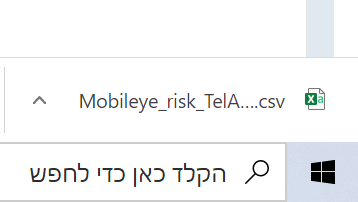
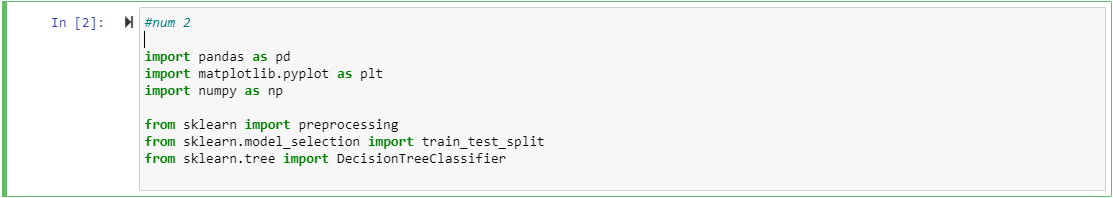
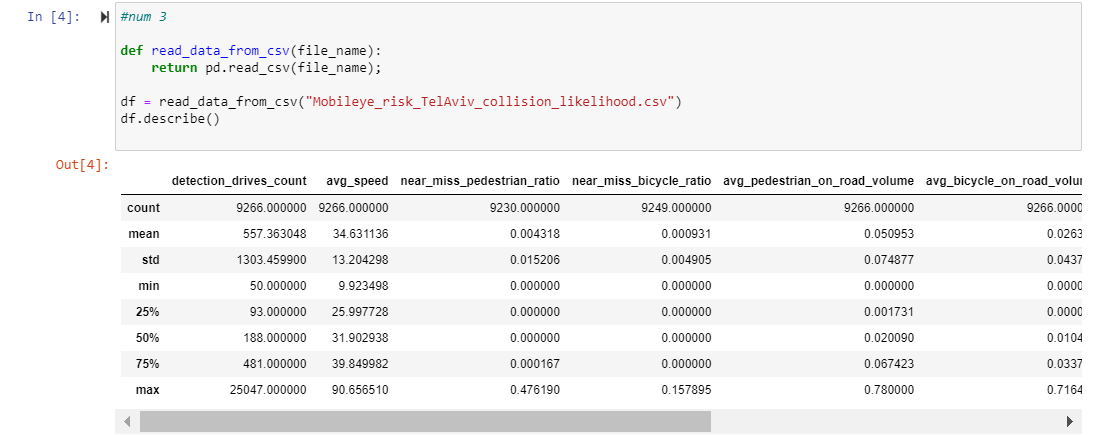
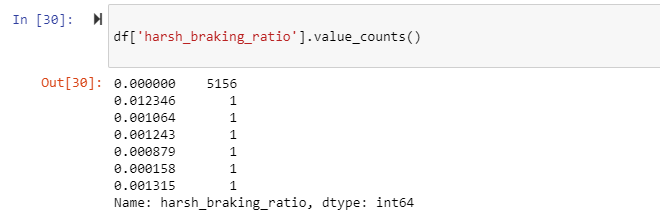
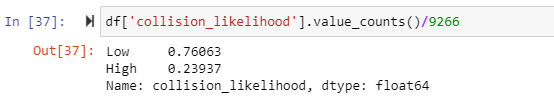
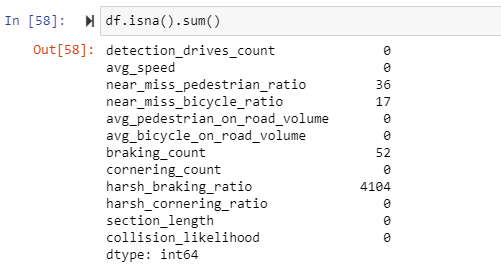
מגיש : ערד קוצר  
ת"ז: 311226203  
מסלול: מדמ"ח יזמות  
**חלק 1**  
**שאלה 1**   


**שאלה 2**

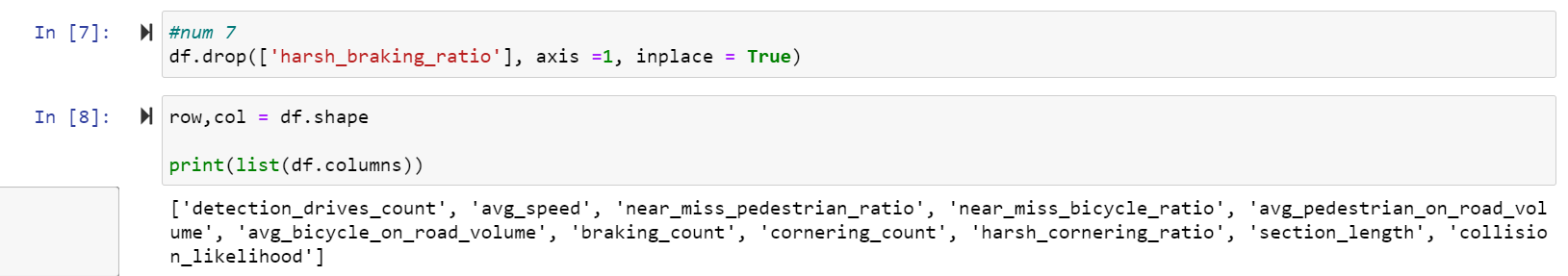
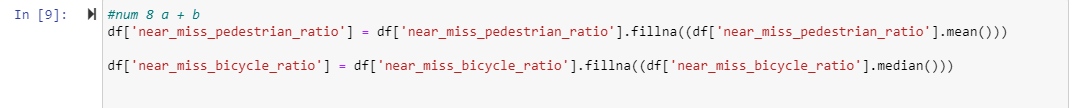


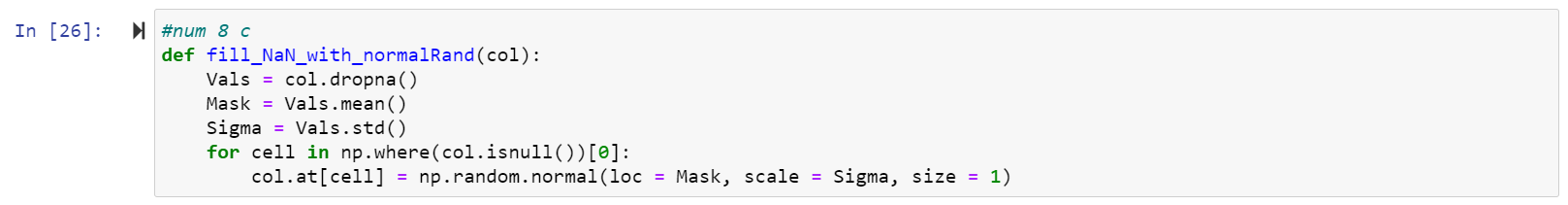
**שאלה 3  
שאלה 4**  
  
ניתן לראות שבעמודה 'harsh\_braking\_ratio' למרות שהערכים אמורים להיות מנורמלים- (לפי המילון) רק - 5162 ערכים קיימים, כלומר חסרים לנו נתונים עבר כמעט חצי מהערכים בעמודה זאת. מעבר לכך שגם בעמודה זאת פרט לשישה ערכים השונים מ- 0 כול הערכים ריקים. לכן לדעתי ניתן להשמיט בהמשך עמודה זאת**.**

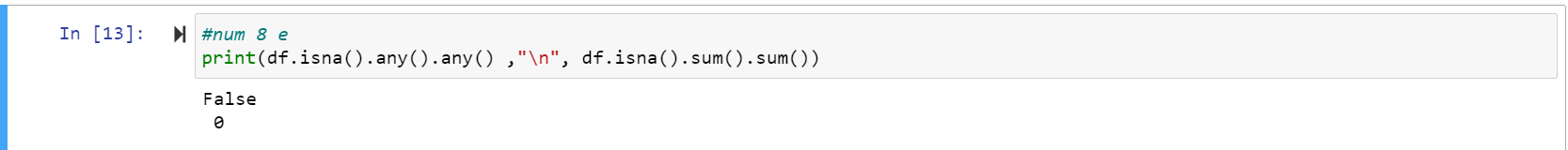
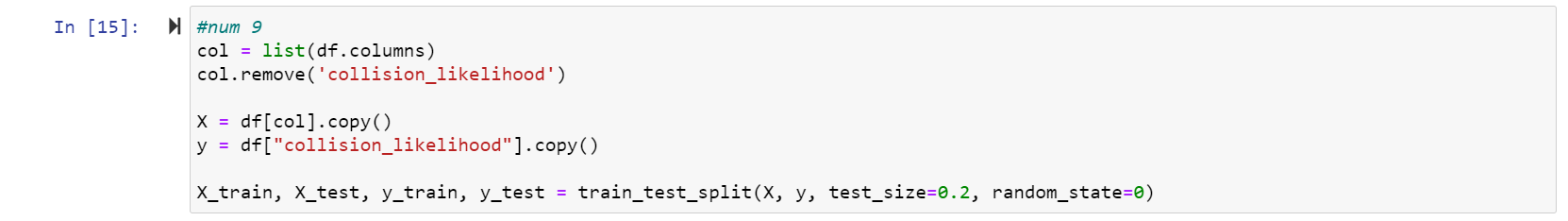
**שאלה 5  
**   
ניתן לראות שיש לנו 76% אחוז של קטעי עם סיכון נמוך וכמעט 24% עם סיכון גבוה להתנגשות**.**

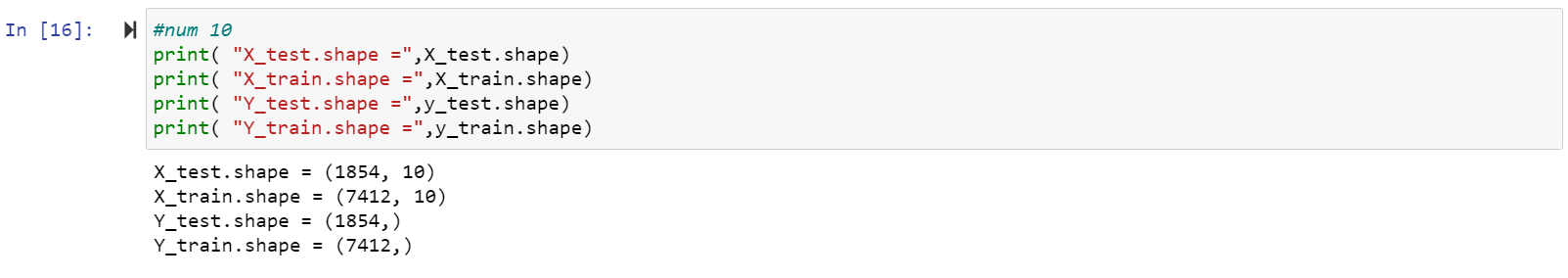
**שאלה 6**a – ניתן לראות שהעמודה 'harsh\_braking\_ratio' מכילה 4104 ערכיי NULL כלומר יש לה הכי הרבה ערכים חסרים.   
b – הדרך הכי נפוצה לעבד את הערכים במצב זה הינה למלא את התאים הריקים על ידי fillna כך שנמלא בתאים הריקים את הממוצע לכול תא בכול עמודה.

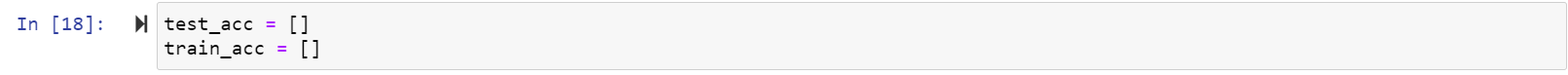
**חלק 2:**

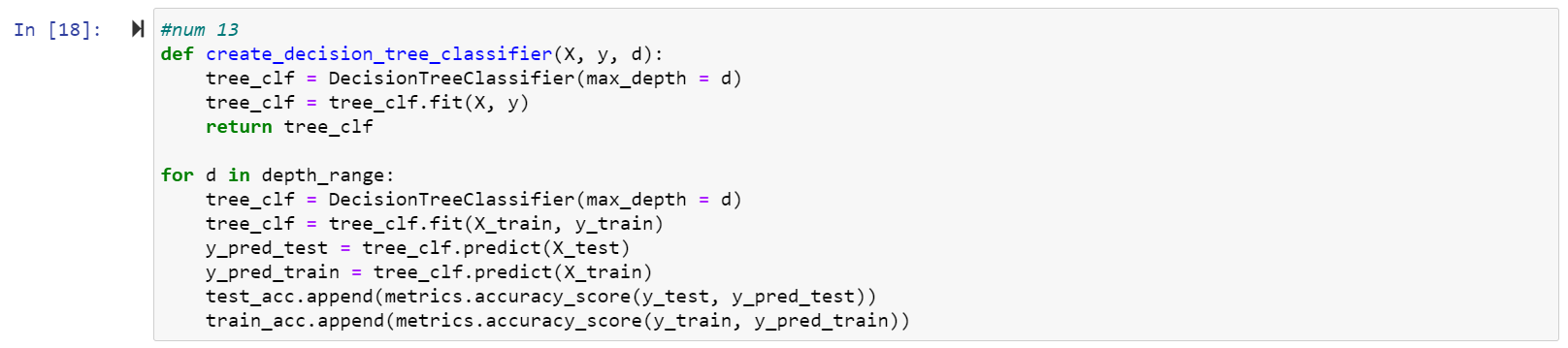
שאלה 7  
  
  
שאלה 8 a + b -   
  
שאלה 8 - c

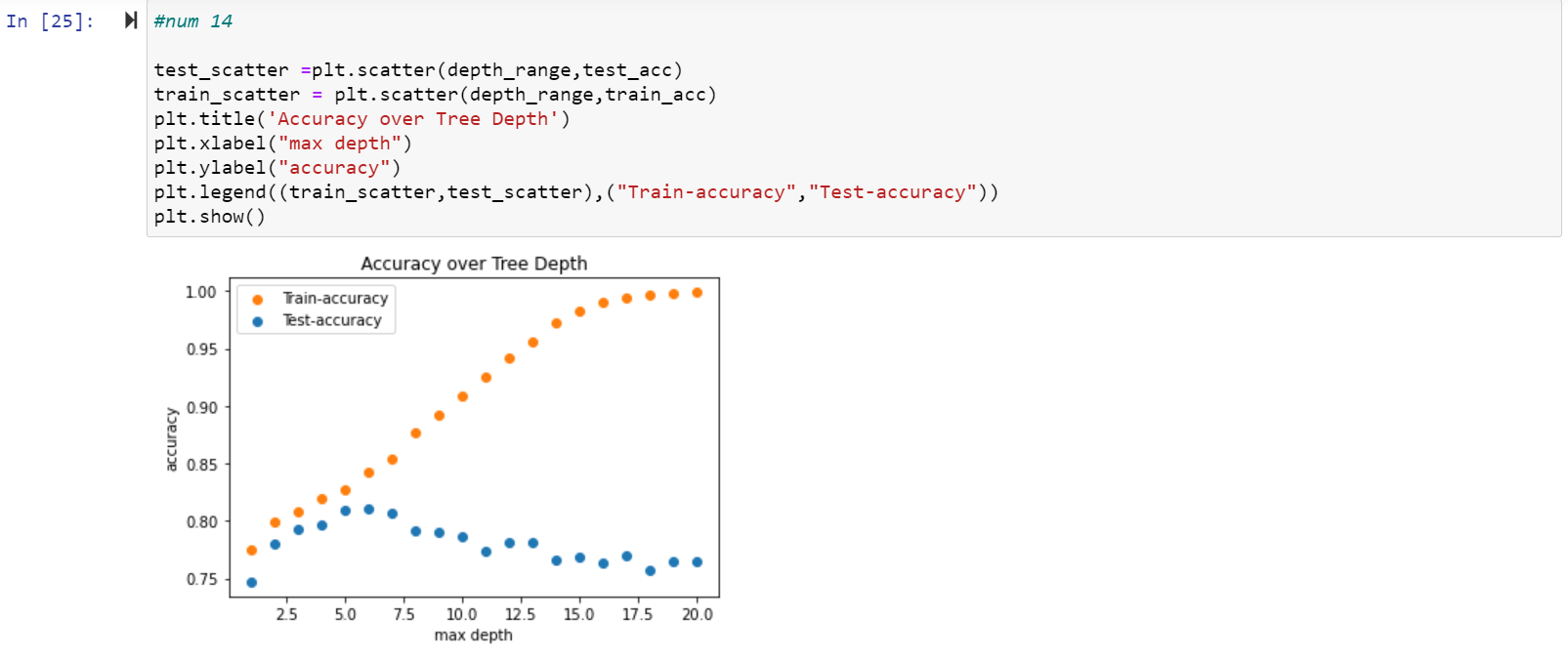


שאלה 8 – d  
  
  
  
שאלה 8 – e  
  
שאלה 9  
 

שאלה 10  


**חלק 3 א' :**שאלה 11  
  
שאלה 12  

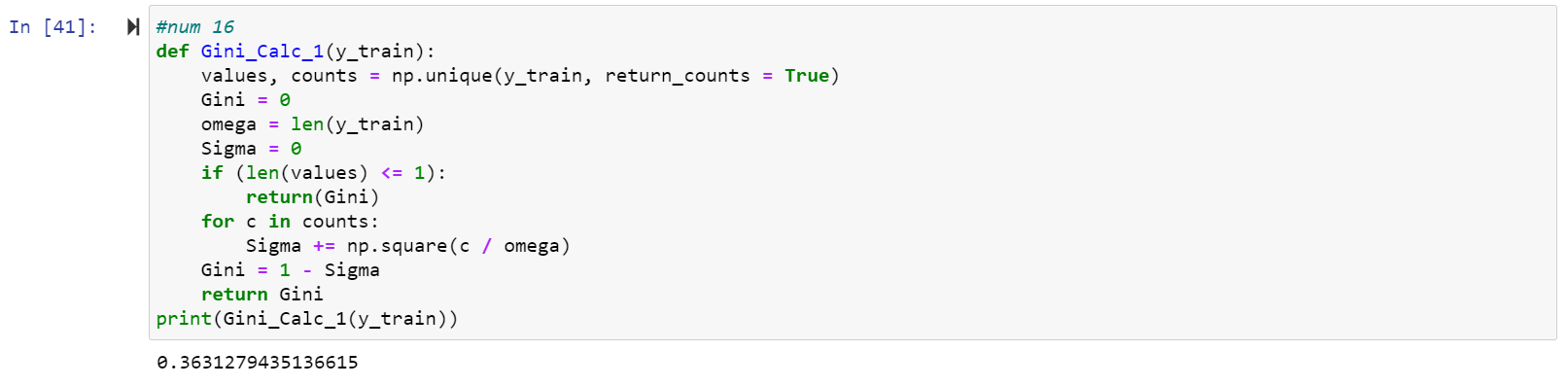
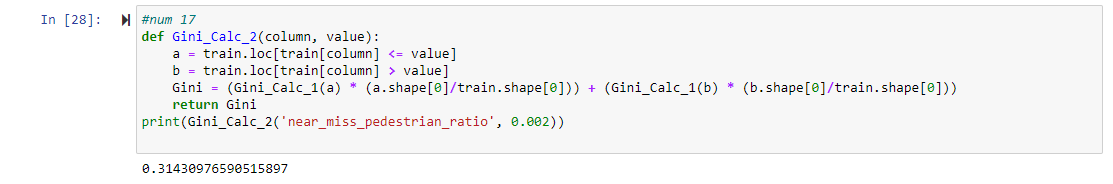

שאלה 13  


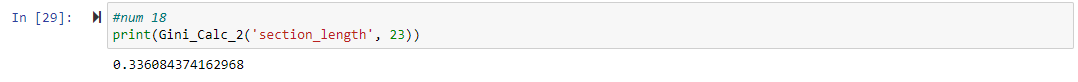
שאלה 14  


שאלה 15-a:  
ניתן לראות שקיימת מגמת עליה בסט של ה- Train לכול איטרציה נוספת בעץ. כלומר במקרה זה אנחנו מגיעים לדיוק של כמעט 100 אחוז לכן כאשר נרוץ על גובה העץ עד לפיצול ה- 20 נגיע לתנאי עצירה. שיכול להיות או שאין לנו עוד עלים, אין עוד דרך לחלק את העץ או שכול פיצול נוסף יכניס אותנו לאותה משבצת שכבר קיימת ובכך לא ישנה את בחירת הענף בעץ.  
לכן כדאי לנו לנסות "לתפוס" את הפיצול המיטבי ביותר למודל (כמה אטרציות יתנו לנו את התוצא הטובה ביותר) נרצה לעצור בנקודות הברך שלו שהן סביב 5 או סביב 13 כדיי להימנע "מאוברפיטינג".

שאלה 15-b:  
ניתן לראות שיש לנו מגמה של עליה עד לגובה של 7 ומשם מגמה של ירידה ודשדוש בסט של ה- Test שלנו. לדעתי מצב זה נובע מ – "אוברפייטינג" ככול שאנחנו ממשיכים לחלק את ענפיי העץ.   
לכן ניתן לראות גם כאן בדומה לסעיף a שכנראה שנקודת "הברך" הטובה ביותר היא לאחר 5 חלוקות, מכוון שרק בנקודה זאת יש לנו "נקודות פיתול" משותפת לשני הגרפים.

