Car Safety Ratings

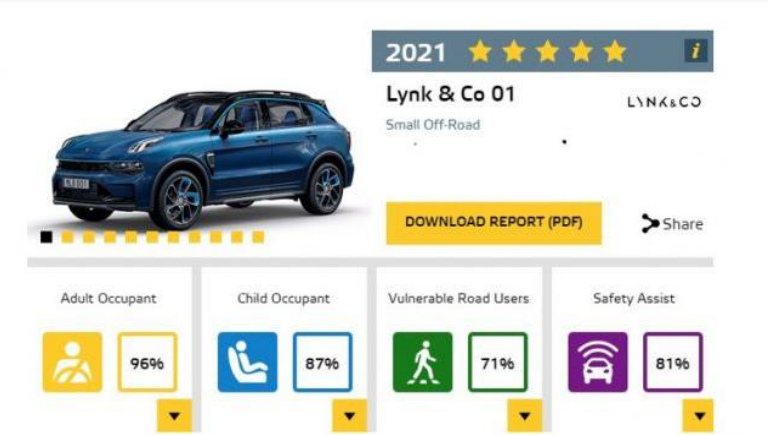
By: Valeria Shafran & Einav Kogut

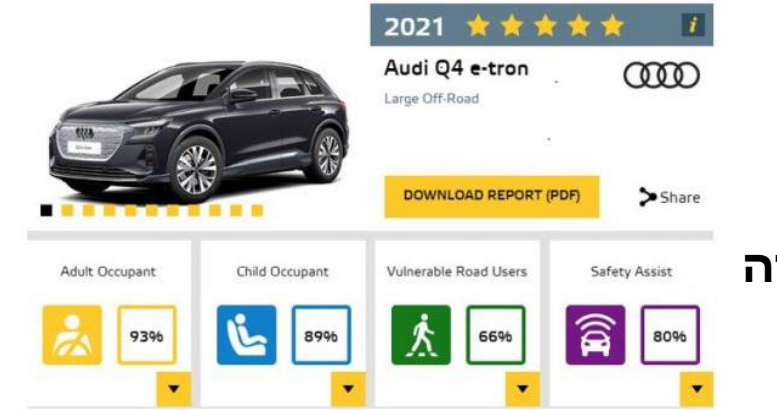
פרויקט סיום בקורס מבוא למדעי הנתונים

שאלת המחקר:

חיזוי חברת הרכב שתיצור את הרכב הכי בטיחותי

Introduction





-בין הדברים הכי חשובים ברכישת רכב חדש הוא רמת הבטיחות של הרכב ומערכות הבטיחות השונות.

-חיזוי ייצור רכב בטיחותי.

-חיזוי חברת הרכב שתיצור את הרכב הכי בטיחותי, על פי מאפיינים של מבחני בטיחות שונים.

-ניתוח נתוני הרכב לצורך הכרחי לרמת בטיחות ומיגון.

-ניתוח נתוני חברות רכב שונות להשוואת מערכות הבטיחות והחברה שעשויה לייצר רכב ברמת הבטיחות הגבוהה ביותר.

