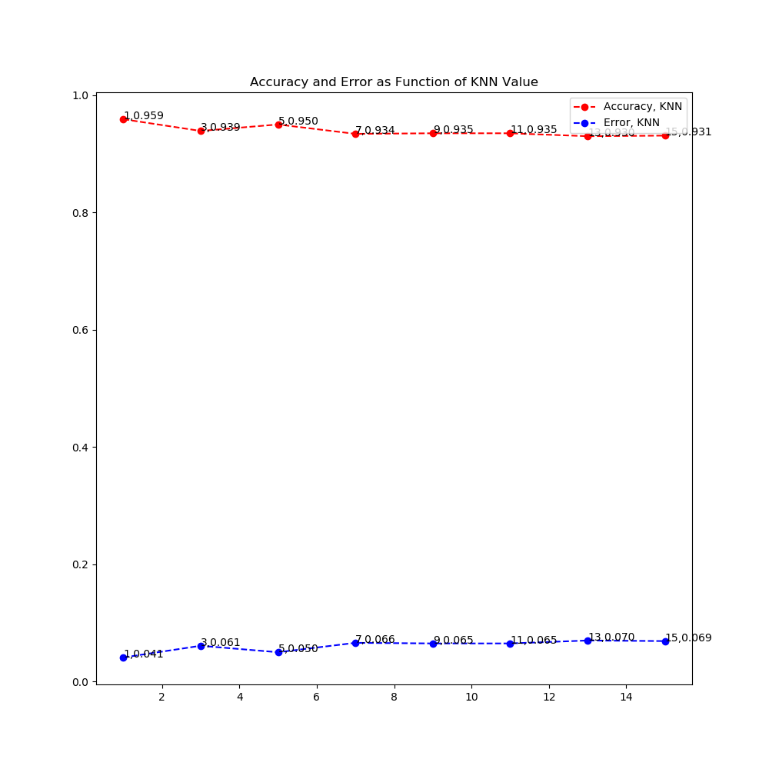
תרגיל בית 3

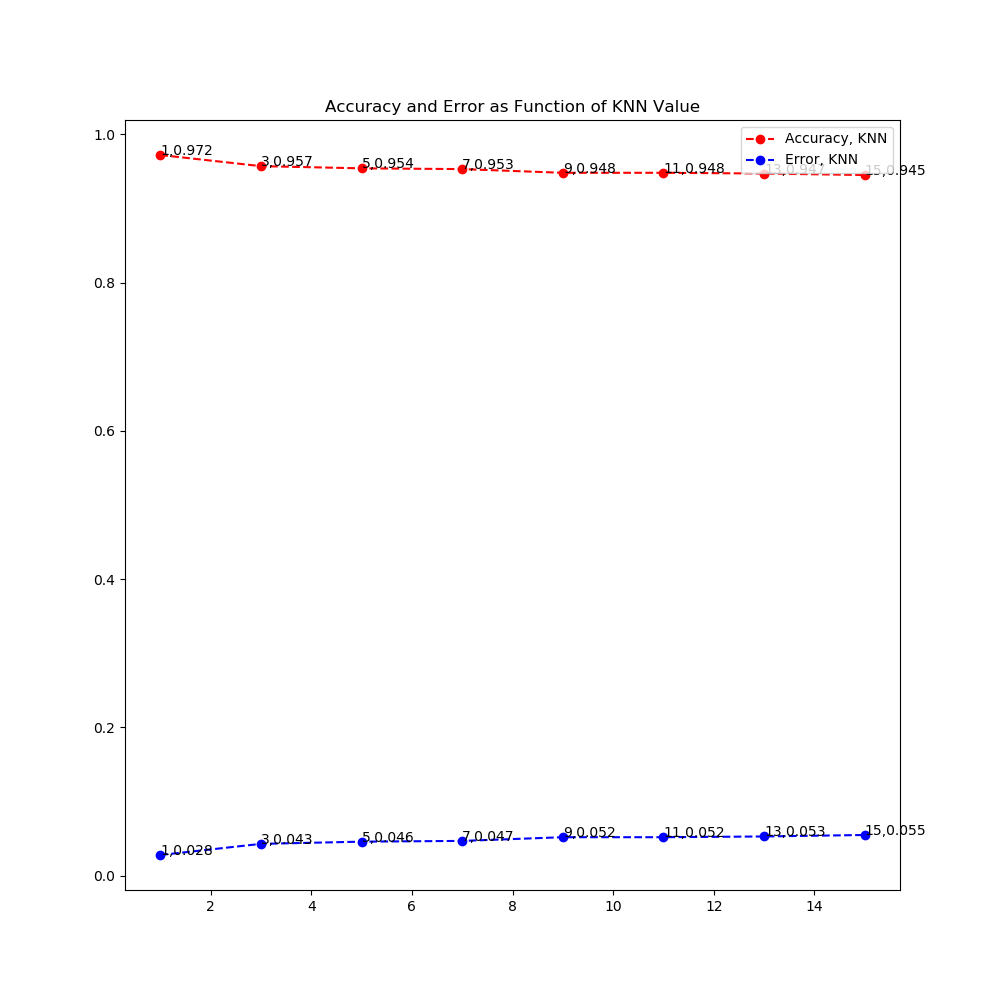
למידה וסיווג הפרעות קצב לב

5. קובץ בשם experiments6.csv מצורף לקבצי התרגיל.

א. גרף accuracy-error עבור הניסוי הנ"ל: יש לשים לב כי accuracy+error=1 לכל ערך, כמצופה.