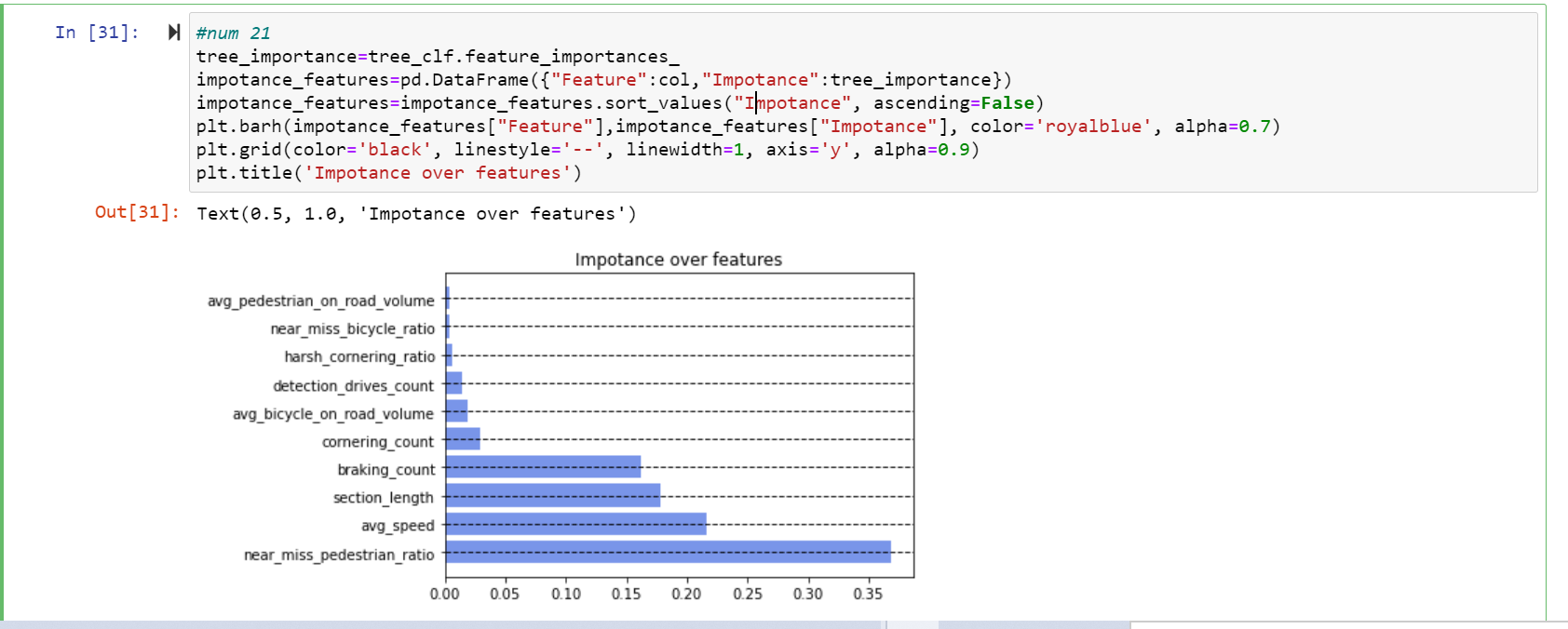
שאלה 15-c :  
כפי שהסברתי בסוף סעיף b , כנראה שהכי כדאי לנו לבחור נקודה בין 5-7 מכוון ששם יש לנו את נקודת הפיתול בשני הגרפים. לדעתי במצב זה מכוון שיש עליה יותר חזקה בגרף האימון וטווח הטעות הוא לא כול כך משמעותי בtest , אז ההייתי לוקח נקודה מימין לנקודת הפיתול – כלומר 7 פיצולים.

**חלק 3 ב':**שאלה 16: ****  
שאלה 17:  
****

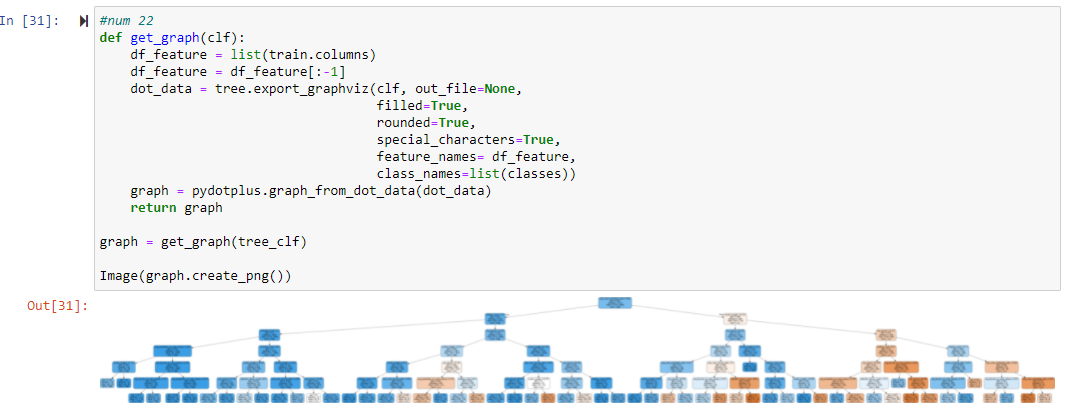
שאלה 18:  
****

שאלה 19:  
הסבר – עץ ההחלטה בוחר בתור התחלה איזה תכונה בעץ יש לנו עליה הכי הרבה מידע. כלומר הוא בודק איזה פיצ'ר הוא הכי רלוונטי – כלומר יש לו נוסחה שבודקת איזה פיצ'ר הכי חשוב מכול הפיצ'רים הקיימים- וזאת הוא בודק ע"י מדד הג'יני וכמה המדד תואם את התוצאה בגרף.  
כלומר התוצאות של מדד ג'יני הן יחסית צפויות והאלגוריתם בודק מהו המרחק המינימלי בין כול תכונה של העץ החלטה שלנו, שבו לכול פיצ'ר ניתן ציון בין 0 – 1 ואז האלגוריתם בודק האם האנטרופיה בין המדד הקבוע של ג'יני לבין התוצאה שקיבלנו בעץ היא הקטנה ביותר.  
בסופו של דבר האלגוריתם בודק לכול ג'יני אינדקס מהו הפיצ'ר הטוב ביותר ונותן לו ציון (בין 0-1 ) ובכך בוחר לכול צומת את הפיצ'ר הטוב ביותר, מוחק את הפיצ'ר מהרשימה וחוזר חלילה- עד לתנאי עצירה.

שאלה 20:  
הערך ג'יני של near\_miss\_pedestrian\_ratio הוא הגבוה ביותר, מיכוון שהתוצאה לחזות אותו היא הכי גבוהה ולכן גם הקורלציה בינו לבין מדד הג'יני היא הגבוהה ביותר. לכן נבחר בפ'יצר הראשון להיות - near\_miss\_pedestrian\_ratio.

שאלה 21:  
 

**חלק 3 ג':**

שאלה 22:  
  
שאלה 23:

הסבר -   
ניתן לראות שככול שהנתיב בעץ יותר כחול כהה אז יש יותר סבירות להצלחה – כלומר אם נבחר את הנתיב הכחול ביותר נצליח להגיע לאחוזים גבוהים ביותר להצלחה – (הסברות להצלחה- גבוהה ביותר). לכן קל לקחת עקרון זה בתור כלל עצבה בעץ שציירנו (תמיד ללכת לכיוון של הקודקוד הכחול).  
כמובן שלהפך גם – ככול שנלך לקודקודים בצבע כתום יותר אז יש פחות סבירות להצלחה.  
קיימים שישה מסלולים כחולים שמגיעים לסבירות גבוהה – כלומר מגיעים לאוברפייטינג על הדאטה סט שלנו.   
בחרתי להראות את השניים עם הסבירות הגבוהה ביותר.  
הנתיבים הטובים ביותר שראיתי בעץ הם –   
ראשון -

near\_miss\_pedestrian\_ratio 0.002 (True) cornering\_count 2.5 (True) avg\_pedestrian\_on\_road\_volume 0.028 (False) avg\_bicycle\_on\_road\_volume 0.019 (False) avg\_pedestrian\_on\_road\_volume 0.076 (False)

שני –

Near\_miss\_pedestrian\_ratio < 0.002(False) -> selection\_length<24.702(False) -> avg\_speed<33.13(True) -> breaking\_count<92.5(False) -> avg\_bicycle\_on\_road\_volume<0.002(False) -> avg\_speed<28.265 (True)

שאלה 24:

