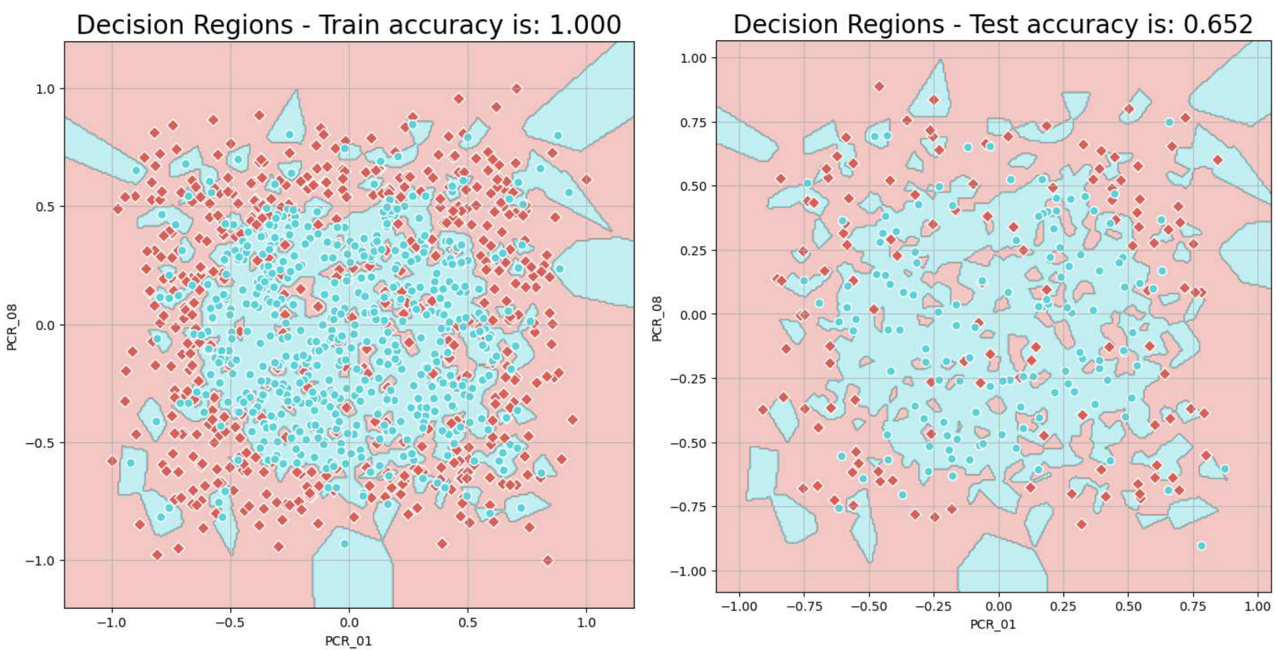
תרגיל בית רטוב 2 – מבוא למערכות לומדות (02360766)

בן הייטנר – 213930175

לילך ביטון – 205764517

חלק 1: בחירת מודל בסיסי עבור kNN

(Q1)

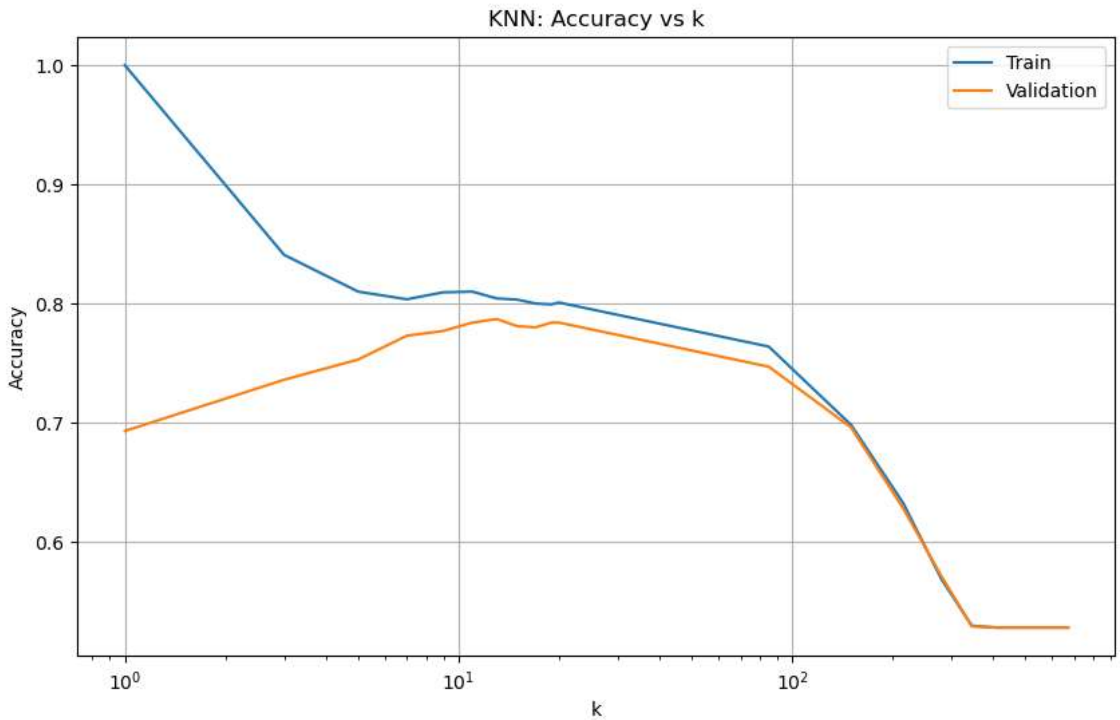


(Q2)

ה-k הטוב ביותר הוא: 13 (הוא האחד שעבורו מתקבל הדיוק הגבוה ביותר על ה-validation data).

עבור k קטן מאוד (לדוג' 1) נקבל overfitting שכן אזורי ההחלטה תלויים במעט נקודות ועל כן רגישים מאוד לסטיות בנתונים - כל סטיה מקבלת יחס של מקרה רגיל ונפוץ בלי קשר במגמה של הנתונים.

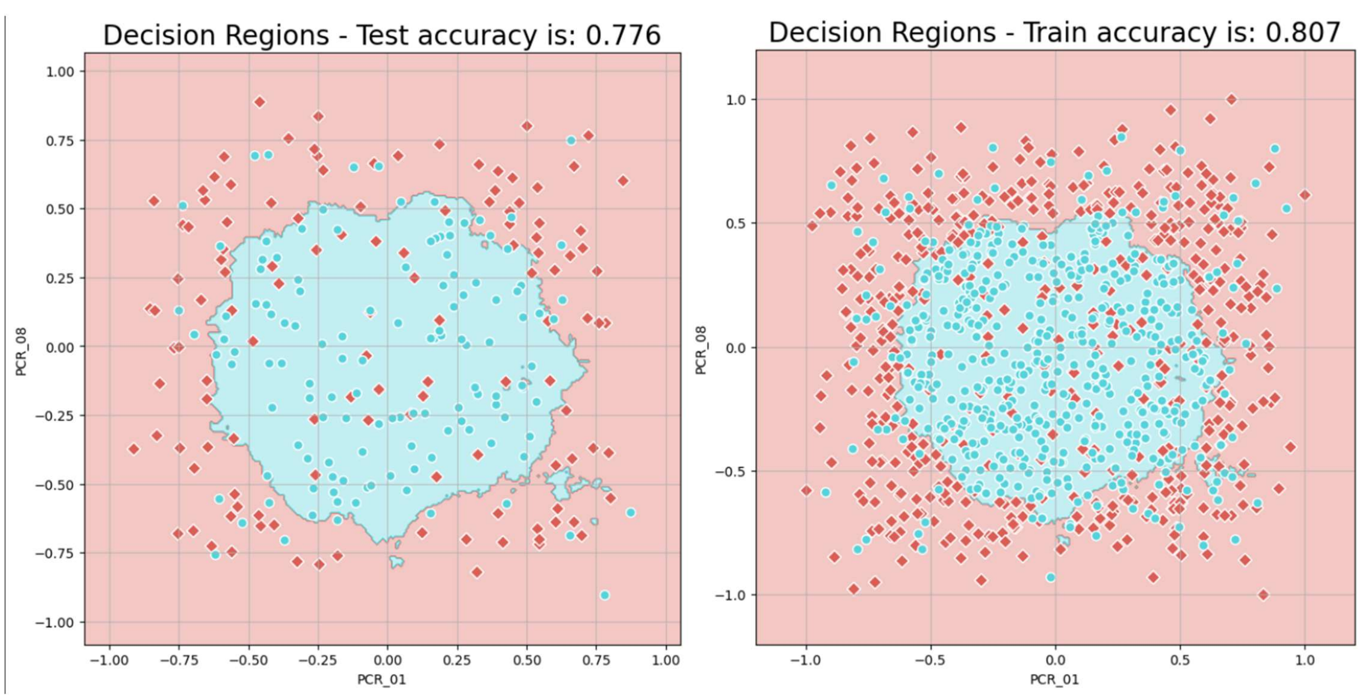
עבור k גדול מאוד נקבל underfitting מהסיבה ההפוכה – כל מקרה שלא רווח מאוד מקבל יחס של סטייה.



דיוק אימון ממוצע: 80.429%

דיוק ולידציה ממוצע: 78.7%

(Q3)



דיוק מבחן: 77.6%

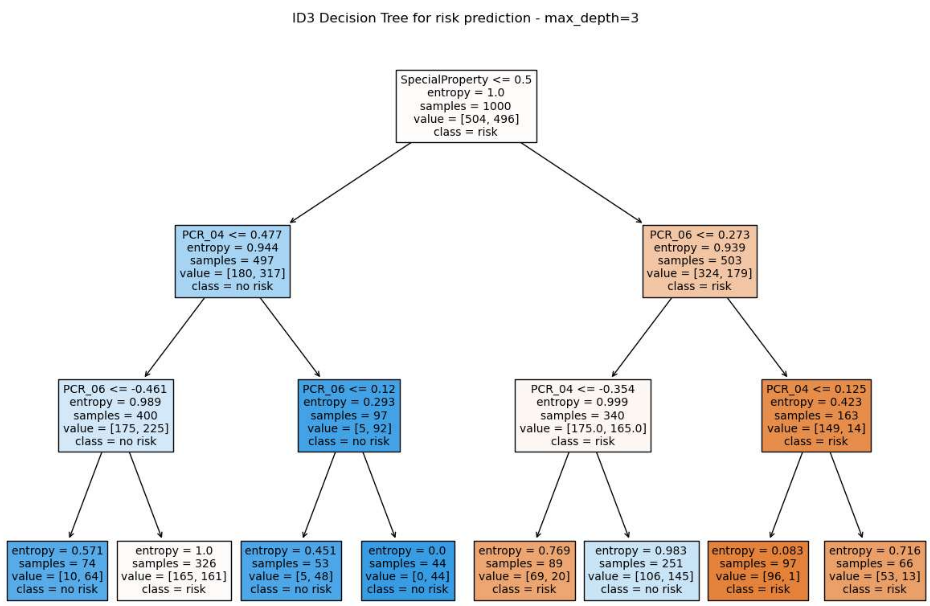
(Q4

ב-Q1 השתמשנו במודל עם k=1 וכפי שציינו ב-Q2 זה גורר ל-overfitting קיצוני כתוצאה מרגישות יתר של המודל לסטיות. לעומת זאת, ב-Q3 ע"י tuning של ה-k קיבלנו אזורי החלטה שמייצגים את המגמה של הנתונים ע"י איזון הרגישות לסטיות נקודתיות. עבור k-ים גדולים מ-13 נתחיל לקבל underfitting וכפי שראינו עבור ב-Q1 עבור k-ים קטנים מ-13 נקבל overfitting.

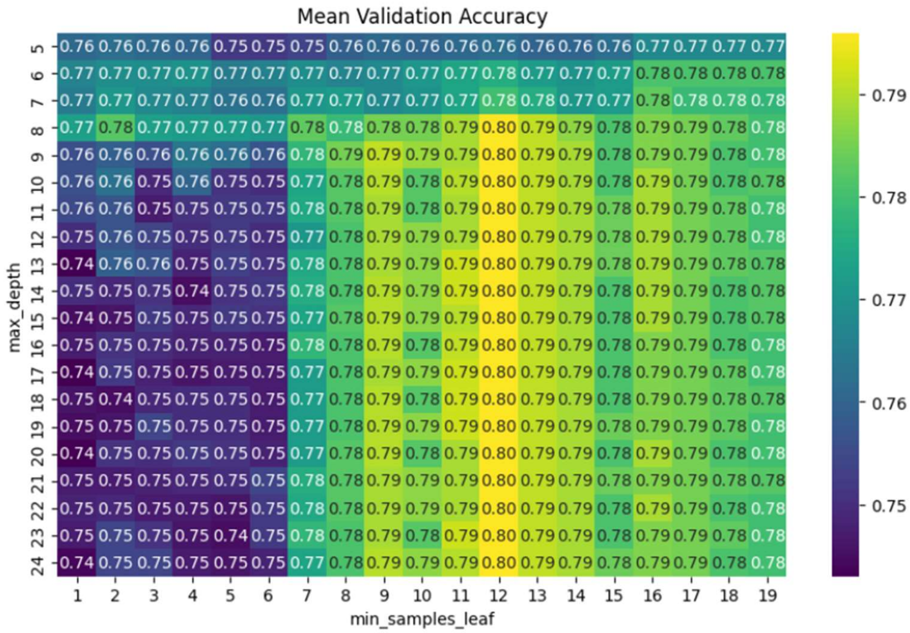
חלק 2: עצי החלטה

(Q5)

דיוק אימון: 68.4%



(Q6)



צירוף מיטבי (מקנה דיוק ולידציה ממוצע גבוה ביותר):

min_sample_leaf: 12.

max_depth: 8.

Underfitting:

min_sample_leaf: 19.

max_depth: 5.

Overfitting:

min_sample_leaf: 1.

max_depth: 24.

min_sample_leaf משפיע על הרגישות של המודל לסטיות. ככל שיותר נמוך, יותר קל ליצור עלה עבור דגימות אשר אינם בהכרח מייצגות את המגמה (סטייה), וככל שהוא יותר גבוה ייווצרו פחות עלים לדגימות שאינן מספיק נפוצות, אפילו אם מייצגות מגמה שאינה סטייה.

max_depth משפיע על סיבוכיות המודל. ככל שהוא יותר נמוך, המודל יכול לעשות פחות פיצולים ויכולתו לתאר התנהגות מוגבלת, וככל שהוא יותר גבוה המודל יכול לעשות יותר פיצולים ולעשות fit ליותר מקרים שונים.

ניתן להבין שעבור min_sample_leaf גבוה ו-max_depth נמוך נקבל underfitting כי יהיה לנו מודל שמסוגל לעשות מעט פיצולים ומאוד מקשה בסלידתו להתנהגויות שאינן רווחות מספיק. כמו כן, במקרה ההפוך נקבל overfitting כי המודל מסוגל לעשות פיצולים רבים ומאוד רגיש לסטיות ועל כן יצור מקרה לכל סטייה. נבחין שאלה אכן מתארים את המקרים שתיארנו לעיל (כאשר נמוך וגבוה זה ביחס לערכים בצירוף המיטבי).

(Q7)

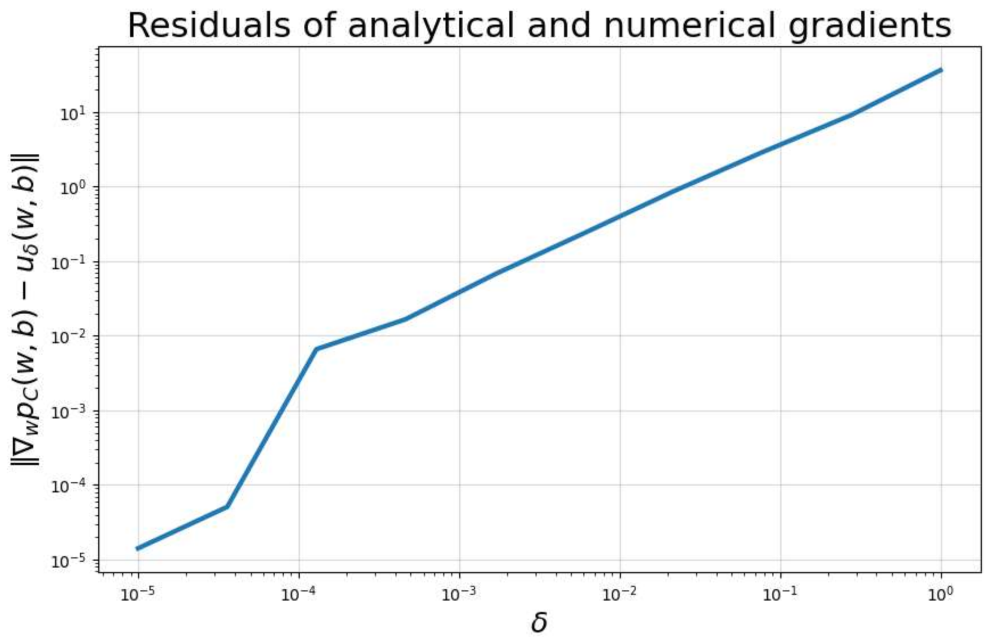
מספר הקומבינציות השונות שווה למכפלת גדלי הטווחים של ההיפר-פרמטרים. במקרה שלנו עבור שתי היפר-פרמטרים עם טווחים 20 ו-19 קיבלנו 380 קומבינציות. בהינתן היפרפרמטר נוסף, מספר הקומבינציות היה מוכפל בגודל הטווח של ההיפר-פרמטר החדש.

(Q8)

דיוק מבחן: 80.4%

חלק 3: SVM לינארי והגרעין הפולינומי

(Q9)



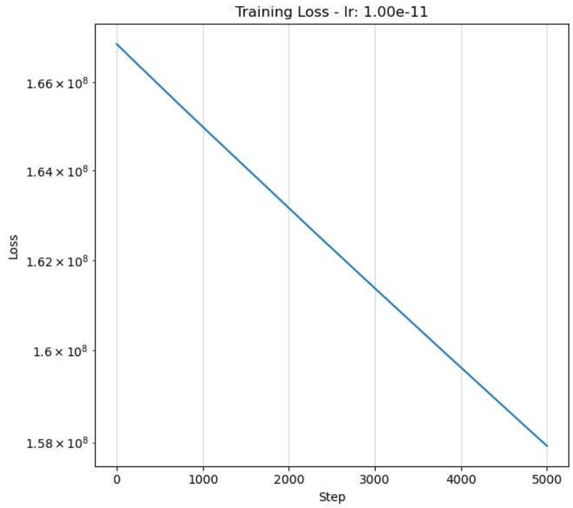
נבחין ראשית כי ההבדלים (ה-residuals) קטנים יחסית עבור כל δ. עם זאת, ניתן לראות שמתקבלת מעיין מגמה לינארית: δ גדל – residual גדל.