מקורות הנתונים וההרכשה

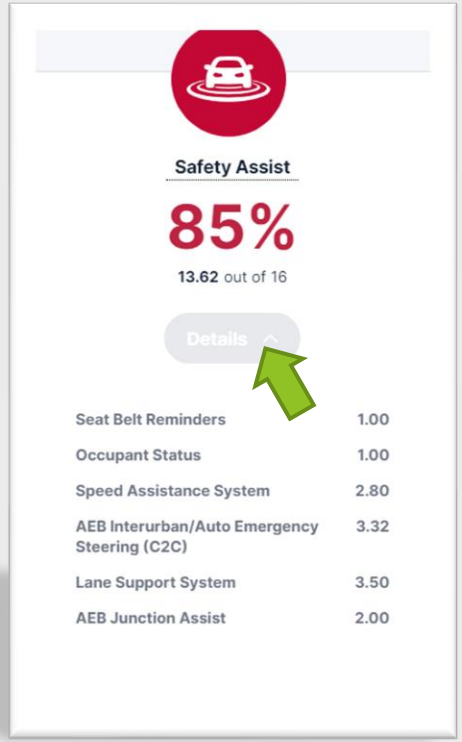
https://www.euroncap.com

https://www.ancap.com.au/safety-ratings

https://www.iihs.org/ratings/top-safety-picks/

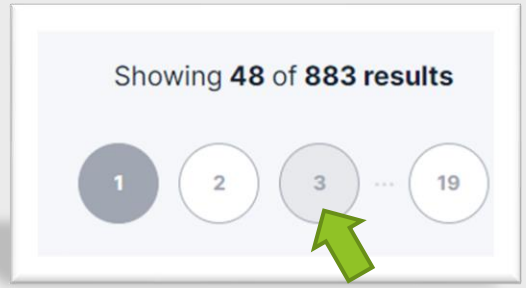
- גישה לשלושה אתרים מגוונים של מבחני בטיחות של דגמי רכב שונים, חילוץ הנתונים בעזרת selenium.
- העברת הנתונים מאתרי האינטרנט לקובץ csv ולDataFrame.
- הרכשת נתוני ID וכלל המפרט של כל רכב על פי חברת ייצור, שנת ייצור, סוג הרכב, משקל, דירוג ומערכות הבטיחות השונות.

מקורות הנתונים וההרכשה



-בתחילת הניסיון להרכשה נתקלנו בבעיה לחילוץ הנתונים ומעבר בין קטגוריות בשימוש ספריית beautiful soup.
נעזרנו בפתרון הבעיה בשימוש ב selenium,
לדוגמא: קליק בהצגת נתונים שונים.

דוגמא נוספת: מעבר בין רכב לרכב ובין עמודים להרכשת כל הנתונים.



-לאחר שלב ההרכשה ביצענו נרמול לערכים הנומריים שלנו על מנת לבצע השוואה ולייצג את ההבדלים בין החברות השונות באמצעות גרפים שונים.

רכב לדוגמה

ASSESSMENT SCORES

Adult Occupant Protection	Child Occupant Protection	Vulnerable Road User Protection	Safety Assist
88%	89%	81%	85%
33.47 out of 38	43.81 out of 49	43.85 out of 54	13.60 out of 16
Details	Details	Details	Details

SAFETY ASSIST TECHNOLOGIES

AEB - City	AEB - Interurban	AEB - VRU	AEB - Junction Assist	AEB - Backover	LSS
✓	✓	✓	✓	✓	✓
Standard on all variants	Standard on all variants	Standard on all variants	Standard on all variants	Standard on all variants	Standard on all variants

GWM

GWM Tank 300

Build Dates	On Sale Dates	Applies To
Sep 2022 - onwards	Dec 2022 - onwards	Hybrid variants
Vehicle Type	**Engine/Fuel Type**	**Price Bracket**
Large SUV	Hybrid vehicles	$35,000 - $65,000
Rating Year/Datestamp	**ANCAP Safety Rating**	**Rating Expires**
2022	★★★★★	Dec 2028

ANCAP SAFETY

TESTED 2022 ★ ★ ★ ★ ★

פרמטרים להשוואה

מודל הרכב	חברת הרכב	שנת יצור	משקל
דירוג כוכבים	מערכת סטייה	בקרת מהירות	מערכת בלימה
דירוג בטיחות ילד	דירוג בטיחות מבוגר	דירוג מערכת בטיחות	דירוג בטיחות הולך רגל
	סוג הרכב	סיווג הרכב	

ניתוח ראשוני וטיוב

-יצירת Dataframe כל שורה מייצגת רכב וכל עמודה מייצגת את הפרמטרים להשוואה של הרכבים.

-הסרת Duplications.

-מילוי ערכים מספריים חסרים: עבור ייצוג העמודות שבהם יש ערך בוליאני (1\0), בהם הערכים היו חסרים הוספנו את הערך 0 בעזרת הפונקצייה fillna.

-המרת ערכי str של אחוזים לint.

‘%64’ -> 64

-בדיקה של עמודות בעלות מידע מועט: כלומר מידע רב של ערכי NaN - רכבים אשר לא קיבלו דירוג בכלל לא היו רלוונטיים עבורנו ולכן ביצענו הסרה שלהם מאחר ולא יכולנו לחשב עבורם את ה score.