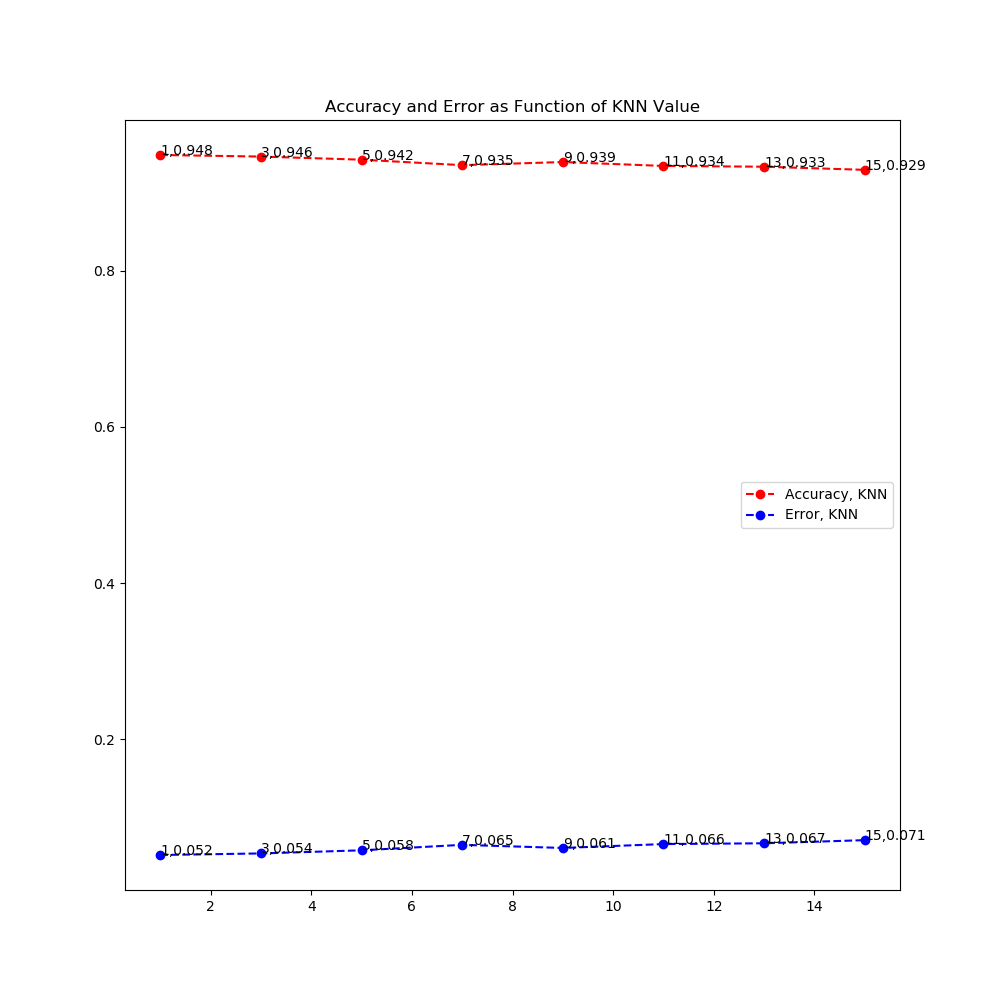


ב. נציין כי עבור הערך K=1 התקבלו הביצועים הטובים ביותר עם דיוק של 95.9%, אם כי עבור k=3,   
 הביצועים מתקרבים. כמו כן, כהעשרה, ביצענו את הניסוי מספר פעמים נוספות עבור חלוקות יותר   
 סטנדרטיות של k (הסטנדרט ברוב המקורות הוא K=10) וקיבלנו תוצאות דומות (97% עבורk=1 )



ג. ניתן להבחין במגמה של ירידה בביצועים ככל שמגדילים את k, כצפוי, כתוצאה מoverfitting.   
 עבור ערכי k גדולים, הערכת הסיווג של כל נקודה מתבססת על מס' שכנים רב, בהכרח חלקם רחוקים  
 יחסית מהנקודה, לכן לא רלוונטיים לסיווג שלה. לכן קיימת sweet-spot תלוית data-set ביחס למספר  
 השכנים, עבורו מתקבל דיוק אופטימלי.  
 על פי: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4978658/>, ניתן להחליש תופעה זו ע"י  
 משקול יחסי של השכנים הקרובים כך ששכנים קרובים **יותר** יקבלו חשיבות רבה יותר בקבלת החלטה.  
 לדוג', ניתן לבצע זאת באמצעות פונ' גאוסיאן (בדומה להורדת למה שנעשה באלג' הורדת ממדיות  
 t-SNE): , אך עקב הזנב הדק של פונ' גאוסיאן (דועכת מהר מאד), יתכן כי נק'  
 רחוקות במידה בינונית ישפיעו באופן אפסי על הסיווג. בעיה שניה עם פונ' similarity כזו היא הערך  
 סטיית התקן, שלמעשה מייצג את צפיפות הנקודות באזור, לכן אינו קבוע עבור כל הdata-set ויש   
 לחשבו מחדש עבור כל סיווג (יקר). אפשרות נוספת ששומרת על הסימטריה היפה של גאוסיאן,  
 אך ללא החסרונות שלו, היא t-student distribution (גם היא בשימוש נרחב בt-SNE). כאשר מדובר   
 בt-student עם דרגת חופש אחת, ההתפלגות מתנוונת להתפלגות קושי. כמו הגאוסיאן גם היא   
 סימטרית ובעלת צורת פעמון, המעניקה משקל רב למרחקים קרובים ל0, אך בשונה מהגאוסיאן  
 דועכת **לאט**, לכן תגרום לפחות הזנחה של נקודות במרחק בינוני:

ד. תוספת: בהקשר הזה נשים לב כי ככל הנראה הdata-set כבר מנורמל, משום שביצוע הניסוי פעם   
 נוספת לאחר ביצוע mean-std normalization של הdata לא הניב שיפור משמעותי:



7. כפי שניתן לראות בגרף הבא, התוצאות הטובות ביותר מתקבלות עבור המסווג knn עם k=1 מתרגיל 6:

