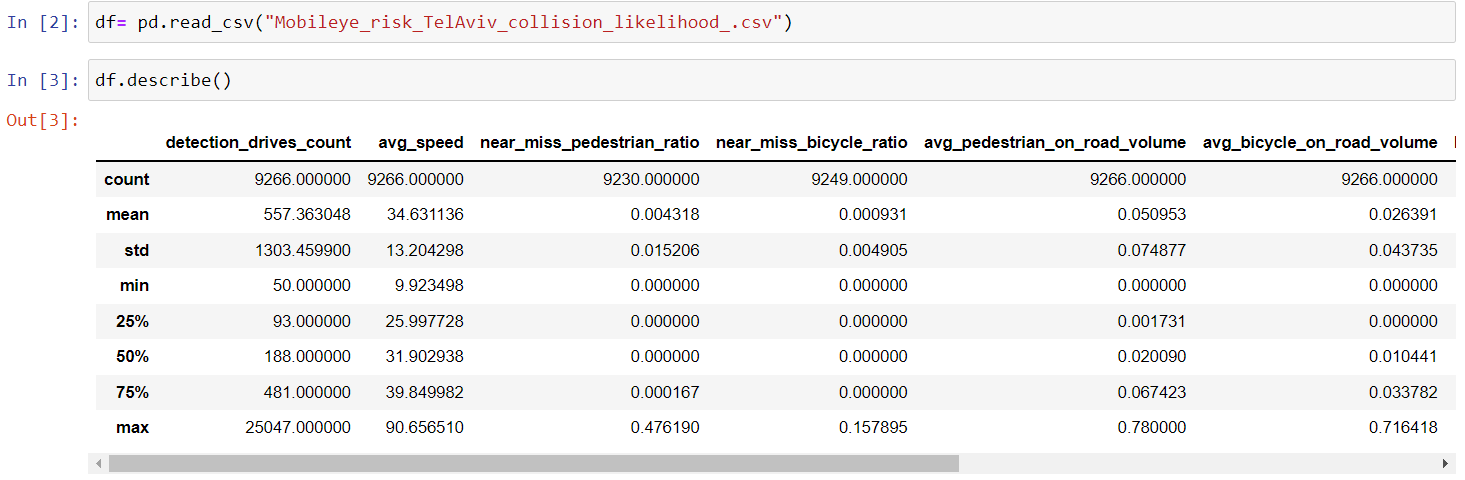
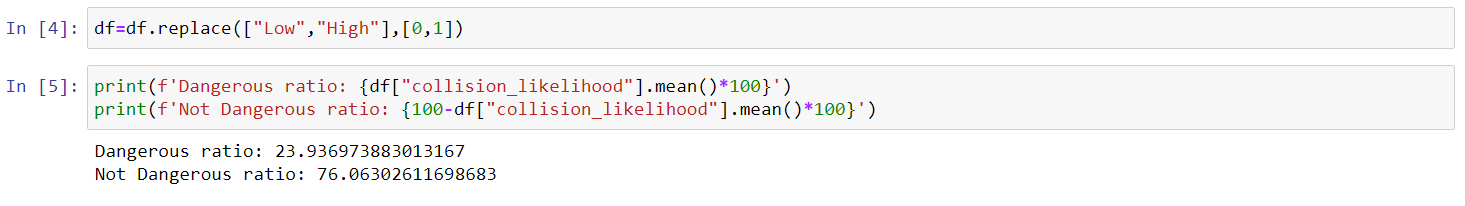
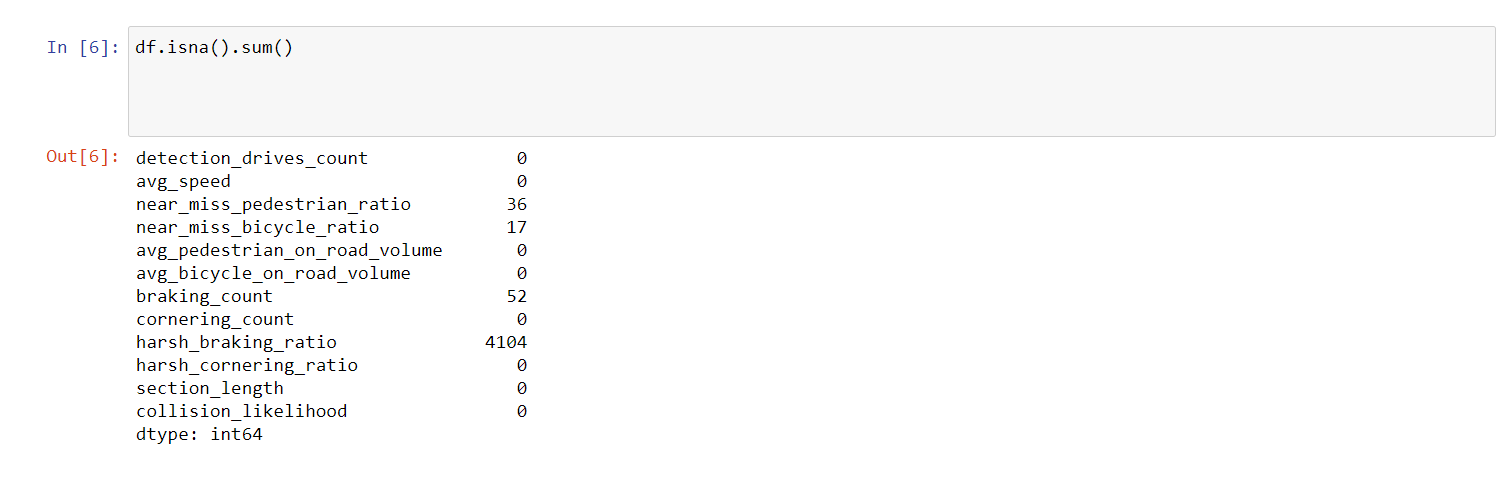
**מטלה 4 - עץ החלטה**  
**מגיש: נדב וילדר 312173776**

