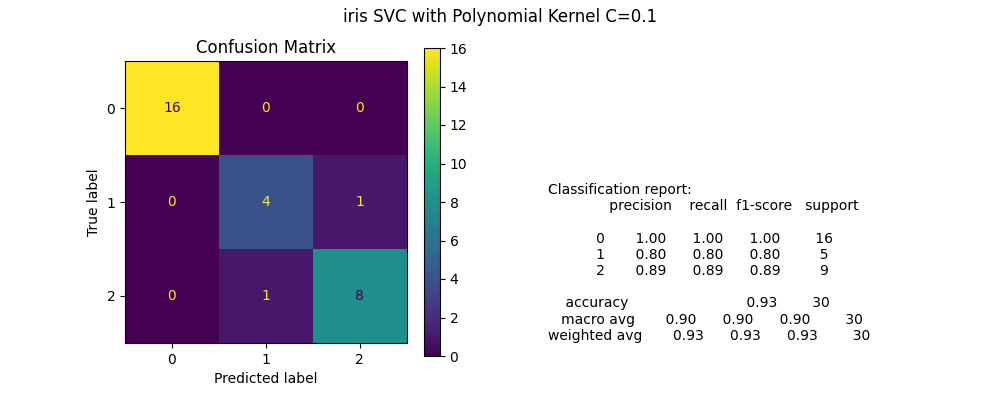
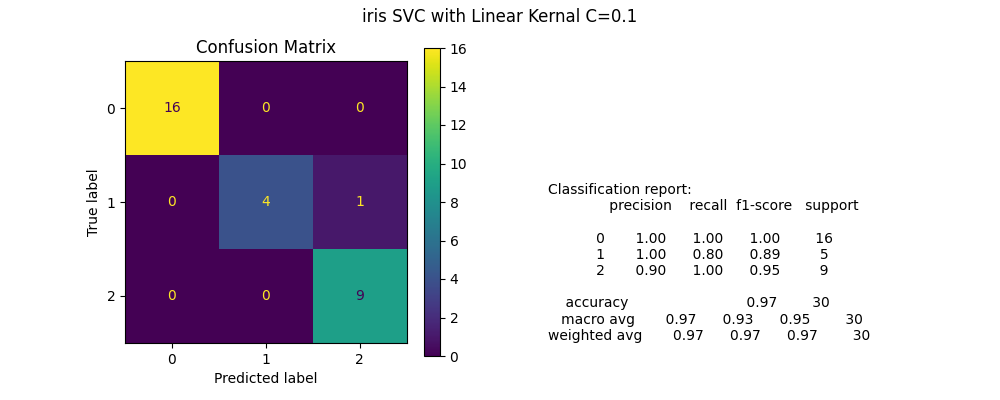
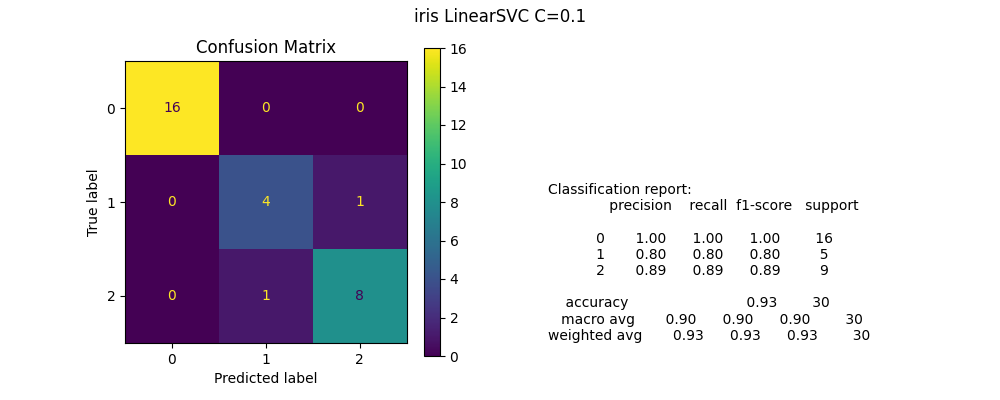
מן הנתונים הללו ניתן להסיק כי מספר השכנים האופטימלי נמוך, שכן כבר כאשר מסווגים בעזרת k=1 רמת הדיוק היא 0.93. יתר על כן, ניתן להגיד בפשטנות כי ככל שמספר השכנים גדל רמת הדיוק יורדת.

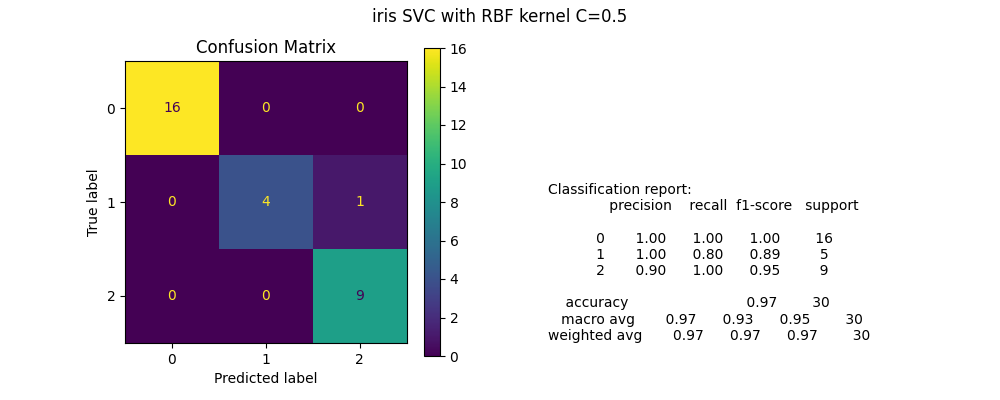
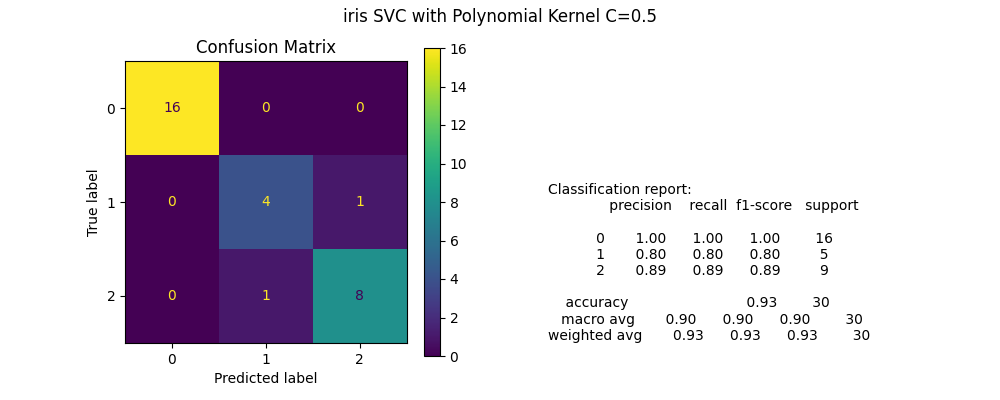
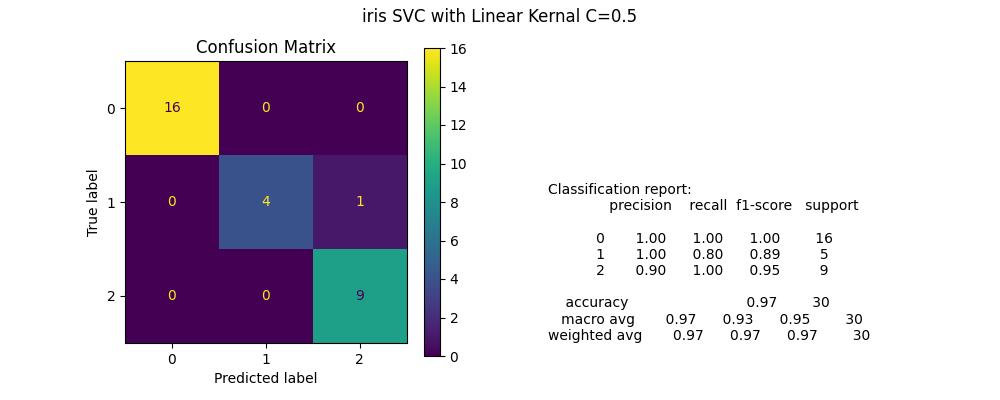
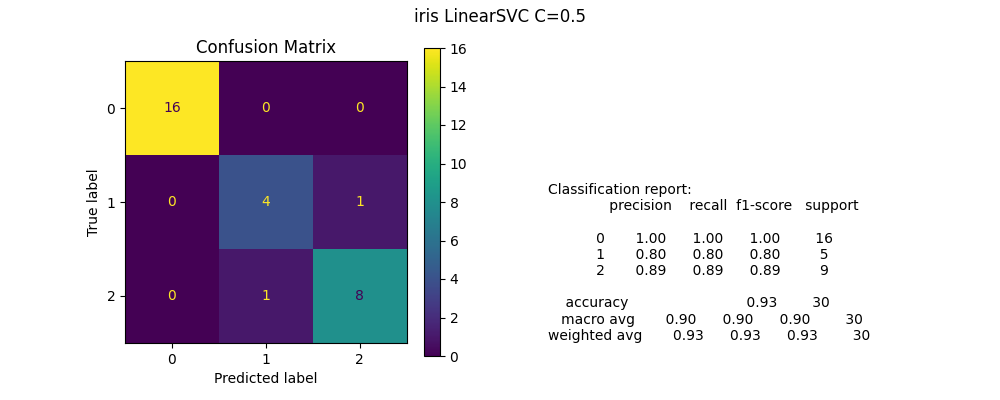
SVMs:

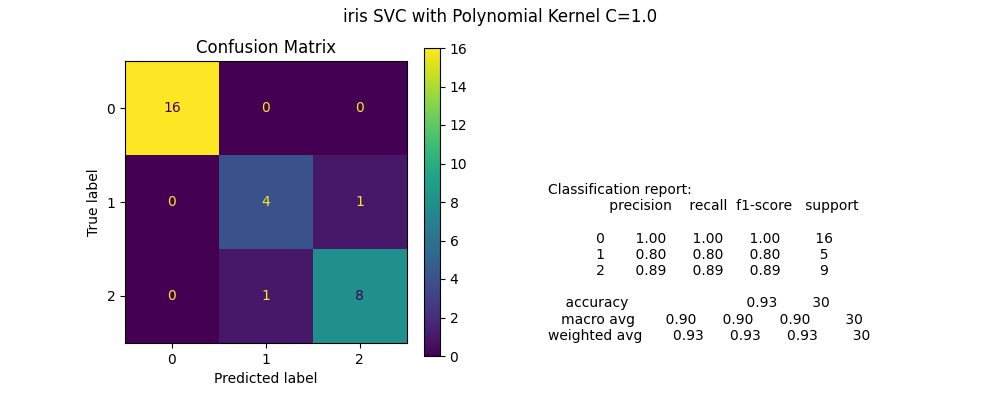
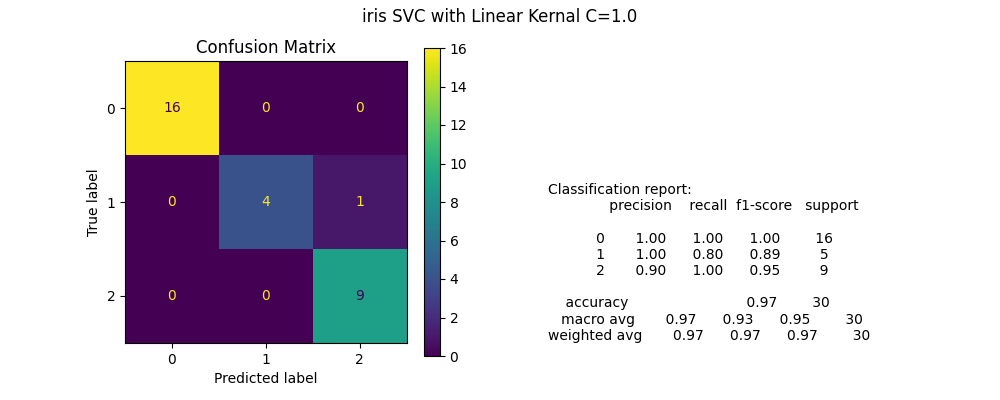
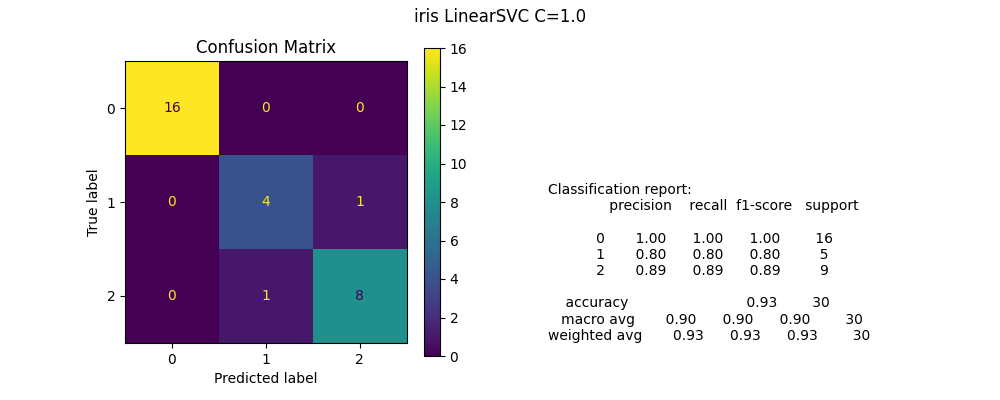


