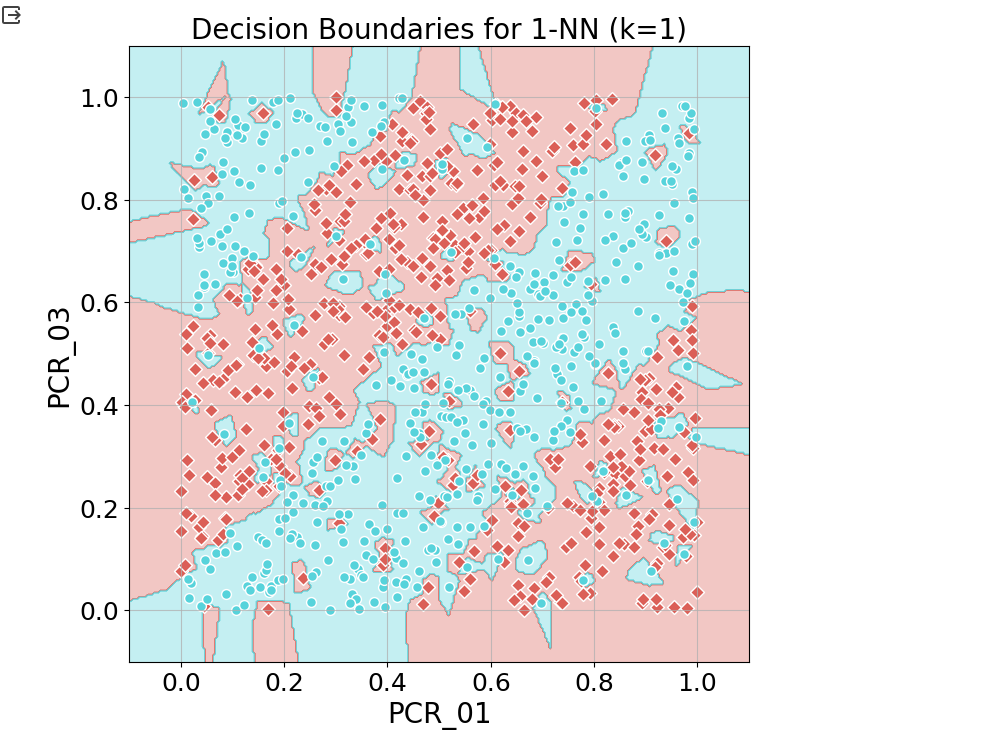
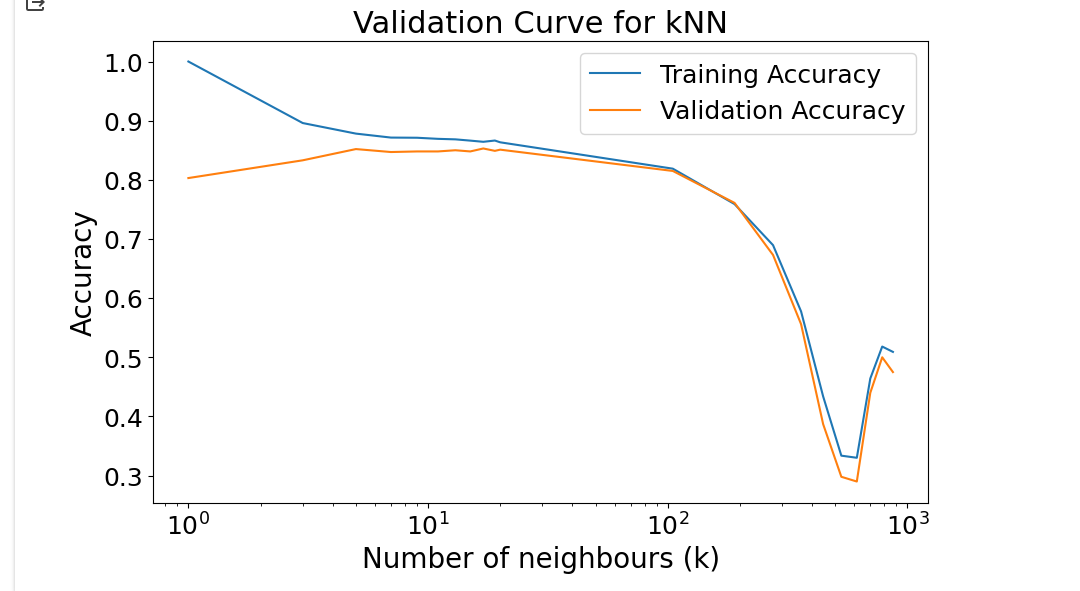
**HW2-Wet**

(Q1)

(Q2)

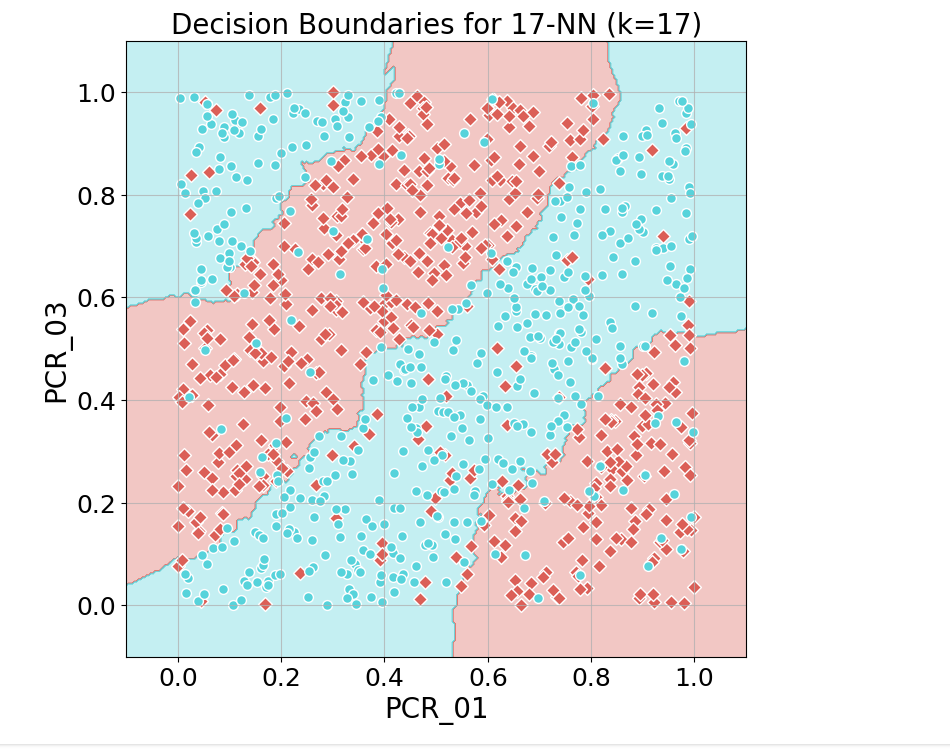


הk הטוב ביותר הוא k=17   
עם מידת דיוק ממוצעת על קבוצת האימון של 0.864 ומידת דיוק ממוצעת על קבוצת הולידציה ממוצעת של 0.853.

ערכי k נמוכים גורמים לתופעת הoverfitting. ניתן לראות כי עבור ערכי k נמוכים מידת הדיוק על קבוצת האימון היא גבוהה ומידת הדיוק על קבוצת הוולידציה נמוכה יותר(כפי שראינו בתרגול ההפרש הזה קשור לoverfit). תופעה זו קורית משום שעבור ערכי k נמוכים, אלגוריתם הknn מחשב את התיוג של נקודה מסוימת לפי מספר שכנים קטן. כתוצאה מכך, האלגוריתם רגיש לנקודות רעש, וניתן לראות גם בplot הקודם שנקודות רעש יוצרות מסביבן אזורי החלטה קטנים שיחזירו תיוג שגוי.

ערכי k גבוהים גורמים לתופעת הunderfitting. ניתן לראות כי עבור ערכי k נמוכים הן מידת הדיוק על קבוצת האימון והן ומידת הדיוק על קבוצת הוולידציה נמוכות. תופעה זו קורית משום שבמצב זה התיוג של נקודה מסוימת מושפע מיותר מדי שכנים, גם משכנים שרחוקים מדי. כתוצאה מכך, ההשפעה של הנקודות הרלוונטיות (הקרובות יותר) בחישוב התיוג יורדת ומקבלים תוצאות שגויות יותר בהרבה מקרים. ראינו בתרגול שאם השגיאה על קבוצת האימון גדולה יחסית, התופעה מקושרת לunderfit ולכך שהמודל שמקבלים הוא פשוט מדי ולא מצליח ללכוד את הקשרים שקיימים בקבוצת האימון.

(Q3)



מידת דיוק על קבוצת האימון: 0.862

מידת דיוק על קבוצת השגיאה: 0.812

4. גבולות ההחלטה של k=1 מתאימים באופן מושלם לקבוצת האימון מה שמרמז על overfitting ושונות גבוהה יותר במידה והיינו מאמנים מסד נתונים אחר . גבולות ההחלטה של k=17 ״חלקים״ יותר ומצליחים לתפוס את ארבעת אזורי ההחלטה העיקריים של הדאטה , מה שמרמז על יכולת הכללה טובה יותר איזון טוב יותר בין variance לbias.