תמונה שמכילה טקסט, צילום מסך, תצוגה, תוכנה

התיאור נוצר באופן אוטומטי

**עבור סעיף א' :**

תמונה שמכילה טקסט, צילום מסך, תוכנה, תצוגה

התיאור נוצר באופן אוטומטי

1. **תשובות :**

**ערך הp,k הכי טובים :**  כפי שניתן לראות בתמונות למעלה הטעות האמיתית הכי טובה התקבלה עבור k = 1 כאשר עבור p = 5 כי קיבלנו את התוצאה הטובה ביותר : 0.3978 .

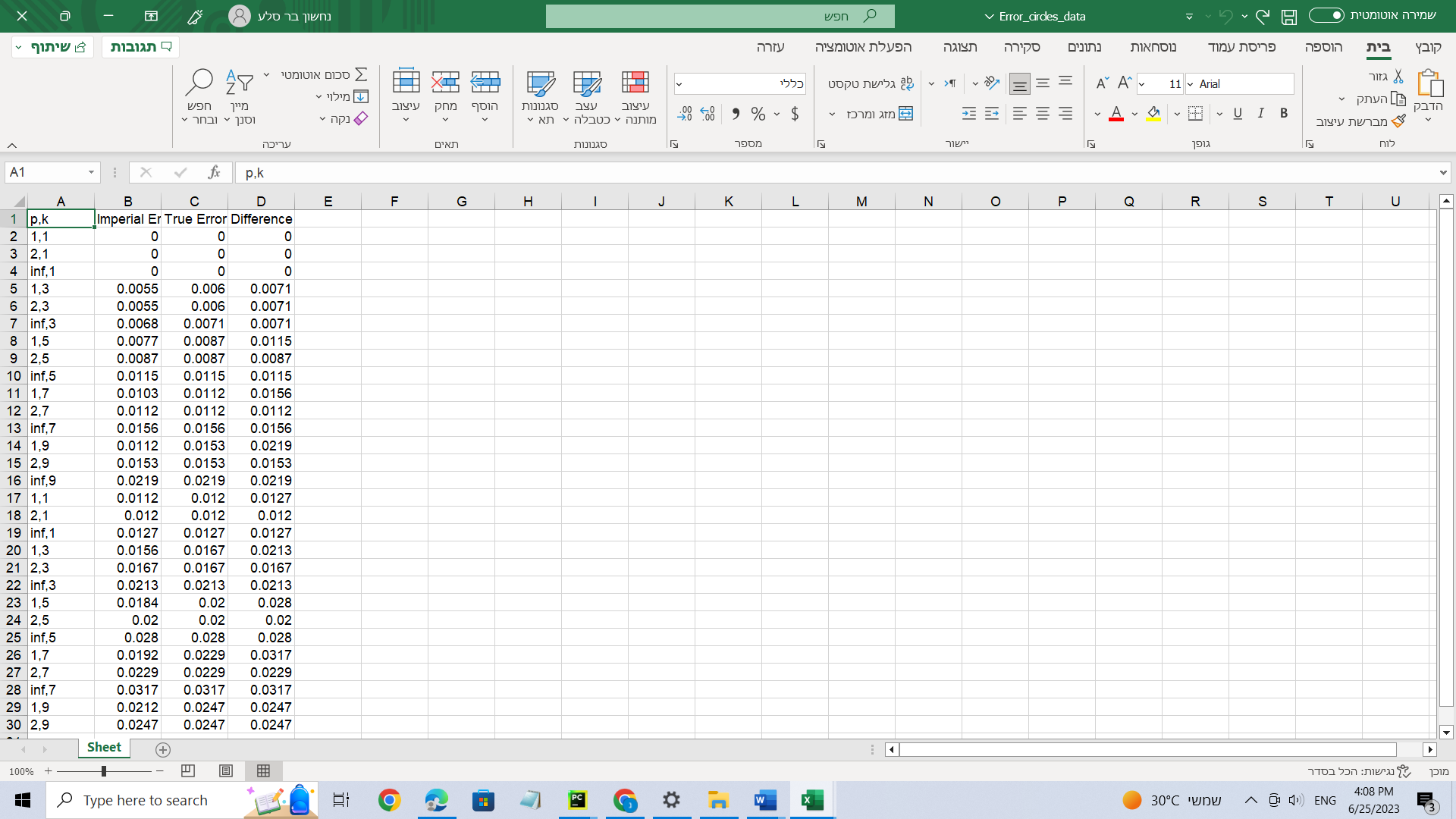
\*בנוסף , ככל שהk גדל כך הטעות האימפירית נעשה גדולה יותר , מהסיבה שהסיווג נקבע על ידי עוד נקודות מלבד הנקודה עצמה שהכי קרובה לעצמה , ולכן ההשפעה שלה על הסיווג שהיא תקבל קטן , כלומר הסתברות נמוכה יותר לתשובה נכונה .

\*דבר נוסף , ניתן להסיק שהמטריקת מרחק הטובה ביותר עבור התוצאות היא p =1 (מלבד עבור k=1 שם p=2 ניצח ) .

\*מבחינת התאמת יתר **: עבור k =1 יש overfiting** כאשר לרוב הנקודה הכי קרובה הייתה הנקודה הנבחנת בעצמה ולכן הimprial error מאוד נמוך . ולעומת זאת הטעות האמיתית הכי גבוהה שם , מהסיבה שאין שום נקודות אחרות שיכולות "לאזן את הרעש" של הנקודה הכי קרובה , לנקודה הנבחנת .:

עבור **k=3**  : גם במקרה זה, בכל שלושת המקרים, אנו יכולים לראות שהשגיאה האימפירית נמוכה משמעותית מהשגיאה האמיתית . אך ההתאמת יתר משמעותית פחות קיצונית מאשר עבור k =1 .

**\*באותו אופן ,** ככל ש – k גדל כך יש פחות התאמת יתר של המודל למערך נתוני האימון , אך עדיין המודל נותן תוצאות טובות יותר למדגם באופן עיקבי .

**תמונה שמכילה טקסט, צילום מסך, תצוגה, תוכנה

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, תצוגה, תוכנה

התיאור נוצר באופן אוטומטיעבור סעיף ב' :**

ההבדלים שונים מהותית מהסעיף הקודם, ראשית כל מבחינת אחוז הדיוק בטעות האמיתית (וגם בטעות האימפירית ) כאשר אלגוריתם הKNN מספק תוצאות **הרבה יותר טובות** עבור הדאטה סט מאשר Caesarian data set עבור כל p,k נתונים .

הסיבה לכך שמערך הנתונים two\_circles בנוי כך שבהינתן נקודות p1 השייכת t wo\_circles ובנוסף נקודה p2 השייכת למערך הנתונים Caesarian data set :

יש הרבה יותר משמעות לקירבה בין הנקודות שקרובות לp1 במערך הנתונים two\_circles על הסיווג של 1 p , ביחס למידת השפעת הקרבה של הנקודות ב Caesarian data set שקרובות לp2 על הסיווג של p2 .

דוגמא קיצונית, ניתן לראות ב ב Caesarian data setשהדגימה :

25,1,0,0,0,1

והדגימה : 25,1,0,0,0,0

שניהן מופיעות במערך הנתונים .

כלומר שני דגימות שונות שהמרחק בינהם עבור על p שבחנו הוא המינימלי ביותר שאפשר (0) אבל עדיין הסיווג שלהן שונה (0 מול 1 )

לגביי overfiting :

ניתן לקבוע שאין .

הדיוק בtrue error מאוד קרוב לדיוק בimprial error לכל k ו- p ולפעמים אף משתווה לו .