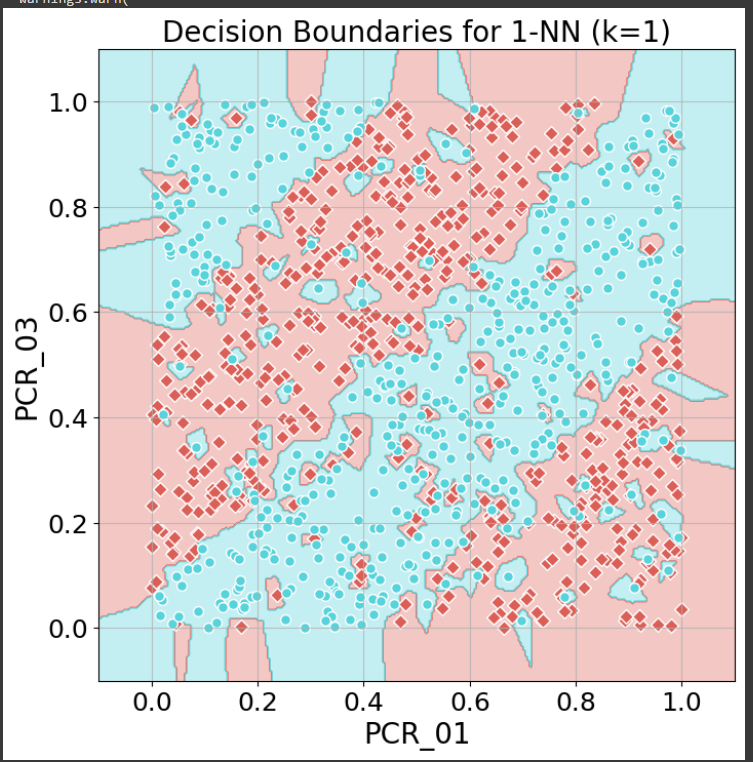
Visualization and basic analysis

(Q1)



(Q2)

A graph with blue and orange lines

Description automatically generated

העקומה הכתומה מציגה לנו את מידת הדיוק הממוצעת ( ממוצעת על קבוצות ולידיציה שקיבלנו מתהליך הcross validation) של מודלי knn בעלי k שכנים.

אנחנו רוצים לבחור את הk שנתן לנו את הדיוק הכי טוב על הvalidation set (מעיד על יכולת הכללה טובה יותר), לכן ניקח את הנקודה המקסימלית על העקומה הכתומה ונמצא את הk שנותן לנו את הדיוק המקסימלי.

עשינו את זה כמובן בעזרת קוד פייתון וקיבלנו :

A screenshot of a computer

Description automatically generated

כלומר הk הטוב ביותר שמצאנו הוא k=20.

**ערכי k נמוכים גורמים לתופעת הoverfitting.** ניתן לראות בגרף הvalidation curve כי עבור ערכי k נמוכים מידת הדיוק על קבוצת האימון היא גבוהה ומידת הדיוק על קבוצת הוולידציה נמוכה יותר. בתרגול ראינו כי מצב כזה מקושר לoverfit, כלומר דיוק טוב מדי על הtraining set הוא לא בהכרח טוב ואף יכול לפגוע ביכולת ההכללה על הtest set.

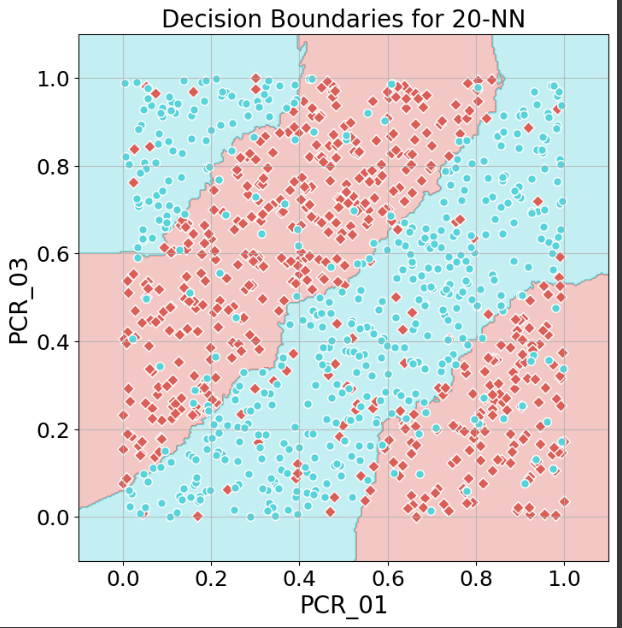
ראינו בשיעורי בית הקודמים שכאשר מס׳ השכנים ב KNN הוא קטן , קורית תופעת הoverfitting מאחר ויש משקל גדול יותר למספר קטן של השכנים הקרובים ביותר ולכן ההשפעה של כל שכן קרוב גדולה יותר. כתוצאה מכך האלגוריתם רגיש לרעש, ובמידה ויש נקודות חריגות בtraining set עלול להיווצר אזור סיווג שגוי מסביב לנקודות הרעש, כפי שרואים גם בplot בשאלה 1.

**ערכי k גבוהים גורמים לתופעת הunderfitting.** ניתן לראות בעקומה שלנו כי ככל שk גדל אנחנו מקבלים שהדיוק הולך וקטן והעקומות של שניהם מתכנסים לאותה נקודה של דיוק נמוך מאוד.

תופעה זו קוריתמאחר וכאשר מספר השכנים גדול מדי התיוג של הנקודות מושפע גם משכנים רחוקים שלא רלוונטים לסיווג של אותה הנקודה. כתוצאה מכך, נוצר מצב שההשפעה של הנקודות הרלוונטיות (הקרובות יותר) יורדת ונקודות חסרות חשיבות נלקחות גם כן בחשבון ועם כמותן גדולה, השפעתן יכולה גם להעפיל על המשקל של הנקודות הקרובות ולייצר סיווג שגוי.

לכן, נוצר מצב של חוסר התאמה של המודל לtraining set וגם לtest set. במהלך חישוב הפרדיקציה הן על קבוצת האימון והן על קבוצת המבחן, המודל לוקח בחשבון נקודות לא רלוונטיות לחישוב הפרדיקציה ולכן טועה על שניהם. ככל שk גדל, המודל נעשה כללי מדי ופשוט ואינו מסתגל היטב לניואנסים של קבוצת האימון. מכיוון שהוא לא מתאים התחזיות של המודל אינן מושפעות במידה רבה מהמאפיינים של נתוני האימון, מה שמוביל לציוני דיוק דומים ונמוכים. על כן, עקומות הדיוק של האימון והולידציה מתכנסות לאותה מידת דיוק נמוכה.

(Q3)

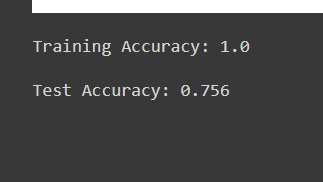


A black background with white text

Description automatically generated

(Q4)

התוצאות של המודל 1NN בסעיף 1 הן :



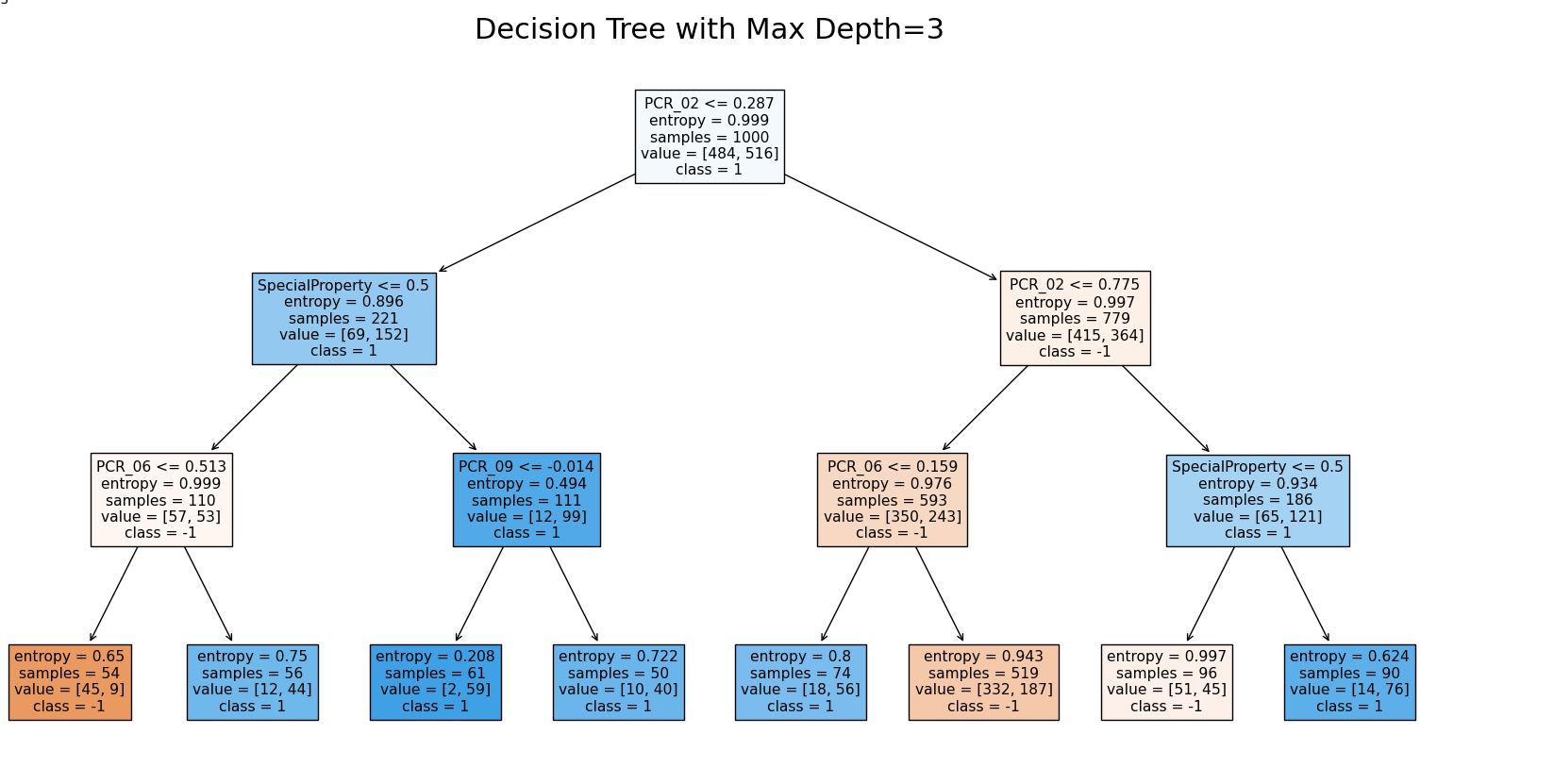
ניתן לראות שהדיוק של קבוצת האימון הוא גבוה בעוד שבדיוק על קבוצת המבחן נמוך. בנוסף גבולות ההחלטה של k=1 מתאימות באופן מושלם לקבוצת האימון. לכן, עבור k=1 מתרחשת תופעת הoverfitting, כלומר למודל אין יכולת הכללה מספיק טובה על קבוצות שונות מקבוצת האימון, והוא מתאים את עצמו לקבוצת האימון (מה שיצור שונות גבוהה של המודל על דאטה סטים שונים).

גבולות ההחלטה של המודל שלנו עם k=20 הם ״חלקים יותר״ ומצליחים לתפוס טוב יותר את ארבע אזורי ההחלטה העיקריים של הדאטה. בנוסף, אומנם מידת הדיוק על קבוצת האימון של המודל קטנה יותר אבל מידת הדיוק על קבוצת המבחן גבוהה יותר. ממצאים אלו מחזקים את הבחירה שלנו בפרמטר שמצאנו בתהליך הtuning ומצביעים על כך שמצאנו k באזור ה”sweet spot” שמאזן בין bias לvariance .

(Q5)

A grey background with white text

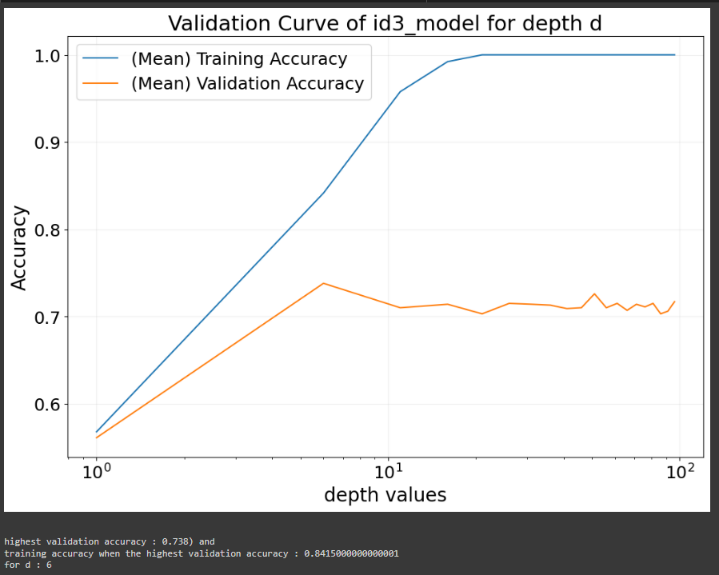
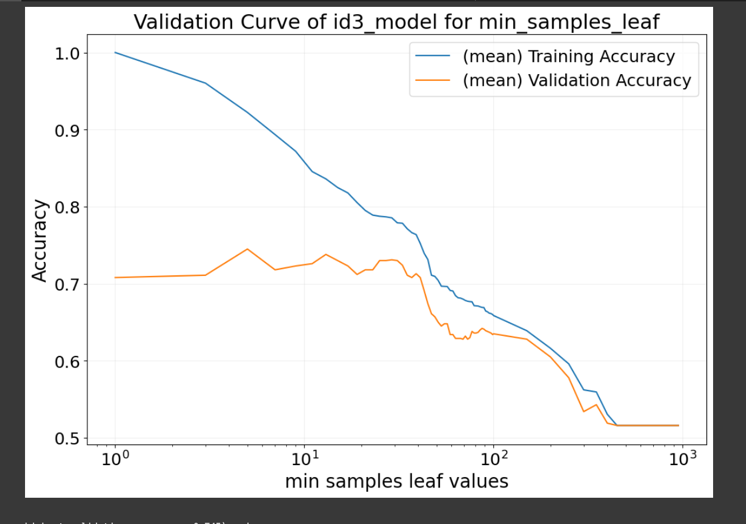
Description automatically generated



(Q6)

תחילה התייחסנו לפרמטרים כבלתי תלויים אחד בשני ובדקנו כל פרמטר בניפרד בשביל לבדוק לכל פרמטר בניפרד מתי מתקבל דיוק מקסימלי כמו שעשינו בשאלה 2 וזה כדי להבין פחות או יותר הטווחים שכדאי להשתמש בהם.

התוצאה עבור max\_depth, והתוצאה עבור min\_samples\_leaf:



A screenshot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generated

ניתן לראות כי עבור min\_samples\_leaf – תחילה יש הפרש גדול בין מידת הדיוק על קבוצת האימון לקבוצת המבחן (overfit) , לאחר מכן יש התכנסות של מידת הדיוק על האימון לכיוון מידת הוולידציה ובאזור של כמות דגימות גדולה מ-40 מתחיל להיות underfit (ירידה הן של הדיוק על האימון והן על הוולידציה).

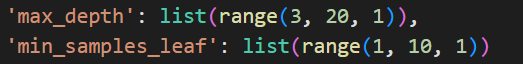
עוד נשים לב כי עם העלייה בכמות הדגימות בעלה באזור 0-40 אין עליה משמעותית ביכולת ההכללה של המודל ובמידת הדיוק על הוולידיציה (למרות שמידת הדיוק על האימון יורדת ויוצאים מאזור הoverfit). עם זאת, נראה כי באזור של 5-6 דגימות בעלה , מקבלים מידת דיוק מעט גדולה יותר.

ניתן לראות עבור max\_depth כי בעומקים נמוכים יכולת ההכללה של המודל היא נמוכה (מידת דיוק נמוכה הן על האימון והן על הוולידציה וככל שעולים עד עומקים של כ5-7 מידת הדיוק על האימון והולידיציה עולות, כאשר לאחר עומקים אלו מגיעים לאזור של overfit- מידת הדיוק על האימון ממשיכה לעלות עד למידת דיוק מושלמת באזור ה-20 ואילו מידת הדיוק על הוולידיציה נשארת פחות או יותר יציבה.

1. לאחר החיפוש הראשוני בחרנו בשני טווחים:

עבור הmax-depth נראה שאנחנו יכולים לתפוס את ה”sweet spot” באזור 3-20 (אזורים של עומק 1-2 הם נמוכים מדי ואזורים מעל 20 הם כבר ממש באזור הoverfit).

עבור הmin\_samples leaf- - בחרנו לעשות חיפוש מדויק יותר על הטווח של 1-10 כדי לנסות לראות אם ניתן לאפטם את מידת הדיוק (מאחר ובאזור 5-6 ראינו את מידת הדיוק הטובה ביותר)



**ב+ג.**

A screenshot of a graph

Description automatically generated

**קומבינציה אופטימלית:**

**ד+ו.**

**קומבינציה שגורמת לunder fitting:**

Max\_depth=3

Min\_samples\_leaf=9

עבור קומבינציה זו קיבלנו מידת דיוק נמוכה הן על קבוצת אימון והן על קבוצת המבחן.

סיבה לתופעת הunderfit על קומבינציה זו- התופעה מתרחשת כאשר למודל אין יכולת לתפוס את קשרים המורכבים בדאטה, היפר-פרמטרים מסוימים עשויים למנוע ממנו ללמוד טוב על קבוצת האימון ולכן מקבלים מודל פשוט מדי שלא מתאים לדאטה. כאשר עומק העץ קטן, המודל לא יכול לבצע מספיק החלטות כדי לבצע פרדיקציה ולכן לא יכול לחלק את הדאטה למספיק קבוצות (למשל עת בגובה 1 יחלק את הדאטה לכל היותר לשתי קבוצות). באותו אופן, כאשר המספר המינמלי שעלה יכול להחזיק הוא גדול , תהליך יצירת העץ יעצור גם כאשר האנטרופיה בעלה גבוהה יחסית וזה לא יאפשר לעץ לבצע מאוד נוספים הדרושים על מנת ללמוד קשרים עדינים בדאטה.

(נשים לב, שערך קטן יותר לmax\_depth וערך גדול יותר לmin\_samples\_leaf כנראה יתן לנו overfitting גדול יותר, כלומר דיוק קטן הרבה יותר, בגלל זה ערכים אלה לא כלולים בטווחים שלנו כי לנו צורך בהם לצורך מציאת הsweet spot) .

**ה+ו.**

**קומבינציה שגורמת לOverfit fitting:**

Max\_depth=19

Min\_samples\_leaf=1

עבור קומבינציה זו קיבלנו מידת דיוק גבוהה על קבוצת אימון ומידת דיוק נמוכה על קבוצת המבחן.

סיבה לתופעת הOverfit על קומבינציה זו-

תופעה זו מתרחשת כאשר המודל מתאים את עצמו בצורה יותר מדי טובה לקבוצת האימון, מה שפוגע ביכולת ההכללה שלו על חיזוי נקודות שהוא לא התאמן עליהם.

max\_depth מאפשר לחלק את הדאטה לקבוצות יותר מדי קטנות (למשל עבור גובה עץ 19 ניתן לחלק את הדאטה ל קבוצות. כתוצאה מכך, יש רגישות -יתר לרעש בדאטה והתאמה יותר מדי מושלמת לדאטה סט אחד , כך שיכולת ההכללה כבר נפגעת.

באופן דומה, מס׳ דגימות קטן בעלה, מאפשר למודל לייצר קבוצות אפילו עבור דגימות בודדות, שעשויות להיות רעש, ולכן גם במקרה נוצרת התאמת- יתר ופגיעה נוספת ביכולת ההכללה.

(\*נשים גם פה לב שאם Max\_depth גדל אז המודל יתאים את עצמו יותר טוב על הtraining set ויקבל דיוק 1 כמו שקיבלנו במקרה ש Max\_depth=19 אבל הדיוק על הvalidation set יהיה קטן יותר, שוב אין לנו צורך בנקודות האלה למציאת הsweet spot ולכן לא בטווח הנבחר)

(Q7)

מספר הקובינציות שנבדקו הוא כמספר הערכים האפשריים לכל פרמטר,

לפרמטר Max\_depth הטווח הוא מ-3 ל-20 (לא כולל )ולכן יש 17 ערכים אפשריים.

לפרמטר Min\_samples\_leaf הטווח הוא מ1 ל10 (לא כולל) ולכן יש 9 ערכים אפשריים.

סה"כ יש 153 קומבינציות אפשריות (17\*9).

הוספת פרמטר נוסף תגדיל את העלות החישובית ואת הזמן הנדרש למציאת השילוב האופטימלי, בנוסף פרטמר נוסף יכול להיות בעל טווחים אפשריים גדולים ואף להגדיל את הטווחים של הפרמטרים האחרים(בגלל התלות שלהם) ולכן זה לא רק להכפיל את מספר הקומבינציות האפשרי של שני הפרמטרים שלנו בטווח המספרים של הפרטמטר החדש, אלא לבדוק מחדש אלו טווחים יהיו מתאימים יותר כדי לקבל דיוק גבוה יותר בcross validation.

(Q8)

קיבלנו test accuracy of 0.772