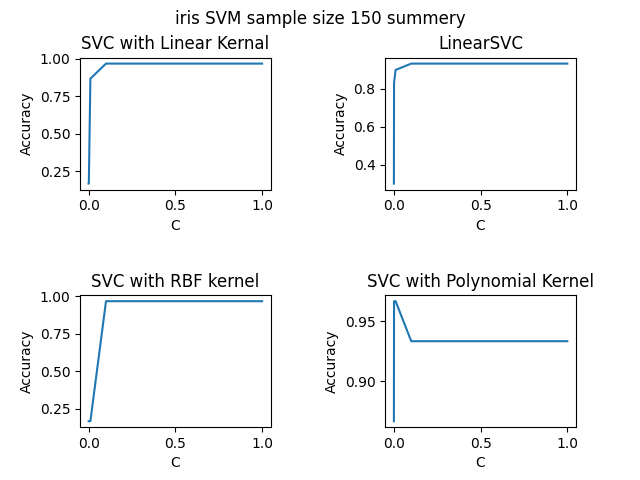












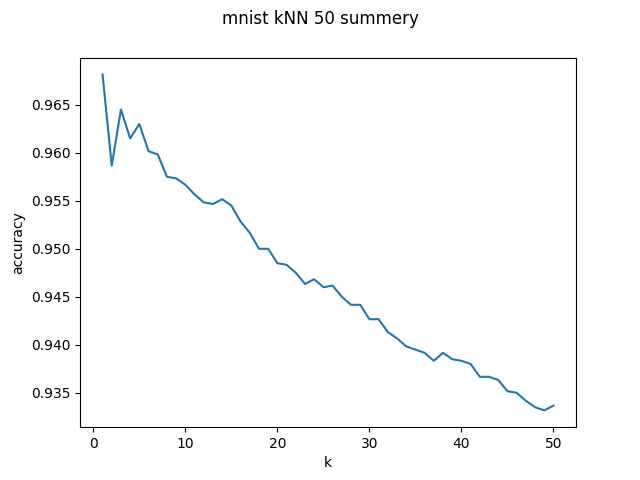
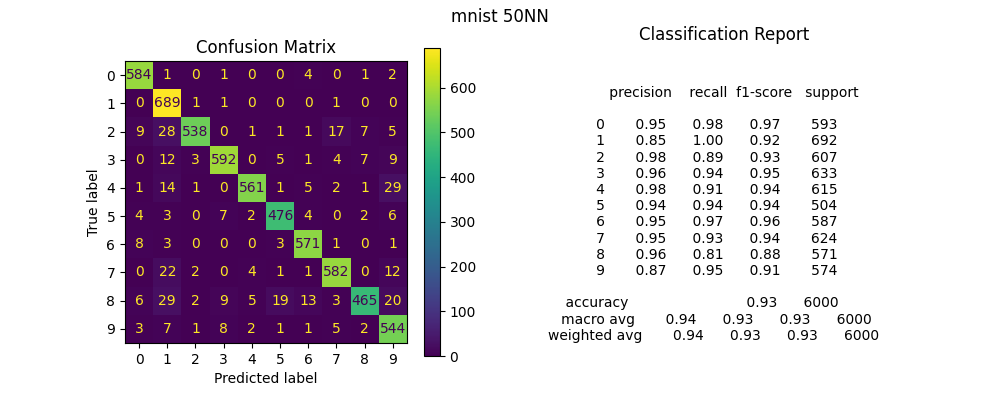
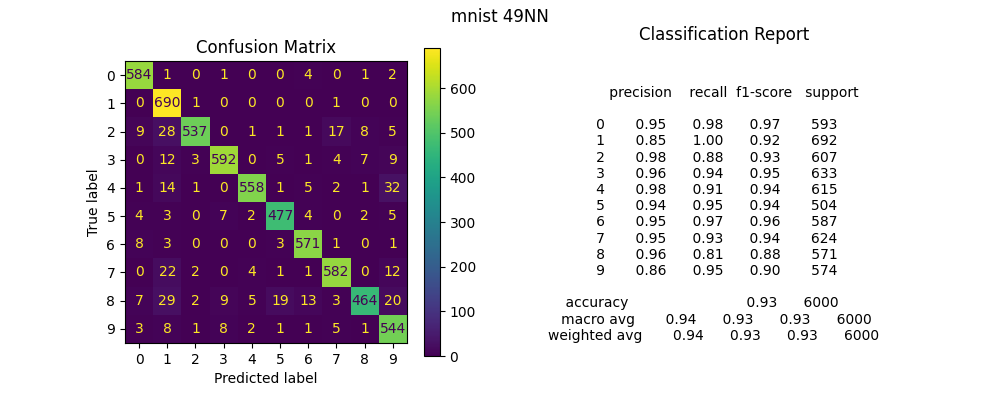
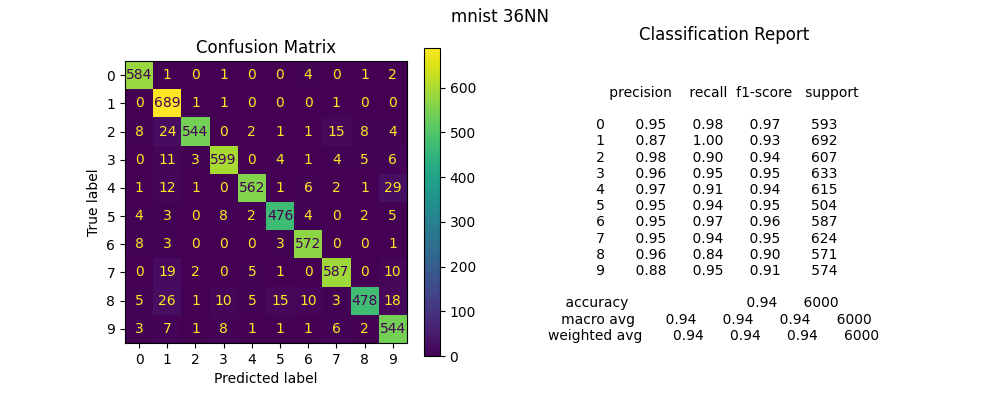
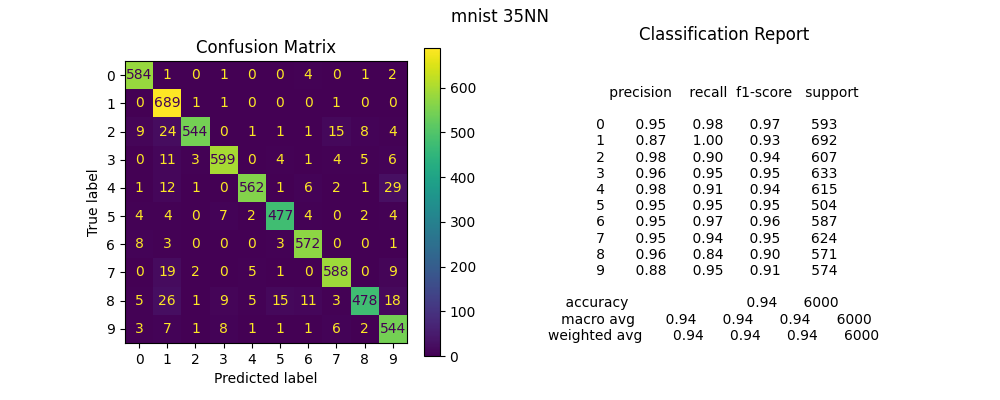
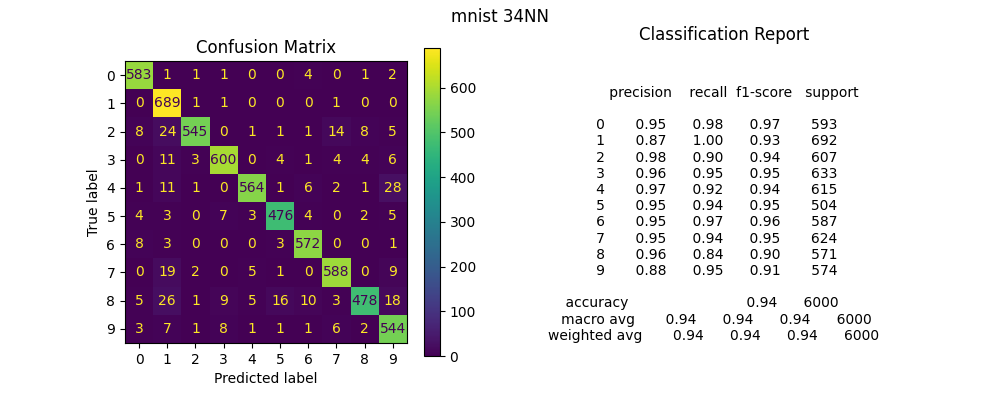
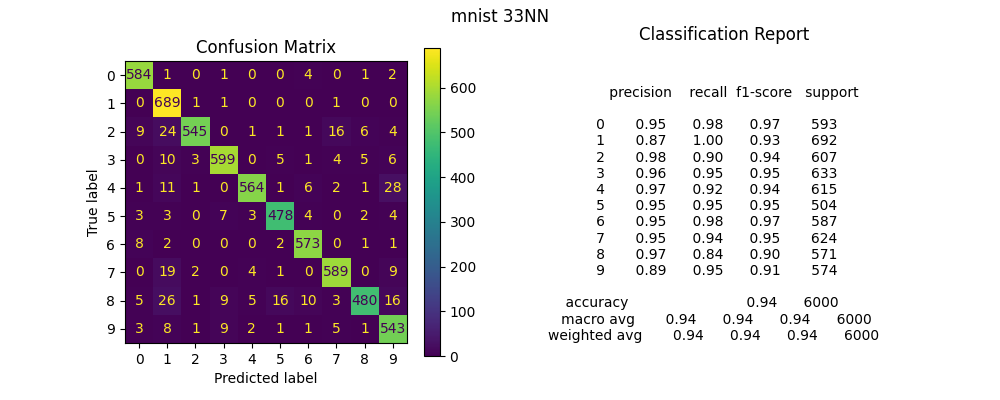
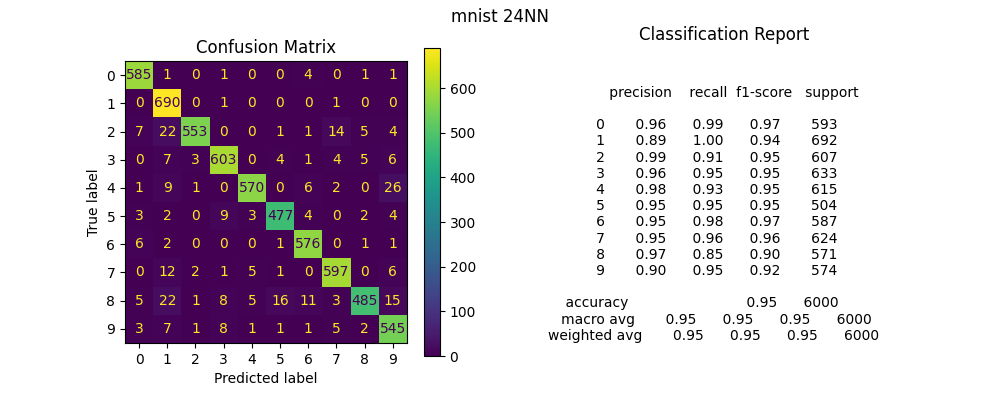
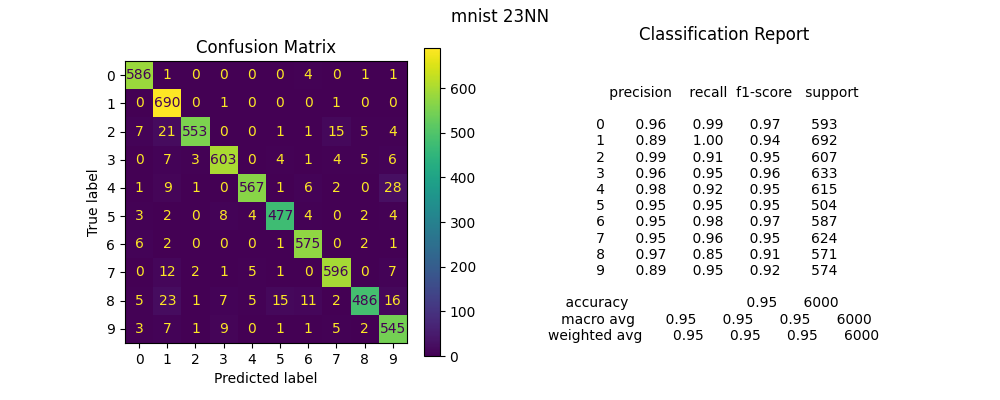
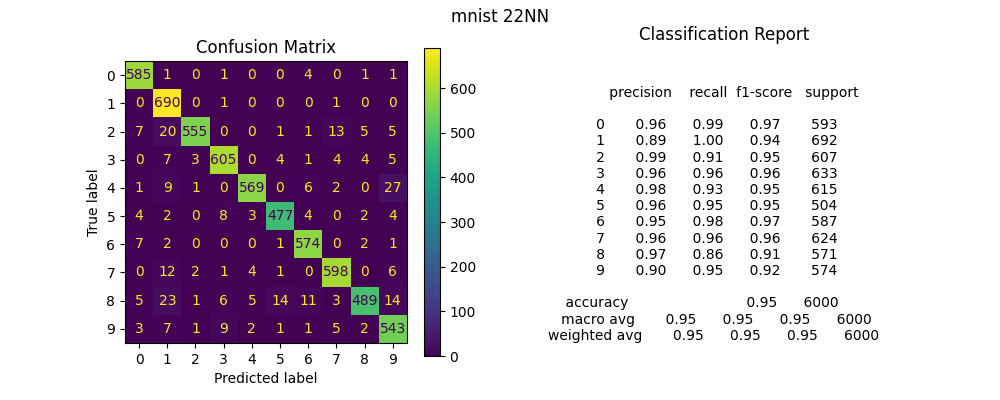
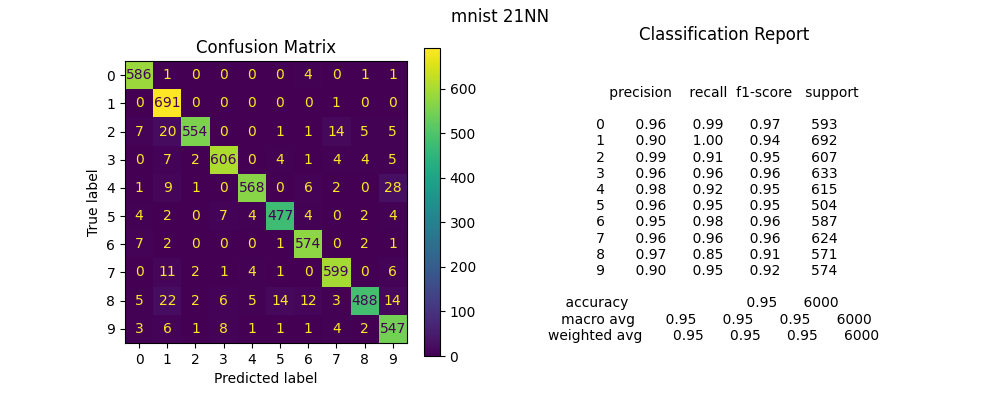
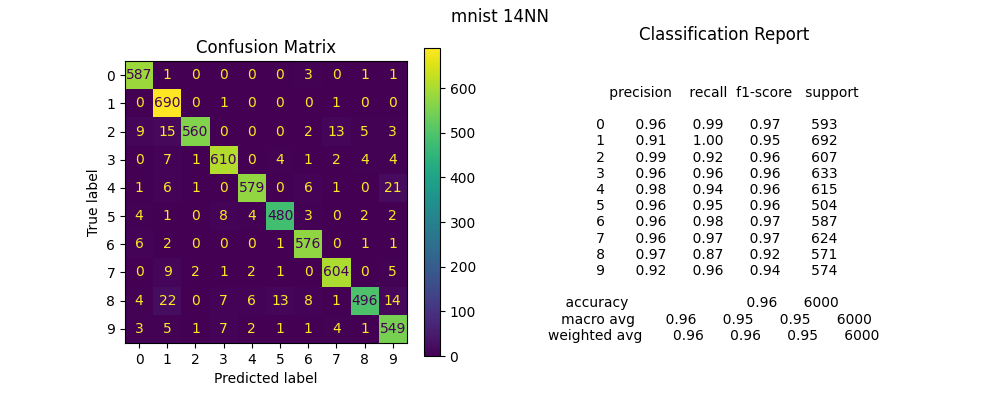
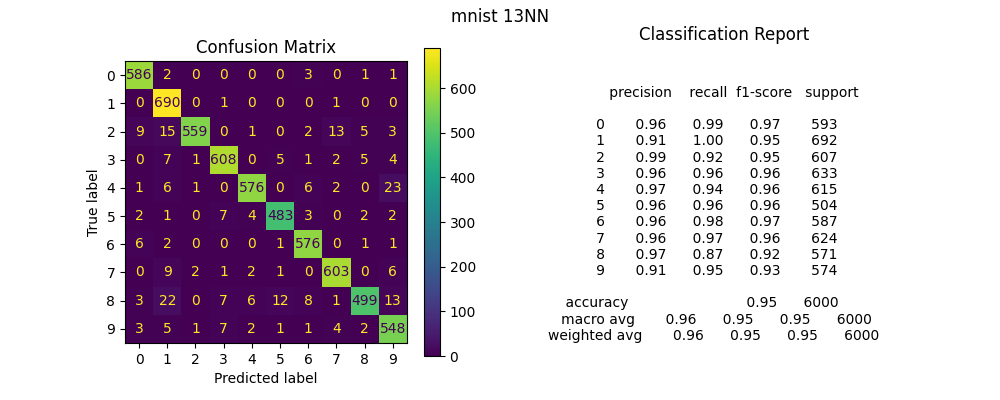
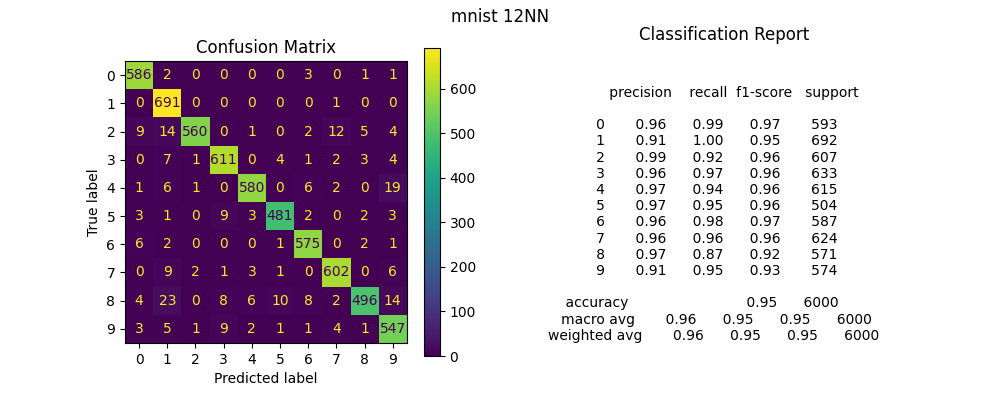
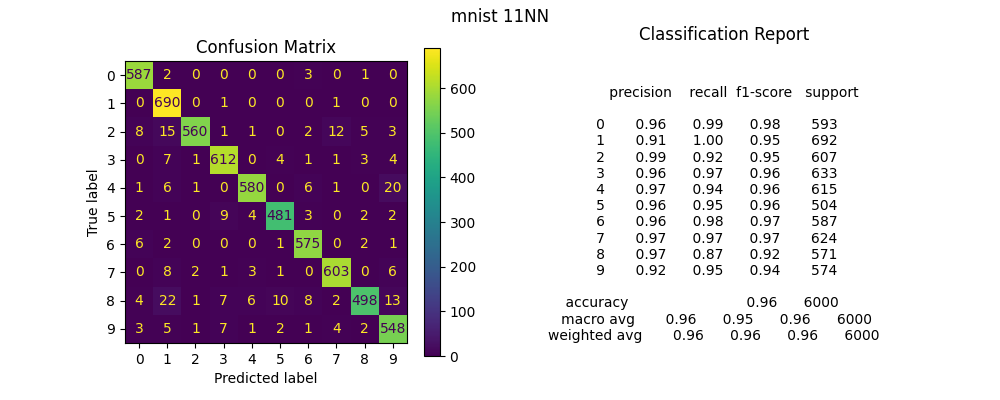
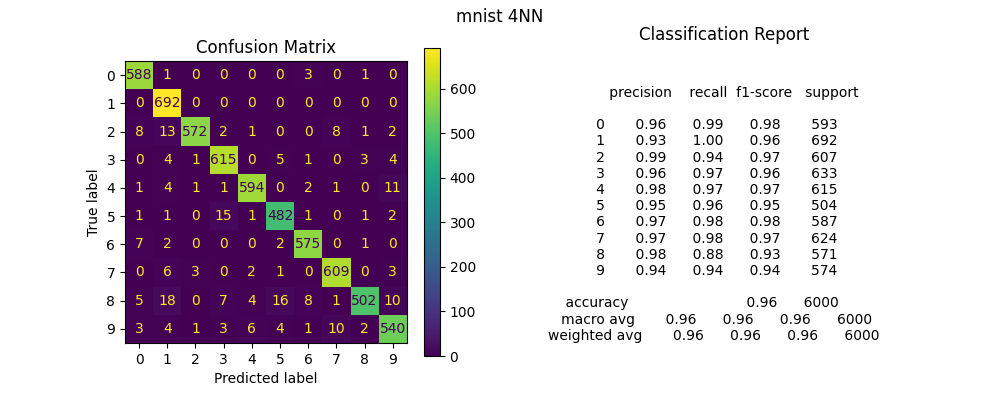
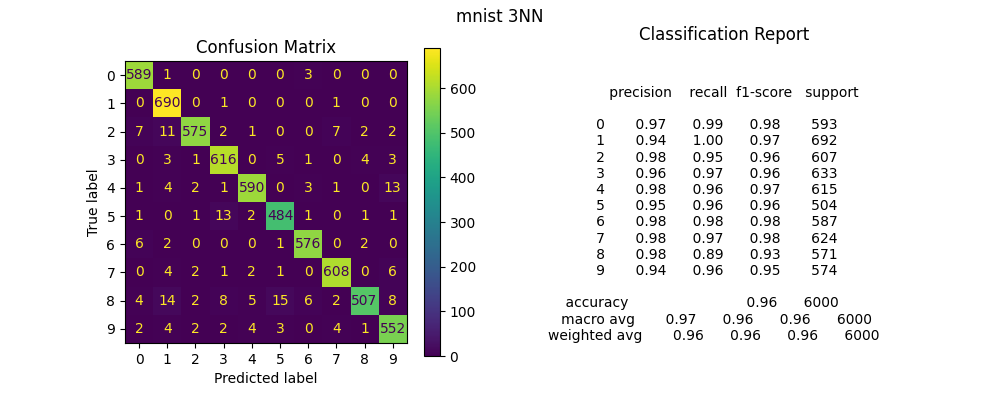
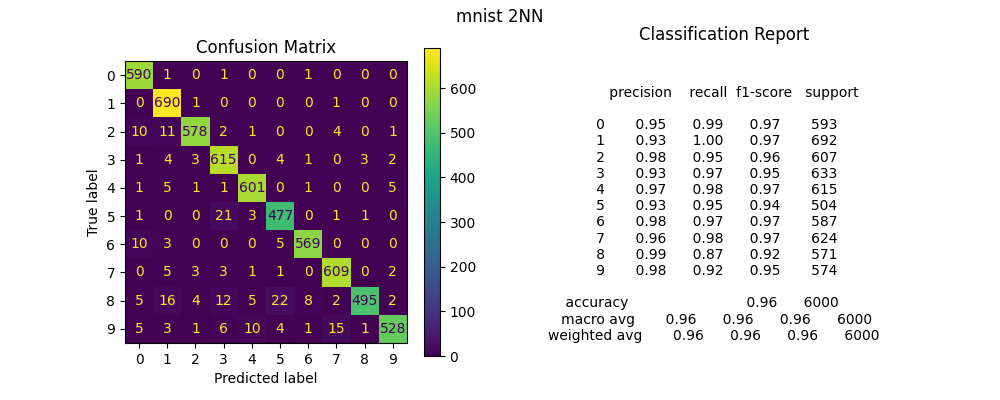
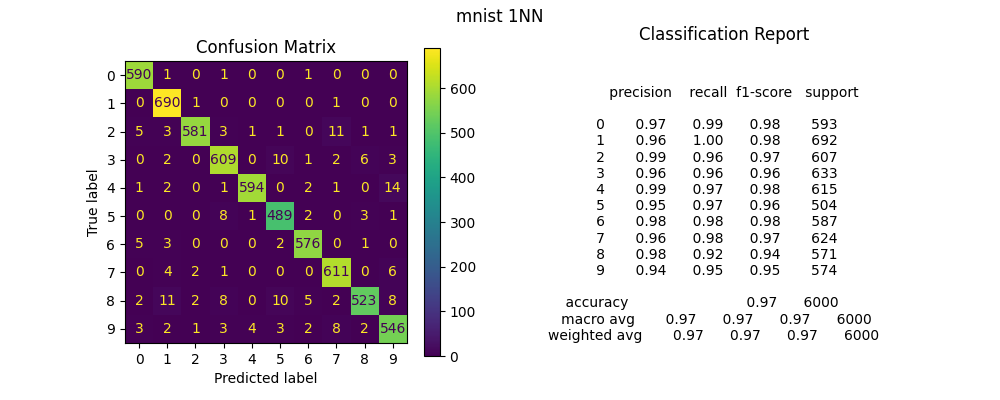
מן נתונים אלו ניתן להסיק כי מעבר ל0.1, במקרה זה, המשקל של הערך של C אינו משפיע רבות על רמת דיוק האלגוריתמים.

יתר על כן, ניתן לראות כי כאשר C=1, 4 האלגוריתמים השונים נותנים תוצאות ברמת דיוק דומה.

לסיכום, ניתן להסיק כי האלגוריתמים שנבדקו יסווגו את מסד הנתונים של IRIS ברמת דיוק זהה לKNN, מאחר וזמן העיבוד של KNN טוב יותר- הוא האלגוריתם המתאים יותר במקרה זה.

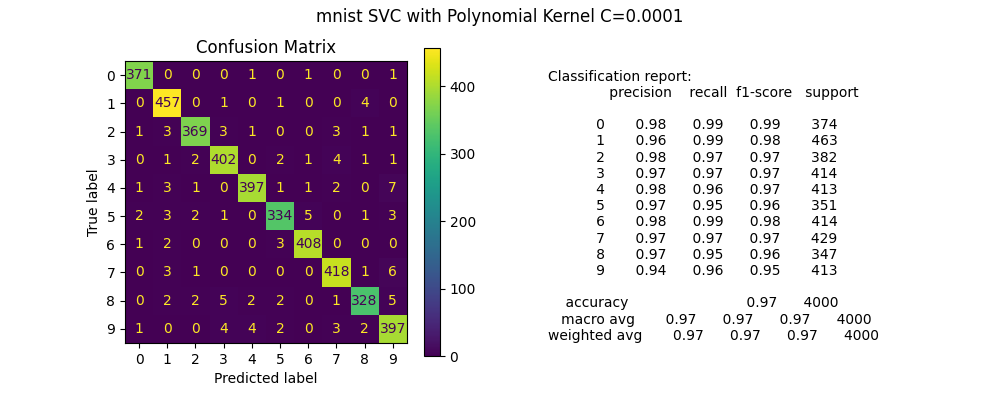
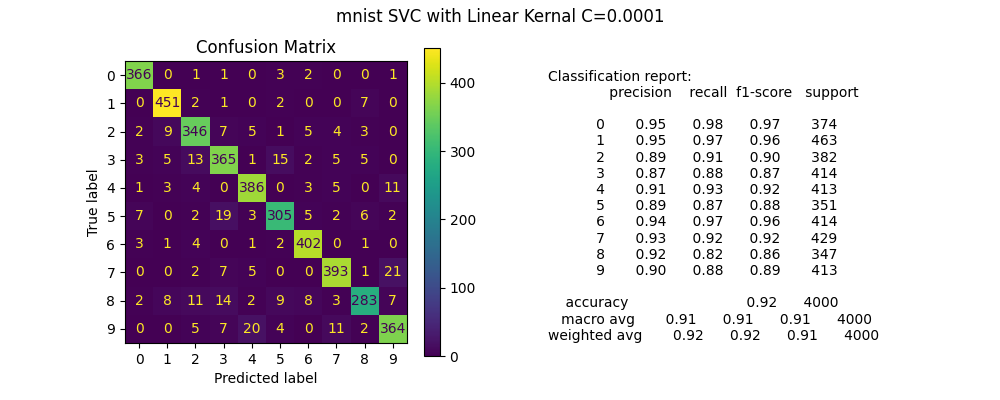
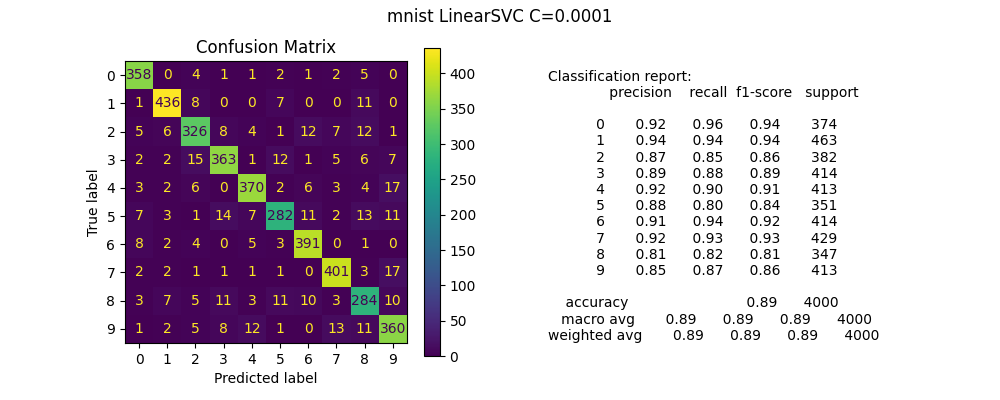
MNIST

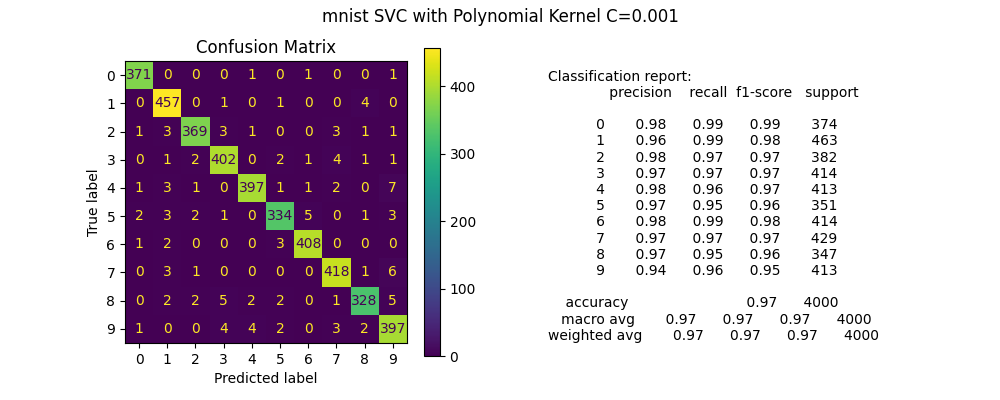
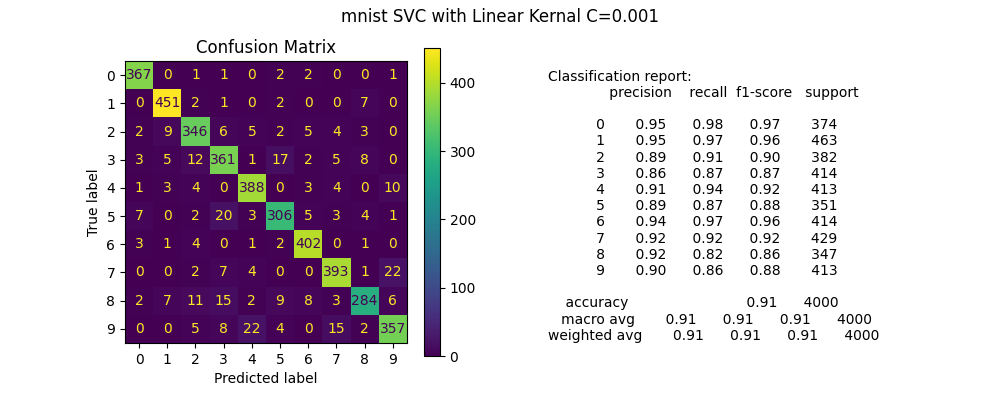
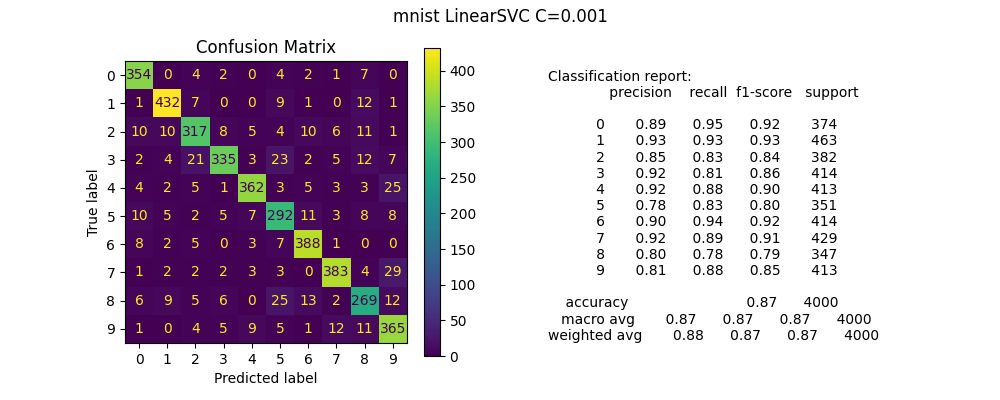
kNN

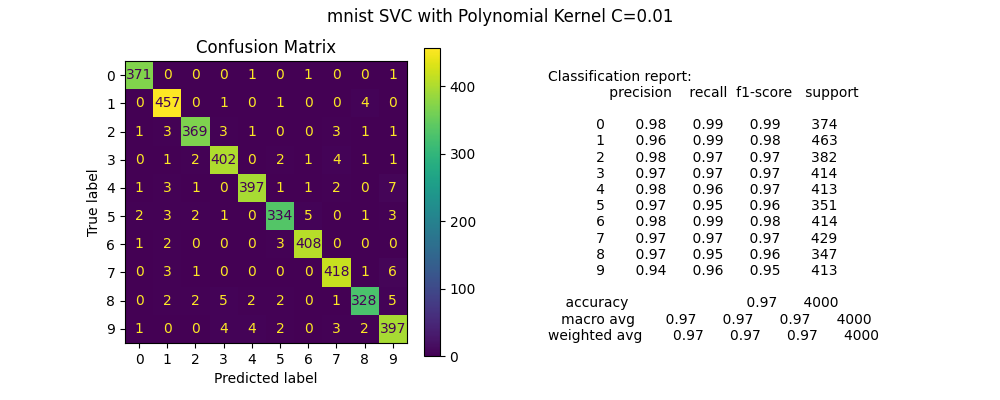
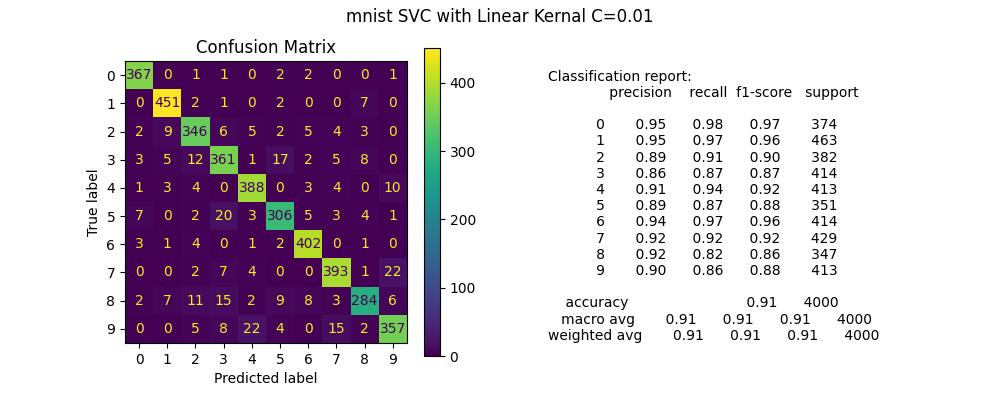
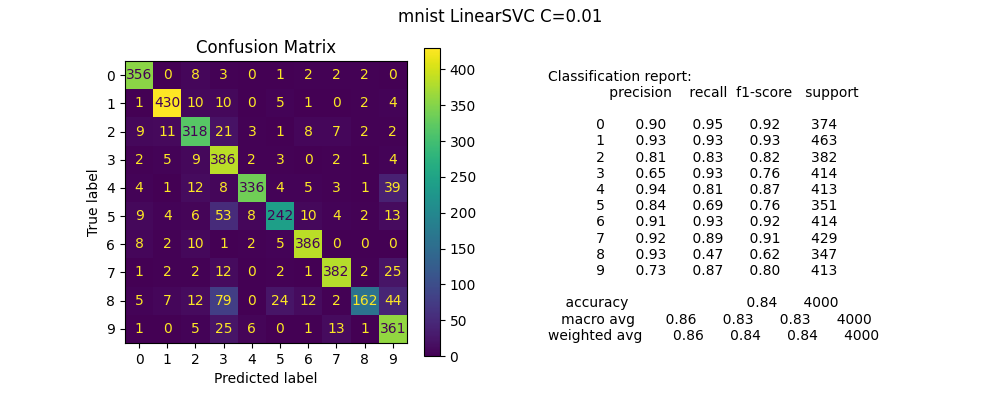


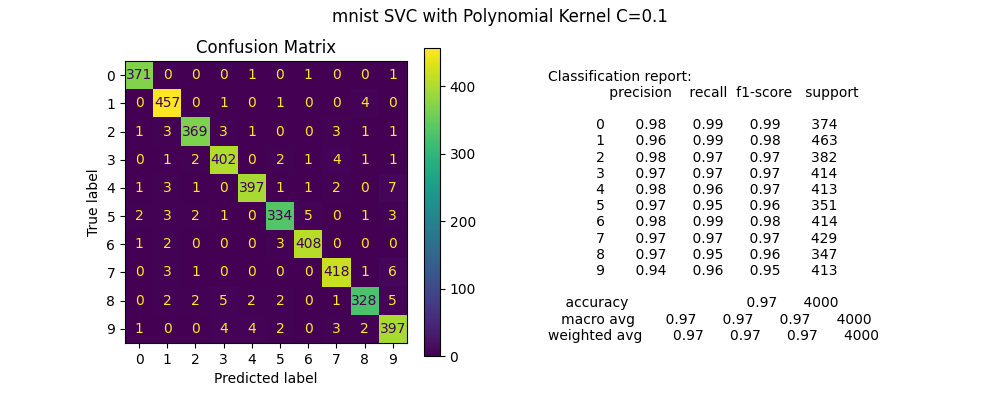
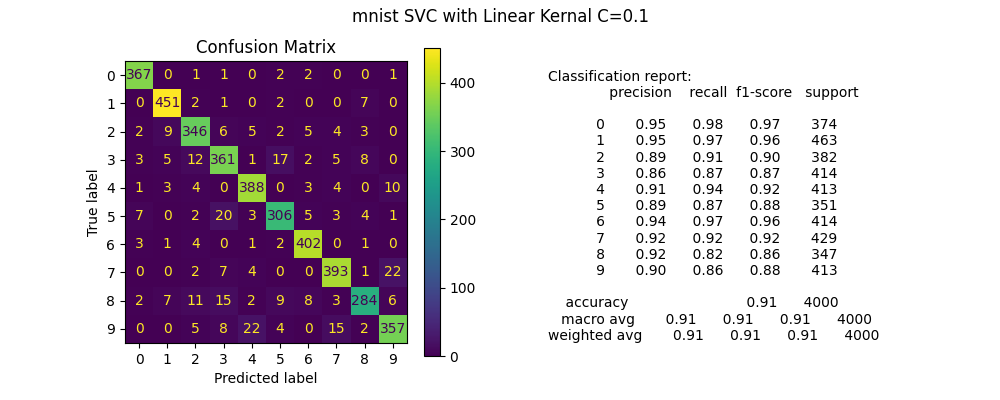
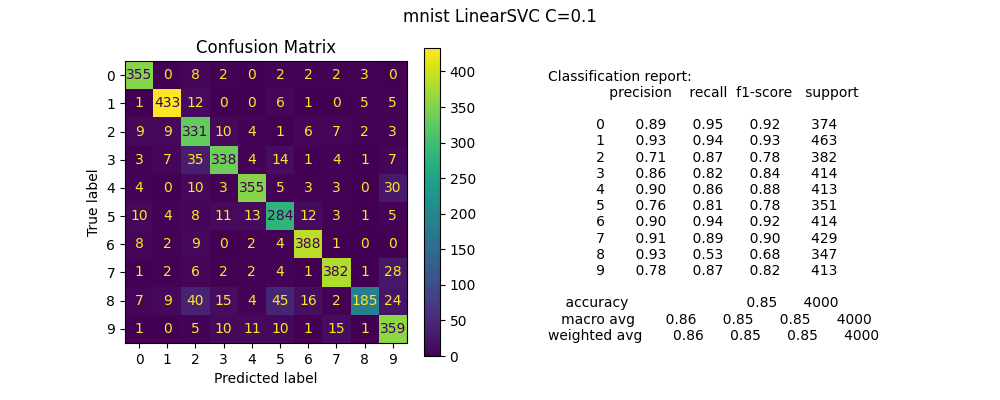
מן הנתונים הללו ניתן להסיק כי מספר השכנים האופטימלי נמוך, כלומר, ככל שנעזרים ביותר שכנים כך רמת הדיוק יורדת (במקרה זה ב0.03%). ניתן לראות שהשיפוע מתון יותר ביחס לkNN עבור IRIS, ייתכן וזאת משום שהערכים (השכנים) אותם נבדוק די דומים האחד לשני- כולם בצבע דומה, עם צפיפות דומה וכו'.

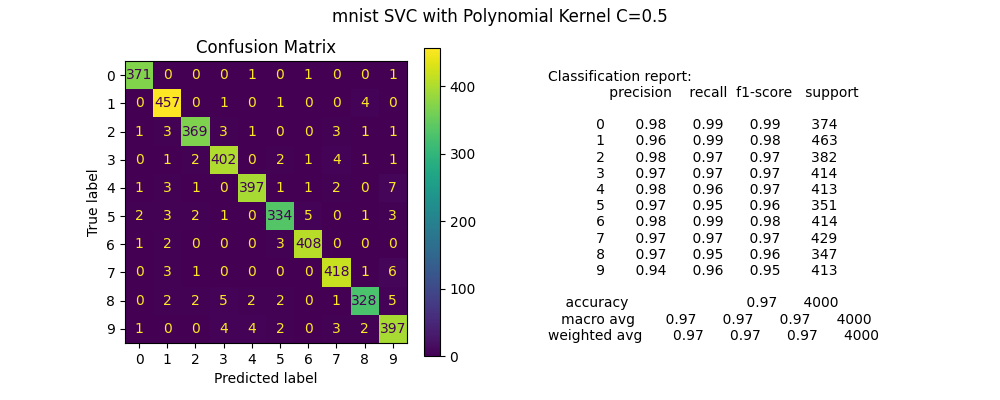
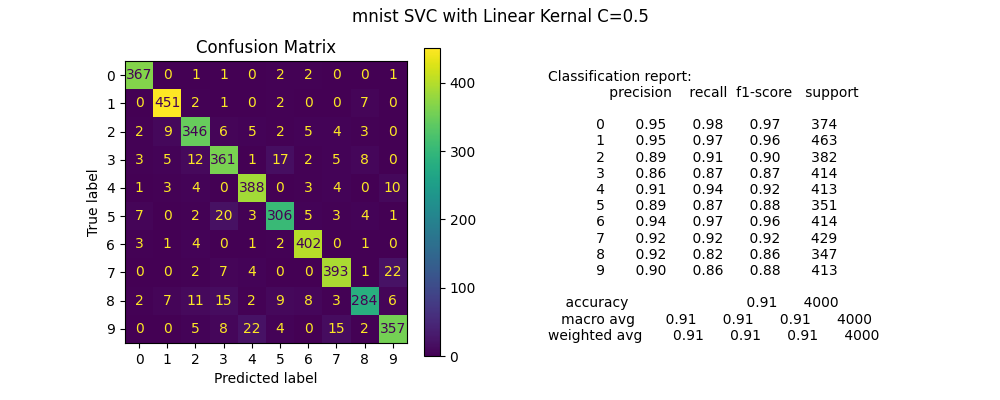
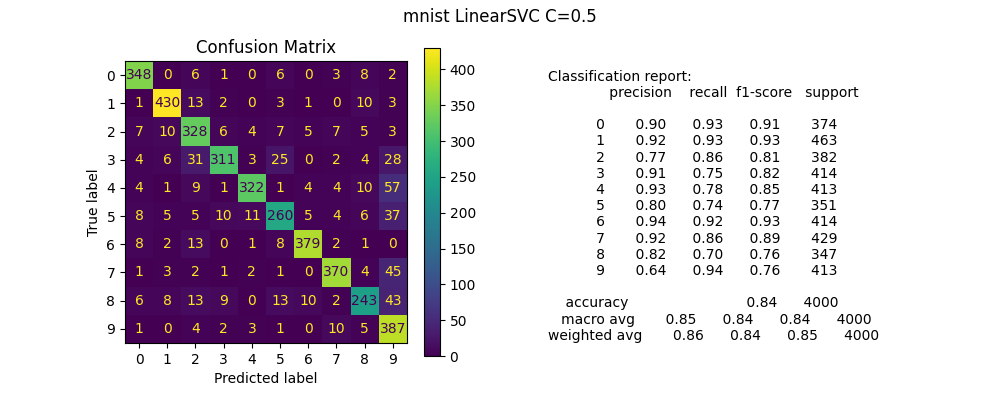
SVM:

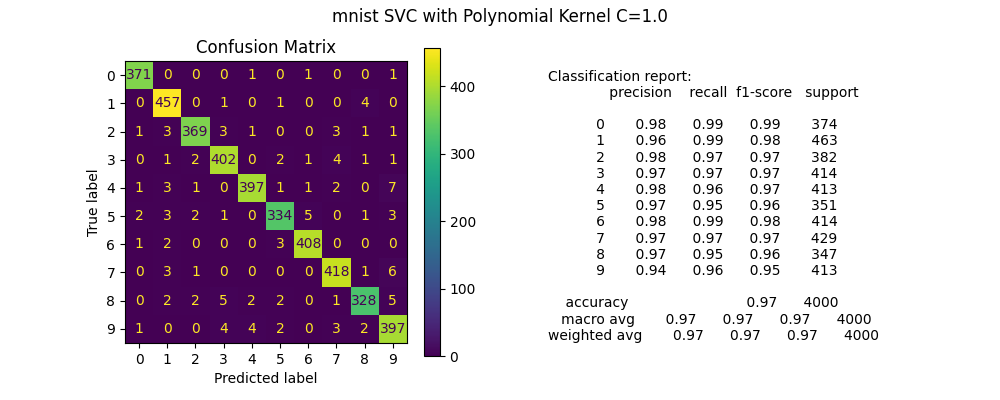
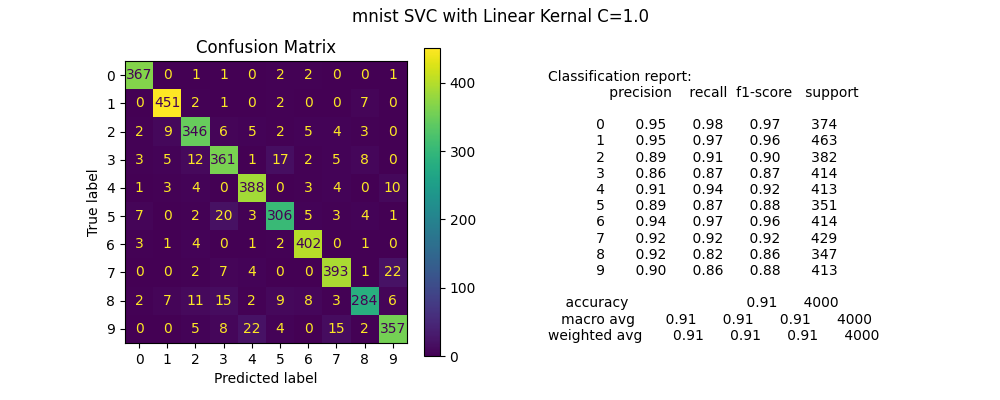
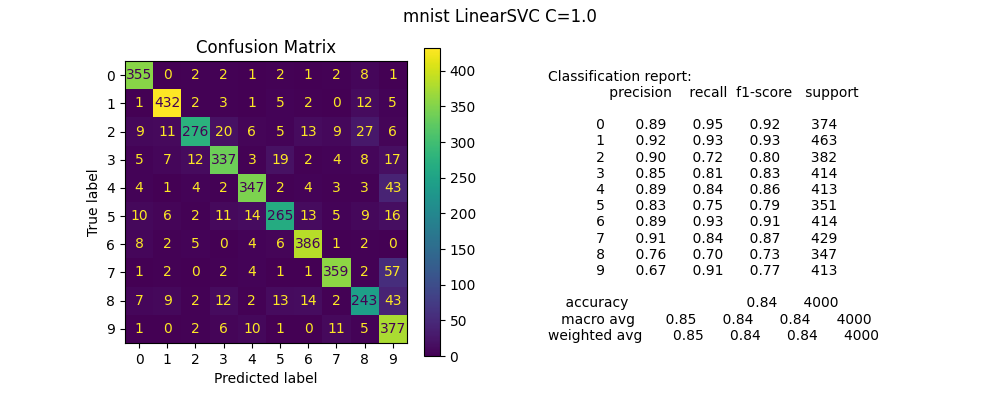


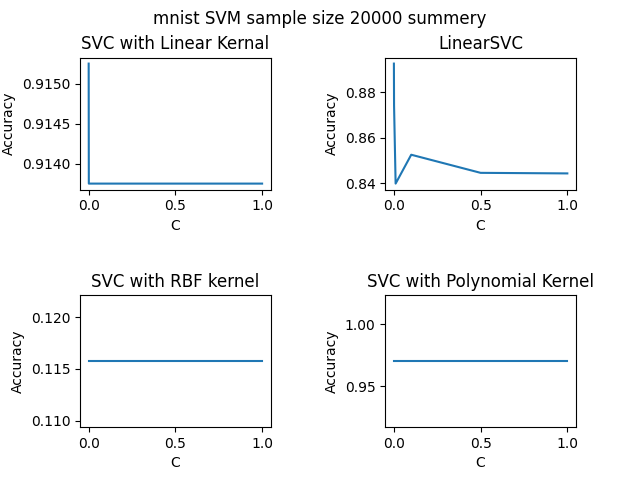












מן הנתונים הללו ניתן לראות כי התוצאות הטובות ביותר מגיעות משימוש באלגוריתם SVC עם גרעין פולינומיאלי (דרגה 3). יתר על כן, ניתן להגיד כי ערכי הC המשתנים אינם משפיעים רבות על רמת הדיוק של האלגוריתמים שנבדקו (SVC Poly לא מושפע בכלל).

למרות שרמת הדיוק של האלגוריתם המדובר טובה יותר משל אלגוריתם kNN, אני הייתי משתמש בkNN בשל זמן הריצה הארוך של האלגוריתמים המוזכרים בחלק זה (כ50 דקות עבור 20,000 דוגמיות של MNIST).