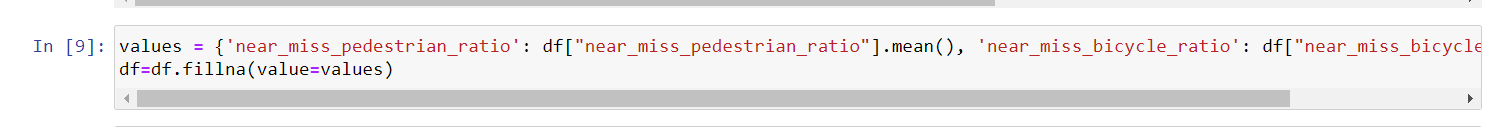
**חלק ( 1 - קריאת נתונים + סטטיסטיקה בסיסית):**

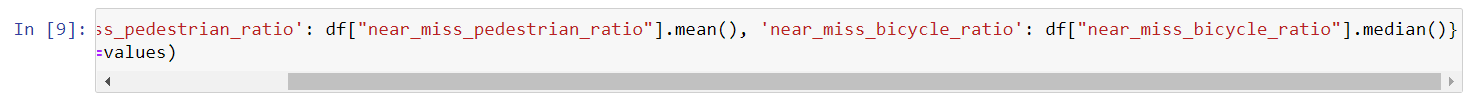
1. הורדתי
2. 
3. 
4. ניתן לראות שלעמודה harsh\_braking\_ratio יש טווח ערכים מאוד קטן סביב האפס. היא לא מספקת לנו הרבה מידע. לכן אולי שווה להוריד אותה.

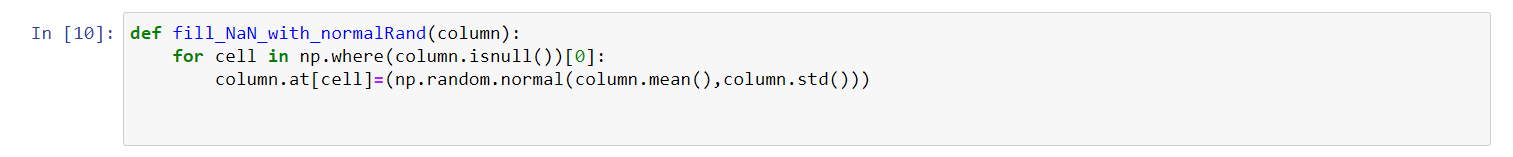
 ניתן לראות שאחוז המקטעים המסוכנים הינו: 24% בעוד שהלא מסוכנים הוא 76%.

1. 
2. harsh\_braking\_ratio חסר הכי הרבה ערכים (4104)
3. את עמודה harsh\_braking\_ratio נסיר מבסיס הנתונים שלנו, היא מלאה בתאים ריקים וגם מה שלא ריק, חסר מידע רלוונטי.   
   עבור שאר העמודות, הייתי מכניס את החציון על מנת לא לפגוע בחישוב.

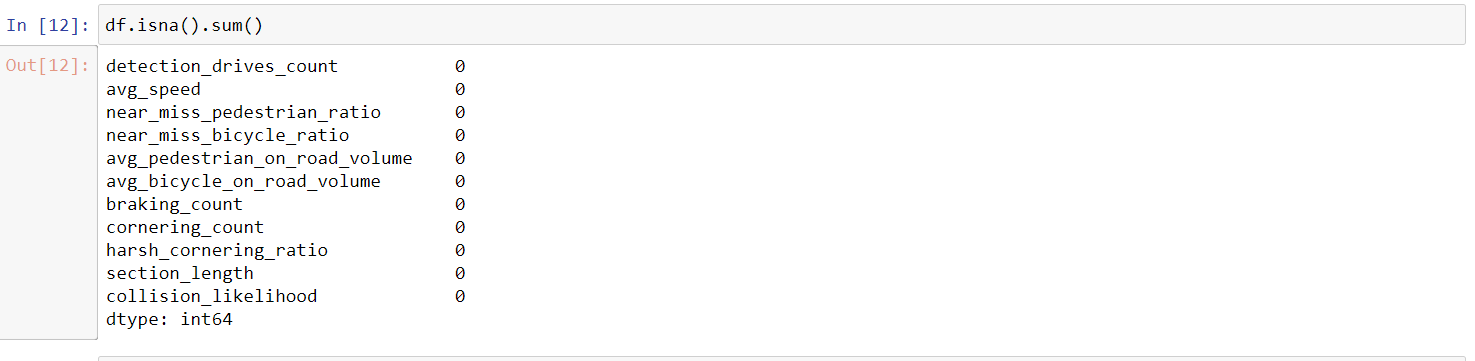
**חלק 2 - Preprocessing :**

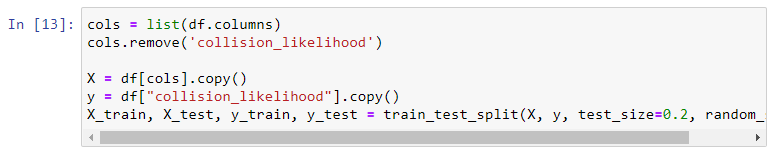
1. 
2. a. 

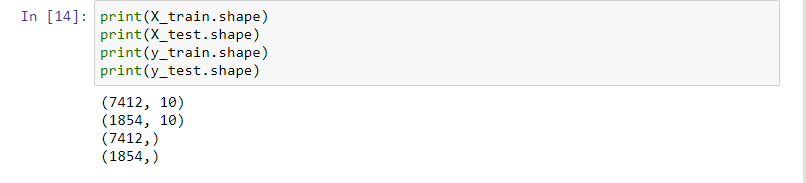
b. 

c. 

d. 

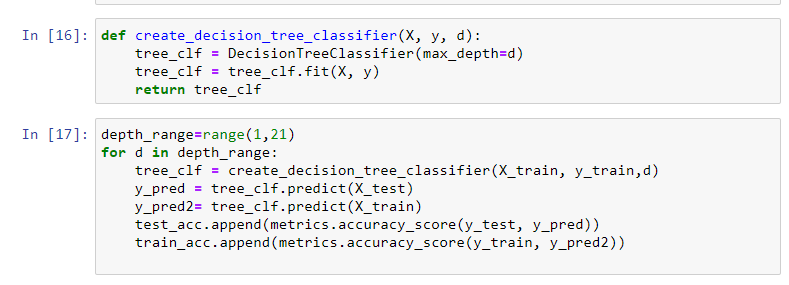
e. 

9. 

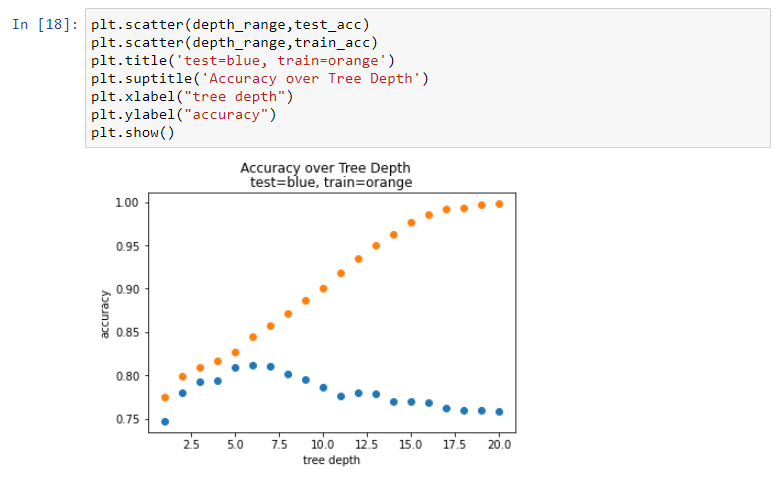
10. 

**חלק 3-Decision Tree:**

11. 

12+13. 

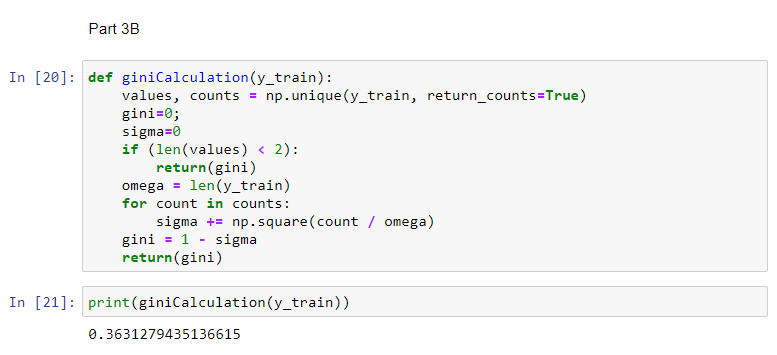
14.

15. a. יש מגמה חיובית בסט האימון(כתום) ככל שעומק העץ גדל, כך רמת הדיוק עולה. זה קורה בשל העובדה שככל שעומק העץ גדל יש "אוברפיטינג" במדד הדיוק.

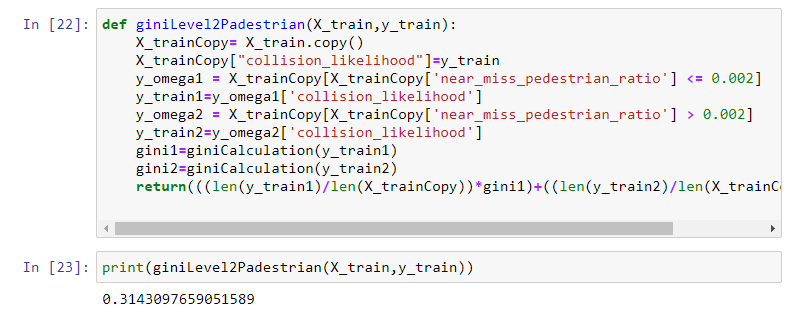
b. אין איזו מגמה מיוחדת, ניתן לראות שמדד הדיוק נשאר פחות או יותר באותו תחום.

c. הייתי ממליץ לבחור בעומק עץ 6 כי ניתן לראות שבעומק זה מדד הדיוק בסט הבדיקה הכי גבוה.

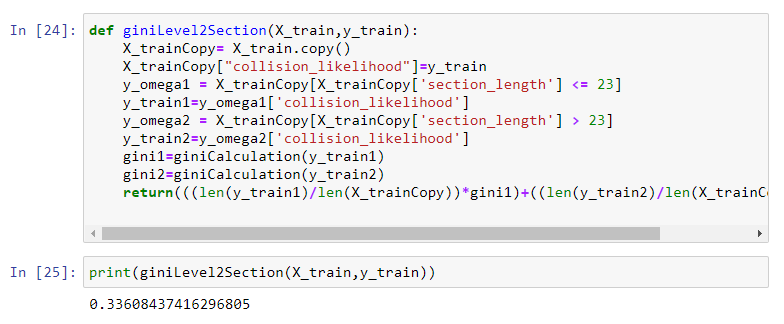
**חלק 3 ב' – חישוב ג'יני - Gini Impurity :**

16.

מדד הג'יני (שלב ראשון) בסט האימון היא: 0.363

17. 

מדד הג'יני שלב 2 עבור המשתנה near\_miss\_pedestrian\_ratio כך ש X<=0.002 הוא 0.3143

18. 

מדד הג'יני שלב 2 עבור המשתנה section\_lengthכך ש 3X<=2 הוא 0.336