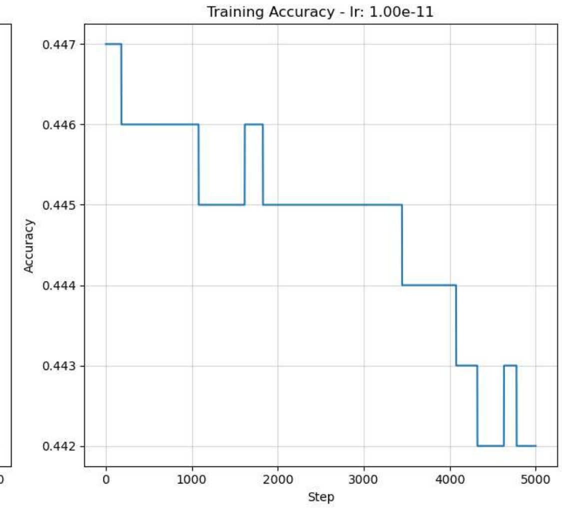
ניתן להצדיק את הקשר הכמעט לינארי הזה ע"י כך שככל ש-δ גדול יותר, כך החישוב הנומרי פחות מדויק ולכן מתרחק מהאנליטי בערך באותו סדר גודל.

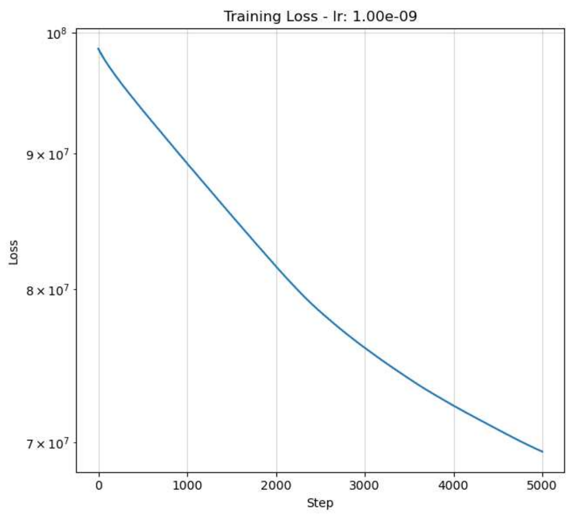
(Q10)

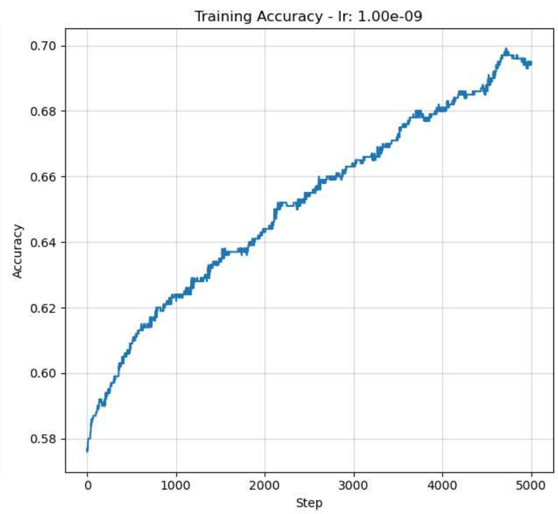
ניתן לראות לפי גרף דיוק האימון כי המודל במהרה מתקבע על רמת דיוק קרובה ל-50% ולמרות תזוזות קטנות ומרובות, לטובה ולרעה, לאורך הריצה, הדיוק נשאר בערך אותו הדבר (נשאר רע). מגרף ה-loss ניתן לראות דבר דומה. התקבעות מהירה על loss גבוה מדי והישארות על ערך דומה לאורך הריצה. זה יחד עם גרף אזורי ההחלטה מביא לתיאום עם הציפיות בהינתן ה-lr. קצב הלמידה קטן הרבה יותר מדי וזה גורר קטנים מדי שלא מקרבים אותנו למעוז חפצינו – מודל עם דיוק גבוה.

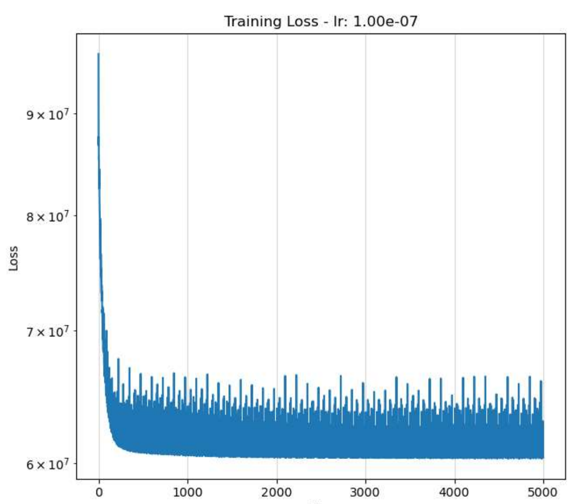
(Q11)

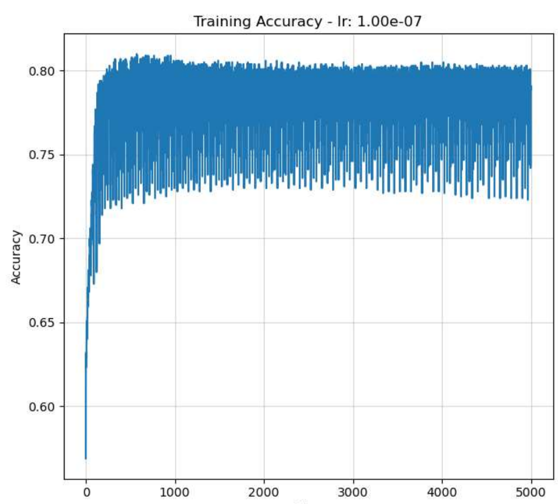

Training Loss - lr: 1.00e-11
Loss
Step
1.66 × 10^8
1.64 × 10^8
1.62 × 10^8
1.6 × 10^8
1.58 × 10^8
0
1000
2000
3000
4000
5000

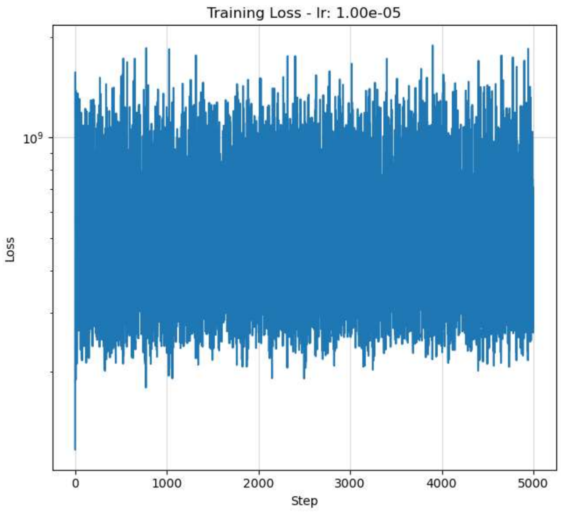

Training Accuracy - lr: 1.00e-11
Accuracy
Step
0.447
0.446
0.445
0.444
0.443
0.442
0
1000
2000
3000
4000
5000

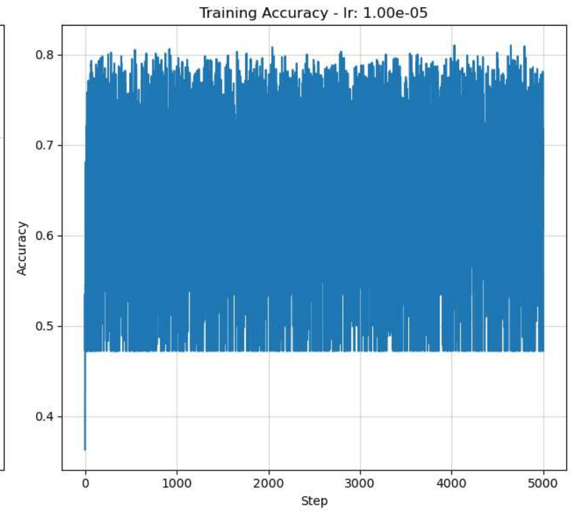

Training Loss - lr: 1.00e-09
Loss
Step
10^8
9 × 10^7
8 × 10^7
7 × 10^7
0
1000
2000
3000
4000
5000

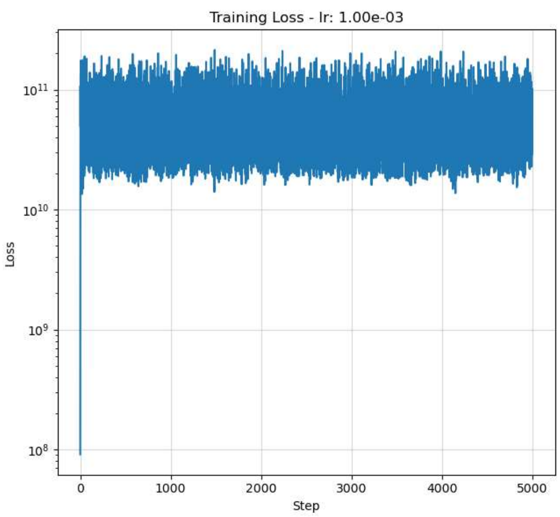

Training Accuracy - lr: 1.00e-09
Accuracy
Step
0.70
0.68
0.66
0.64
0.62
0.60
0.58
0
1000
2000
3000
4000
5000

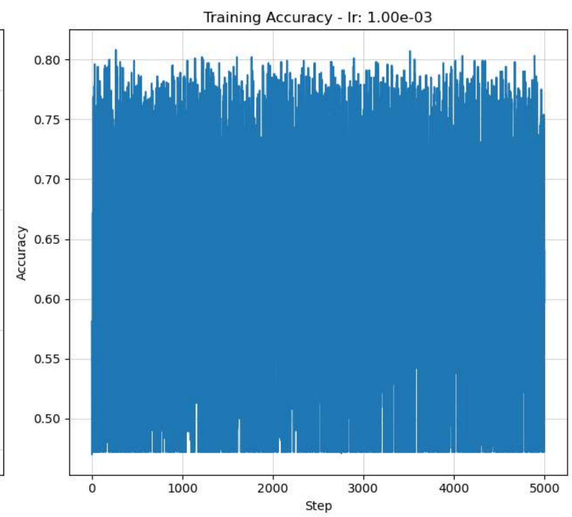

Training Loss - lr: 1.00e-07
Loss
Step
9 × 10^7
8 × 10^7
7 × 10^7
6 × 10^7
0
1000
2000
3000
4000
5000


Training Accuracy - lr: 1.00e-07
Accuracy
Step
0.80
0.75
0.70
0.65
0.60
0
1000
2000
3000
4000
5000



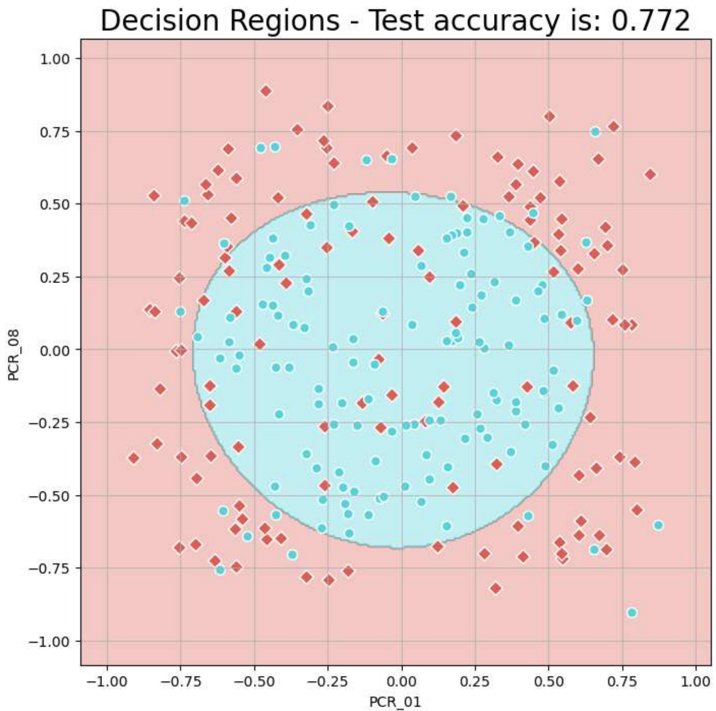






קצב הלמידה שהיינו בוחרים הוא 1×10^{-7} כי הוא מקנה התכנסות יציבה ל-loss נמוך ודיוק גבוה ביחס לאחרים. עבור ערכים קטנים יותר לא מקבלים תוצאות טובות באותה המידה ועבור רערכים גדולים יותר נקבל רעש אקססיבי שגם פוגע בתוצאות.

(Q12)



דיוק אימון: 78%

דיוק מבחן: 77.2%

<u>חלק 4: גרעין RBF</u>

(Q13)

<u>סעיף 1:</u>

a. ראינו כי $\phi(S)$ ממפה את הקבוצה S לווקטור בינארי בגודל d לכן כאשר נסתכל על המכפלה הפנימית נקבל מהגדרה סכום של מספר האינדקסים (המילים) שבהם $\phi(S_i)$ וגם $\phi(S_j)$ שווים ל-1:

$$\langle \phi(S_i), \phi(S_j) \rangle = \sum_{k=1}^{d} \phi(S_i)_k \cdot \phi(S_j)_k$$

מאחר ו-$\phi(S)_i$ שווה ל-1 כאשר $i \in S$ ול-0 אחרת, בהכרח מתקבל כי המכפלת הפנימית הנ"ל סופרת את מספר המילים הנמצאים בחיתוך בין שני הקבוצות, זאת אומרת גודל החיתוך בין שני הקבוצות:

$$\langle \phi(S_i), \phi(S_j) \rangle = \sum_{k=1}^{d} \phi(S_i)_k \cdot \phi(S_j)_k = |S_i \cap S_j|$$

קרנל נחשב מוגדר היטב כאשר הוא תואם למכפלה פנימית של מיפוי כלשהו, הראינו במפורש כי בהכרח הקרנל הנ"ל מקיים מכפלה פנימית של המיפוי $\phi(S)$ ולכן קרנל זה מוגדר היטב.

b. הראינו בסעיף a כי $K_1(S_i,S_j)=|S_i\cap S_j|$ הינו קרנל מוגדר היטב, ולכן מחוק 1 הנתון בשאלה מתקבל בהכרח כי הקרנל $K_{spam}(S_i,S_j)=e^{|S_i\cap S_j|}=e^{K_1(S_i,S_j)}$ הוא גם קרנל מוגדר היטב.

<u>סעיף 2</u>:

c. תחילה, נפתח את הביטוי של $\|\phi(S_i)\|_1$ מהגדרת נורמת l_1 ומהגדרת $\phi(S)$ כוקטור בינארי בגודל d:

$$\|\phi(S_i)\|_1=\sum_{k=1}^{d}|\phi(S_i)_k|$$

מאחר ו-$\phi(S)_i$ שווה ל-1 כאשר $i\in S$ ול-0 אחרת, בהכרח מתקבל כי נורמת l_1 הינה גודל הקבוצה S_i:

$$\|\phi(S_i)\|_1=\sum_{k=1}^{d}|\phi(S_i)_k|=|S_i|$$

לכן, נקבל את הפונקציה: $f(S_i)=e^{-\|\phi(S_i)\|_1}=e^{-|S_i|}$.

מחוק 2 הנתון בשאלה נתון כי מכפלה של קרנל מוגדר היטב בפונקציה המקבלת כל הקבוצה הינו גם קרנל מוגדר היטב:

$$k_2(S_i,S_j)=f(S_i)k_1(S_i,S_j)f(S_j)$$

נציב בחוק זה את הפונקציה שקיבלנו ואת הקרנל $K_{spam}(S_i,S_j)$ שהראינו בסעיף b שהוא מוגדר היטב:

$$k_2(S_i,S_j)=e^{-|S_i|}e^{|S_i\cap S_j|}e^{-|S_j|}=e^{|S_i\cap S_j|-|S_i|-|S_j|}$$

מתורת הקבוצות אנחנו יודעים כי מתקיים:

$$|S_i\cup S_j|=|S_i|+|S_j|-|S_i\cap S_j|$$

נעביר אגפים ונקבל:

$$|S_i\cap S_j|=|S_i|+|S_j|-|S_i\cup S_j|$$

כעת נציב חזרה ב-$k_2(S_i,S_j)$ ונקבל:

$$k_2(S_i,S_j)=e^{|S_i\cap S_j|-|S_i|-|S_j|}=e^{|S_i|+|S_j|-|S_i\cup S_j|-|S_i|-|S_j|}=e^{-|S_i\cup S_j|}$$

לכן הוכחנו שגם $e^{-|S_i\cup S_j|}$ הינו גם קרנל מוגדר היטב.

d. תהיי $\phi_1(x)$ פונקציית מיפוי עבור הקרנל המוגדר היטב $k_1(x,x')$ ו- $\phi_2(x)$ פונקציית מיפוי עבור הקרנל המוגדר היטב $k_2(x,x')$. נגדיר $\phi_1(x)_i$ להיות ערך הפיצ׳ר ה-i תחת המיפוי $\phi_1(x)$ ונגדיר $\phi_2(x)_j$ להיות ערך הפיצ׳ר ה-j תחת המיפוי $\phi_2(x)$ ונקבל:

$$k_1(x,x')k_2(x,x')=\left(\phi_1(x)\cdot\ \phi_1(x')\right)\left(\phi_2(x)\cdot\ \phi_2(x')\right)$$

$$=\left(\sum_{i=1}\phi_1(x)_i\phi_1(x')_i\right)\left(\sum_{j=1}\phi_2(x)_j\phi_2(x')_j\right)=\sum_{i,j}\phi_1(x)_i\phi_1(x')_i\phi_2(x)_j\phi_2(x')_j$$

$$=\sum_{i,j}\left(\phi_1(x)_i\phi_2(x)_j\right)\left(\phi_1(x')_i\phi_2(x')_j\right)$$

נגדיר כעת פונקציית מיפוי $\phi_3(x)$ שבה ערך התכונה $\phi_3(x)_{i.j}$ בעבור כל זוג (i,j) מוגדר כך:

$$\phi_3(x)_{i.j}=\phi_1(x)_i\phi_2(x)_j$$

ואז נקבל שבעבור הקרנל החדש המקיים $k(x,x')=k_1(x,x')k_2(x,x')$ קיימת פונקציית מיפוי $\phi_3(x)$ אשר המכפלה הפנימית שלו תואמת לקרנל:

$$k(x,x')=k_1(x,x')k_2(x,x')=\sum_{i,j}\left(\phi_1(x)_i\phi_2(x)_j\right)\left(\phi_1(x')_i\phi_2(x')_j\right)$$

$$=\sum_{i,j}\left(\phi_3(x)_{i.j}\right)\left(\phi_3(x')_{i.j}\right)=\left(\phi_3(x)\cdot\ \phi_3(x')\right)=\langle\phi_3(x),\phi_3(x')\rangle$$

ולכן, הקרנל החדש $k(x,x')=k_1(x,x')k_2(x,x')$ הינו קרנל מוגדר היטב אם גם $k_1(x,x')$ הינו קרנל מוגדר היטב וגם $k_2(x,x')$ הינו קרנל מוגדר היטב.

e. בסעיף קודם הראינו כי מכפלה סקלרית של קרנלים מוגדרים היטב היא בעצמה קרנל מוגדר היטב. כמו כן, בסעיפים b ו-c, בהתאמה, הראינו ש-$e^{|S_i\cap S_j|}$ ו-$e^{-|S_i\cup S_j|}$ הם קרנלים מוגדרים היטב. נבחין כי מתקיים:

$$K_{spam\ pro\ max}=e^{|S_i\cap S_j|-|S_i\cup S_j|}=e^{|S_i\cap S_j|}\cdot e^{-|S_i\cup S_j|}$$

ולכן, $K_{spam\ pro\ max}$ הוא קרנל מוגדר היטב.

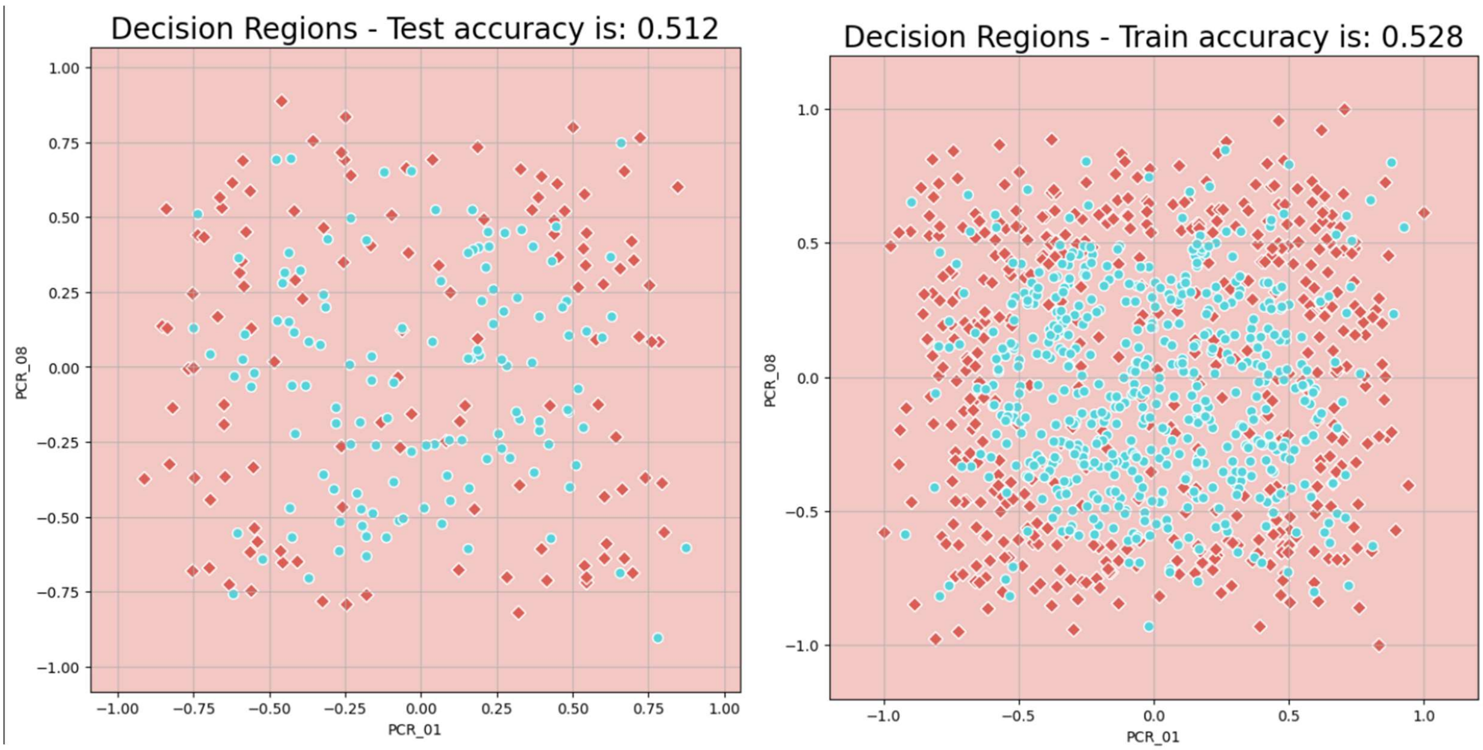
f. הקרנל שהחבר הציע עדיף כיוון שהוא משלב את האיחוד והחיתוך של הקבוצות ובכך למעשה יכול לתפוס את הדמיון בין קבוצות באופן מחוכם יותר. כלומר, יש לקרנל מידע על דמיון וכן על אי דמיון בין הקבוצות ועל כן הוא יכול "להסיק" יותר על הקשר בין הקבוצות. לדוג' ניתן לדעת את הגדלים של הקבוצות.

g. נציע את הקרנל הבא:

$$K_{spam\ too\ muc}\ =e^{\gamma\left(\frac{|S_i\cap S_j|-|S_i\cup S_j|}{|S_i\cup S_j|}\right)}$$

נרצה לנרמל את כמות המילים כתלות בגודל האיחוד של שני הקבוצות כך שנוכל לקבל מספר בין 0 ל-1 ללא תלות בגודל הקבוצות. כך למעשה נקבל ערך של דמיון בין הקבוצות שאינו תלוי בגודל המייל או בכמות המילים השונות.
בנוסף אם נוסיף היפר-פרמטר לערך האקספוננט נאפשר לקרנל יכולת **tuning** על שיפוע הדעיכה שסביר שתחזק אותו אף יותר.

(Q14)

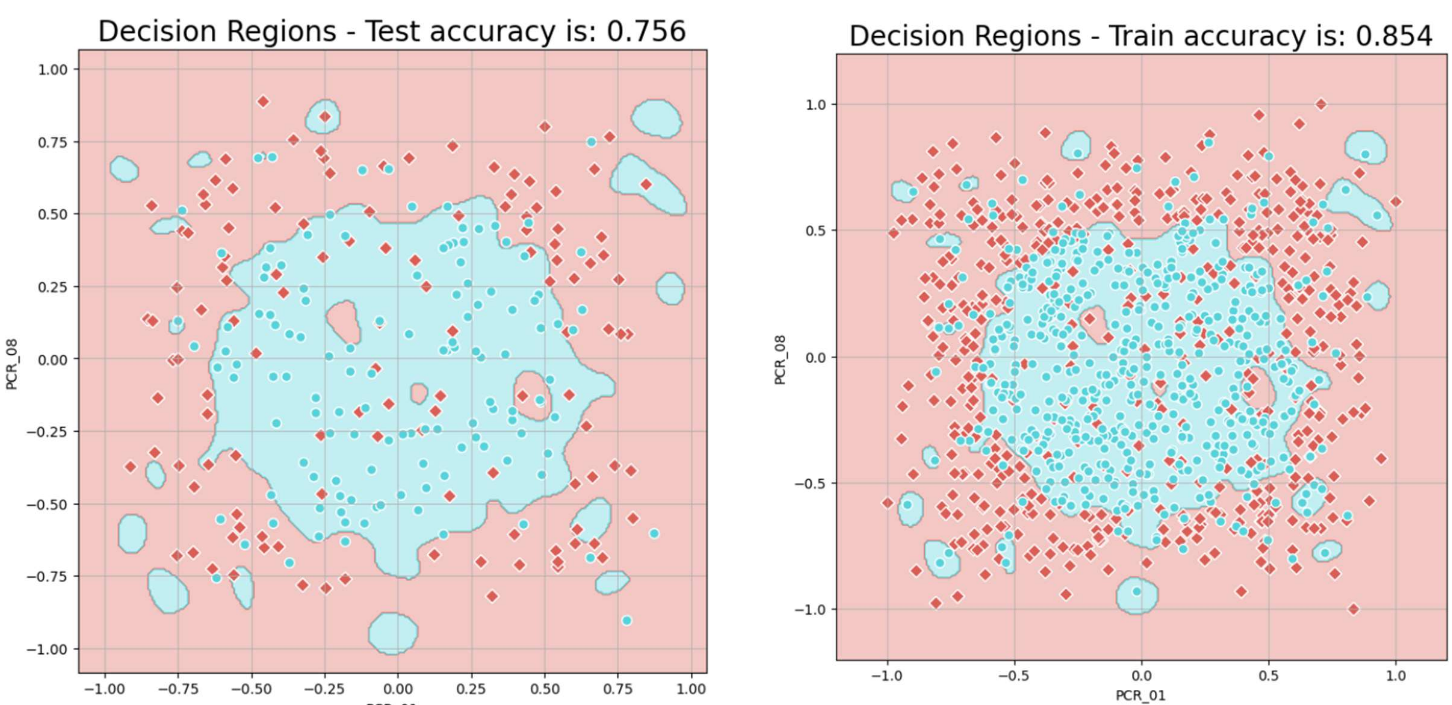


דיוק אימון: 52.5%

דיוק מבחן: 51.2%

ניתן להבחין כי יש underfitting. מצד אחד יש דיוק נמוך על נתוני המבחן אבל מצד שני יש דיוק דומה עבור נתוני האימון – המודל לא מצליח לתאר את מגמת הנתונים בכלל. כלומר, underfitting.

(Q15)

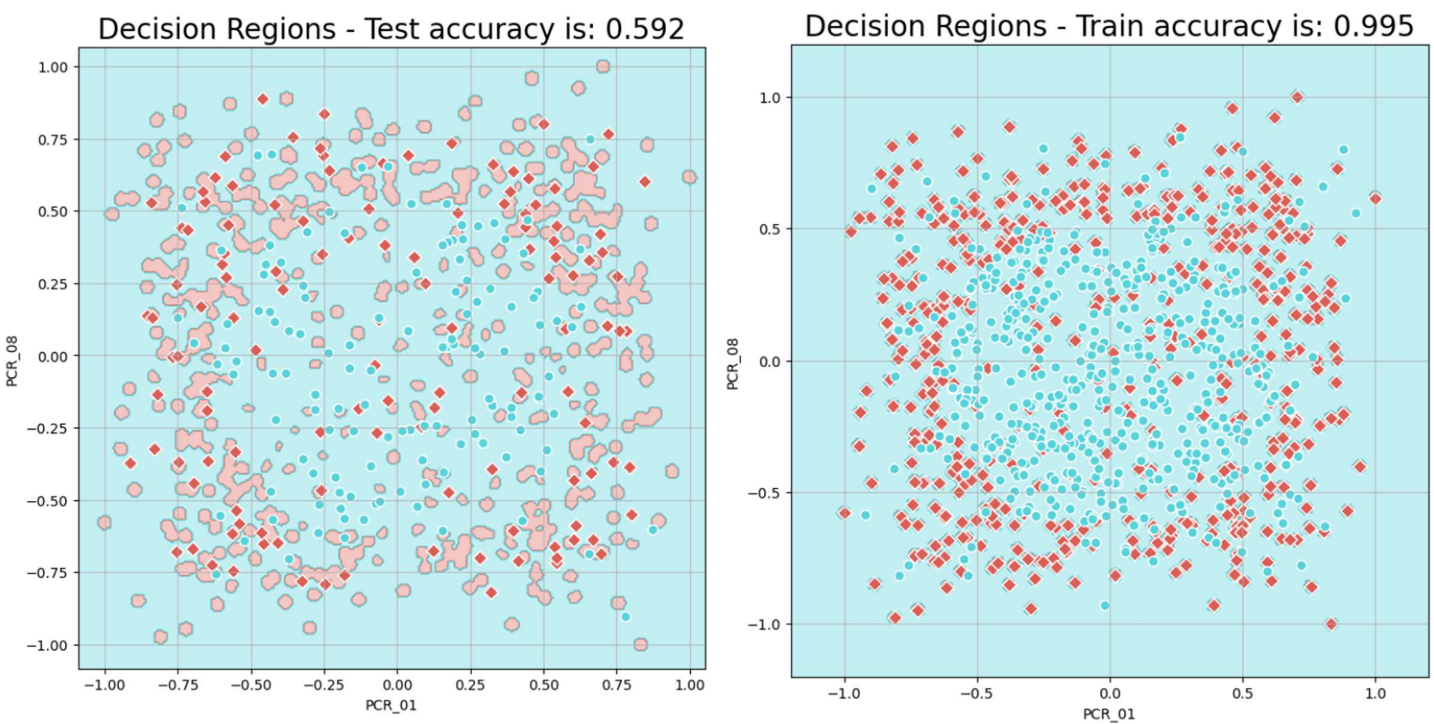


דיוק אימון: 85.4%

דיוק מבחן: 75.6%

המודלים אכן נותנים אזורי החלטה דומים יחסית עם דיוק קרוב. ההבדל הניכר בין תוצאות המודלים הוא ה"איים" שיש בגרף אזורי ההחלטה של ה-RBF. איים אלה נוצרים מהתהליך בו RBF מוצא את הפונקציה המתאימה ליצירת אזורי ההחלטה ולא יכלו להיווצר ע"י kNN (לדוג' אי כחול שקרוב להרבה יותר נקודות אדומות מכחולות). האיים האלה הם overfitting שנוצרים מהעוצמה של גרעין RBF ופוגעים מעט בדיוק מודל זה ביחס למודל ה-kNN.

(Q16)



דיוק אימון: 99.5%

דיוק מבחן: 59.2%

ניתן להבחין כי יש overfitting. מצד אחד יש דיוק נמוך על נתוני המבחן אבל מצד שני יש דיוק גבוה מאוד (כמעט מושלם) עבור נתוני האימון – המודל מתאר את נתוני האימון ואותם בלבד. כלומר, overfitting.