ניתוח ראשוני וטיוב

Data Frame ראשוני לפני סינון המידע

	CarCompany	TestedModel	VehicleType	CarClass	YearOfPublication	KerbWeight	Stars	AEBCar2Car	SpeedAssistance	LaneAssistSystem	AdultOccupantR	ChildOccupantR	VulnerableRoadUsersR	SafetyAssistR
0	Aiways	Aiways U5	- 5 door SUV	Small Off-Road 4x4	2019	1750kg	5	1.0	1.0	1.0	73%	70%	45%	55%
1	Alfa	Alfa Romeo Tonale	- 5 door SUV	Small Off-Road	2022	1626kg	5	1.0	1.0	1.0	83%	85%	67%	85%
2	Alfa	Alfa Romeo Giulietta	- 5 door hatchback	Small Family Car	2017	1355kg	5	0.0	0.0	0.0	72%	56%	59%	25%
3	Alfa	Alfa Romeo Stelvio	- 5 door SUV	Large Off-Road	2017	1745kg	5	1.0	1.0	1.0	97%	84%	71%	60%
4	Alfa	Alfa Romeo Giulia	- 4 door saloon	Large Family Car	2016	1449kg	5	1.0	1.0	1.0	98%	81%	69%	60%
3391	Volvo	Volvo S80 (2000)	NaN	Executive	2000	NaN	5	0.0	0.0	0.0	NaN	NaN	NaN	NaN
3392	Volvo	Volvo S70 (1998)	NaN	Executive	1998	NaN	5	0.0	0.0	0.0	NaN	NaN	NaN	NaN
3393	Volvo	Volvo S40 (1997)	NaN	Large Family Car	1997	NaN	5	0.0	0.0	0.0	NaN	NaN	NaN	NaN
3394	WEY	WEY Coffee 02	- 5 door SUV	Small Off-Road	2022	2100kg	5	1.0	1.0	1.0	94%	87%	73%	93%
3395	WEY	WEY Coffee 01	- 5 door SUV	Large Off-Road	2022	2365kg	5	1.0	1.0	1.0	91%	87%	79%	94%

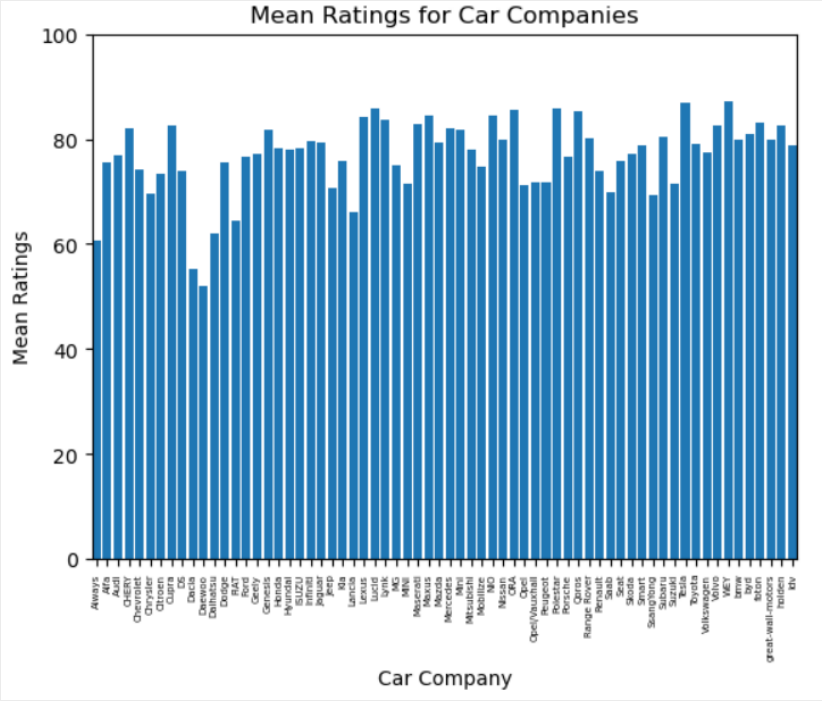
ניתוח ראשוני וטיוב

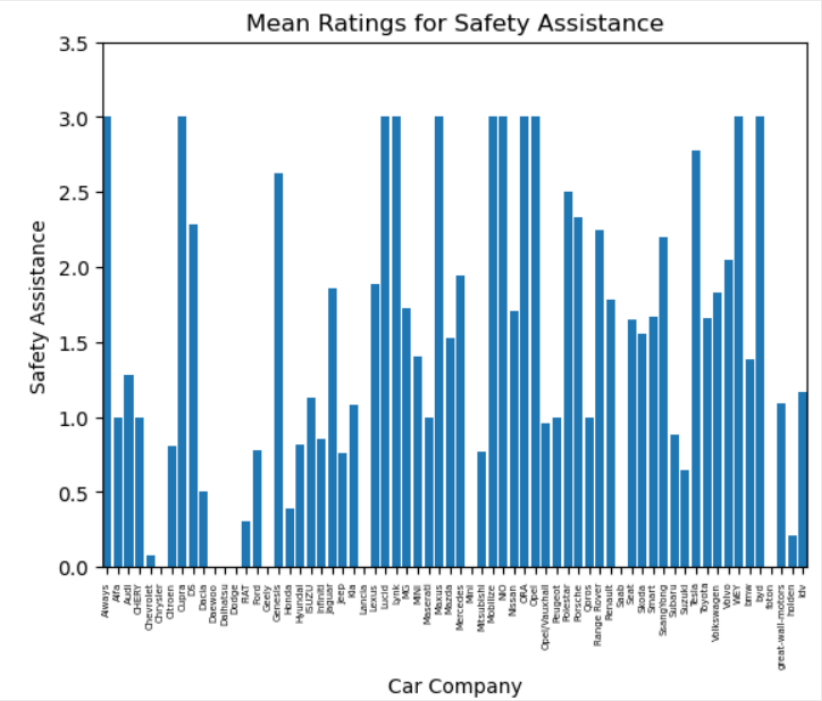
Data Frame אחרי סינון המידע

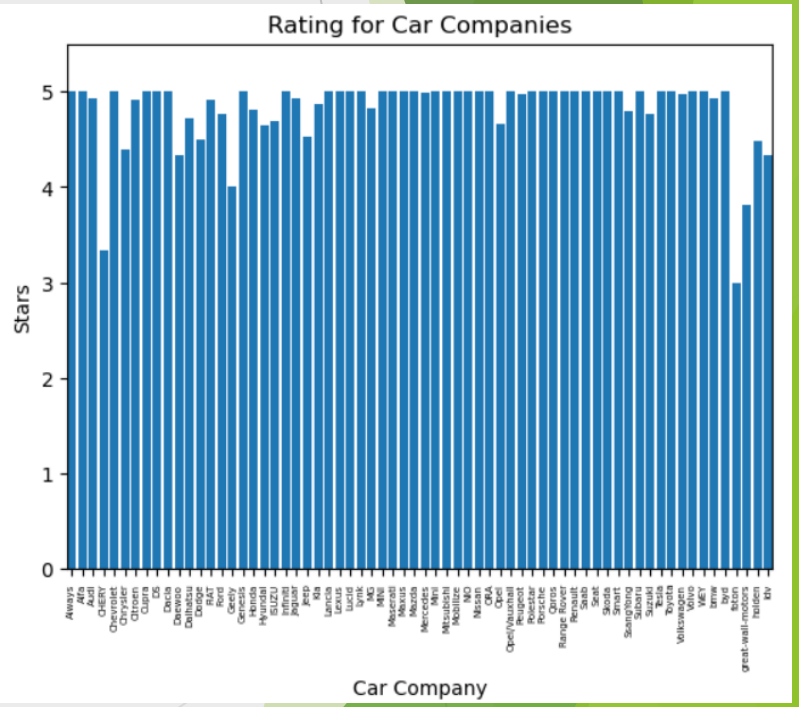
	CarCompany	TestedModel	VehicleType	CarClass	YearOfPublication	KerbWeight	Stars	AEBCar2Car	SpeedAssistance	LaneAssistSystem	AdultOccupantR	ChildOccupantR	VulnerableRoadUsersR	SafetyAssistR
0	Aiways	Aiways U5	- 5 door SUV	Small Off-Road 4x4	2019	1750kg	5	1.0	1.0	1.0	73	70	45	55
1	Alfa	Alfa Romeo Tonale	- 5 door SUV	Small Off-Road	2022	1626kg	5	1.0	1.0	1.0	83	85	67	85
2	Alfa	Alfa Romeo Giulietta	- 5 door hatchback	Small Family Car	2017	1355kg	5	0.0	0.0	0.0	72	56	59	25
3	Alfa	Alfa Romeo Stelvio	- 5 door SUV	Large Off-Road	2017	1745kg	5	1.0	1.0	1.0	97	84	71	60
4	Alfa	Alfa Romeo Giulia	- 4 door saloon	Large Family Car	2016	1449kg	5	1.0	1.0	1.0	98	81	69	60
1300	Volvo	Volvo C30	- 3 door hatchback	Small Family Car	2009	1352kg	5	0.0	0.0	0.0	91	78	26	65
1301	Volvo	Volvo V70	- 5 door estate	Large Family Car	2009	1725kg	5	0.0	0.0	0.0	88	84	43	88
1302	Volvo	Volvo XC60	- 5 door SUV	Small Off-Road	2009	1850kg	5	0.0	0.0	0.0	94	79	48	64
1303	WEY	WEY Coffee 02	- 5 door SUV	Small Off-Road	2022	2100kg	5	1.0	1.0	1.0	94	87	73	93
1304	WEY	WEY Coffee 01	- 5 door SUV	Large Off-Road	2022	2365kg	5	1.0	1.0	1.0	91	87	79	94

EDA ויזואליזציה

BoxPlot מסוג Bar - המתאר השוואה בין חברות הרכב בפרמטרים שונים

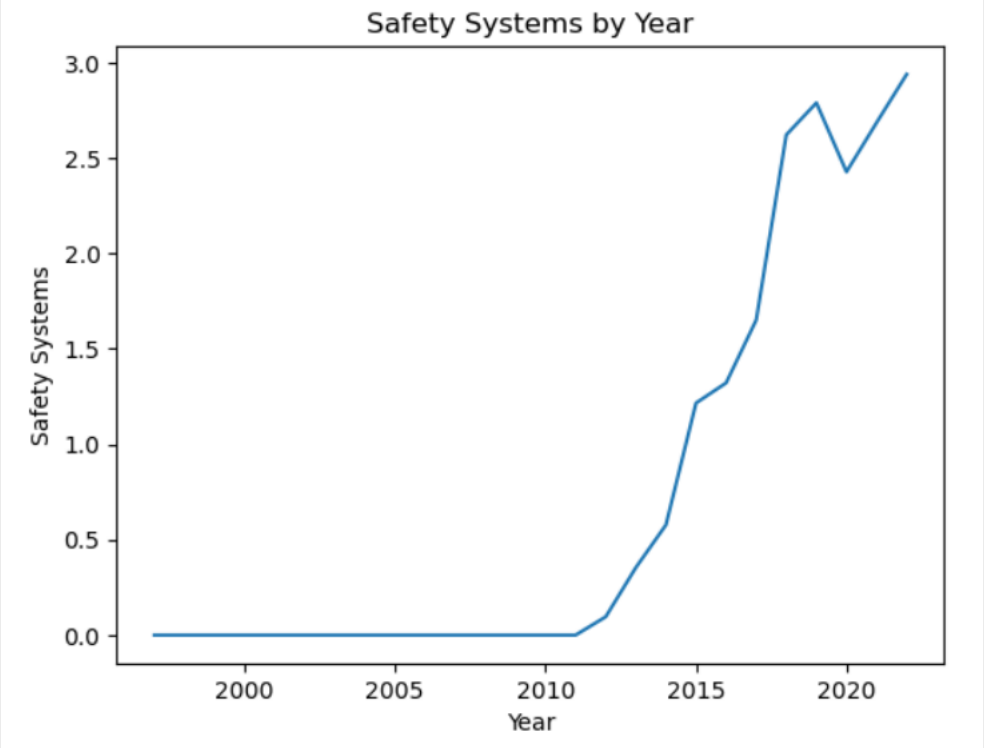






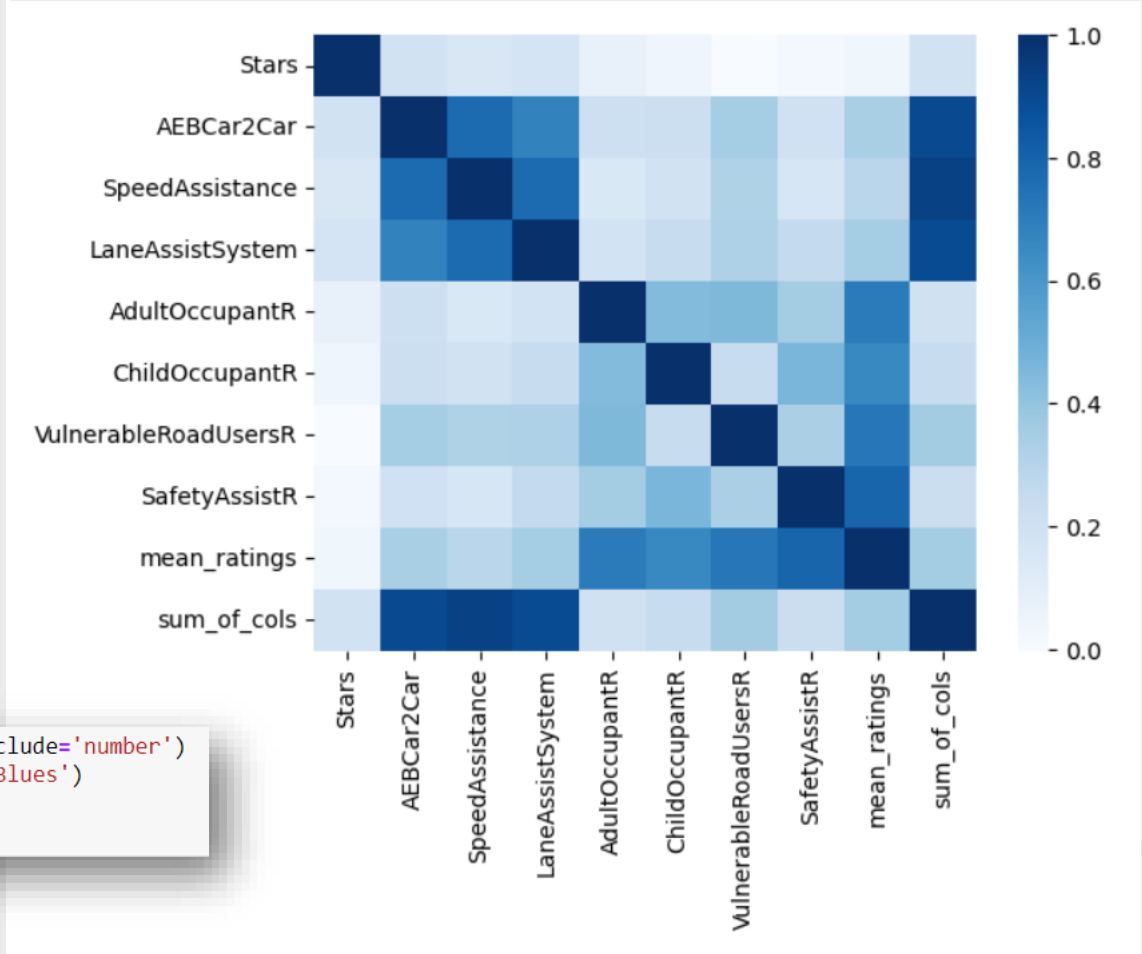
EDA ווויזואליזציה

BoxPlot מסוג Line - המציגה את עליית קיום מערכות הבטיחות לרכב לאורך השנים



EDA ווויזואליזציה

BoxPlot מסוג HeatMap - המציגה קורולציה בין פרמטרים מנורמלים שונים




```
numeric_cols = df_norm.select_dtypes(include='number')
sns.heatmap(numeric_cols.corr(), cmap='Blues')

plt.show()
```

ניתוח נתונים מתקדם

בחירת שיטת עבודה - בהינתן הנתונים השונים שלנו ועמודת המטרה שלנו, החלטנו לבחור ב2 שיטות שיניבו לנו את התוצאה המדויקת ביותר.

ביצענו factorize על עמודת המטרה שמקטלגת כל חברה לפי ערך המיוחד לה, על מנת שיהיה לנו נוח לעבוד עם ערכים נומריים בלבד.

```
0                Aiways
1                  Alfa
2                  Audi
3                 CHERY
4             Chevrolet
            ...
61                   byd
62                 foton
63    great-wall-motors
64               holden
65                   ldv
Name: 0, Length: 66, dtype: object
```

```
codes, uniques = pd.factorize(All_data_copy['CarCompany'])
All_data_copy['CarCompanyCode'] = codes
```

ניתוח נתונים מתקדם

קלסיפיקציה

DecisionTreeClassifier

Accuracy score: 0.44

RandomForestClassifier

Accuracy score: 0.52

יישום והערכת ביצועים

-ניסינו לבדוק אחוזי דיוק עבור השיטה של RandomForestClassifier וראינו שקיבלנו אחוזי דיוק של 0.52 שלא מספקים לנו תוצאה יעילה.

-לאחר מכן עברנו לשיטה של DecisionTreeClassifier על מנת לבדוק האם נוכל לשפר את אחוזי הדיוק, וקיבלנו תוצאה של 0.44 והבנו ששיטה זו לא עוזרת לשיפור.

-הבנו ששיטות המבוססות על רגרסיה לא נותנות לנו פיתרון יעיל עבור השאלת מחקר שלנו, אמנם לא קיבלנו תוצאה גבוהה כפי שציפינו בתחילת הפרויקט אך מספקת.

```
predicted lables:
 [12 10 10 59  9 40 10 59 40 59 62 63 63 53 18 17 18  9  8 27 18  6 39 19
 17 10 63 64  8 60 18  3  8 44 10 63 24 10  3 18 18 10 62  2 19 59  9  2
 18  3 10 18  3 52  3 19 63 40 62 18  3 10 63  9 10 63 10 55 18 19 18  2
 17  9 10 61 10  8 19 28 64 27 53 38 48 19 53 17 24 37 10  1  2 17 18  6
 10 56 24 24  8 44  8 10  2 17 24 17 48  2  2 55 38  8  3 17 24 63 56  9
  8  2 24  2 32  2 23 10 46  8 10  9 10 52 61 28 37  8 63 18 19 58 62 18
 62 58 24 24 29  2 55 55 48 10 10 27 28 37 22 53 18 10 17 19  3  8 62 56
 17  8 10 18 18 60  1  2 19 62 37 19 10  8  6  1 14 46  2 27  6  8 17 62
 62  2 19  3 63 10  2  8 27  3 46  8 23 48 40 52 46  9 24 32  8  3 28 24
 18 19 48 46 19  1 46 63 64  9 10 17 61 24  8 18  8  8 18 38  8 24 10 10
  9 10 10  3 56 27 61  2  8 62 52 27 59 27 10 18 10 27 59 19  2  8 31  3
 63 10  2  9 18 56 18 38 19 63 32  9 48 19  6 18 19  8 52 17  9 10  2 17
 22 24 26 19 24  8 19 18 61 10  9 63  6 62 63 29  8 46  3 48 48 18 32  2
 48 13 59  2 12  9  8 19  1  8 25 38  8 63  3  2  6 18 38 55  9 27 63  3
  8 18  9 56 24 16 62  2  9 19 63 46 18  3  3 10 18  2 10 46 18  8 18  3
 24 38 38 38 62 18 38 10 18  7  8  9  2  3 32 14 29 37 53 38 53  2 17 27
 52 29 38  9  8 62 46 48]
```

```
actual lables:
 [14 62 59 32 29 53 10 61 17 65 32  9 22 53 18 15 18 46  8  5 18  6 39 10
  9 10 41 64  8 48 48  3  9  9 10 63 41 10  3 18 18  8 60  2 19 59 63  2
 63  3 59 18  3 63  3  9 63 53 29 45 56 17 63  9 19 63 19 50 18 37 18 10
 17  8 17 64 10 59 22 28 64 37 44 38 19 10 53 17 24 31 44  1 37 17 18  6
 44 56 24 24  3 52 53 10  2 17 39 17 48  2  2 55 56 25 60 55 24 63 11 12
  8 18 24 10 32  2 23 23 46  8 10 62 10 52 61 62 37 56 32  9 53 58 62 10
 62 19 24 37 29  9 11 55 48 59 10 27  7 37  9 53  9 10 17 44  3 57  3 19
 17  8 48 18 18 37  1 56 19 62 41 44 10  8 53 17 14 46  2 27 64  8 29 62
 62  2 19  3 19 10  2 37 52  3 59  8 23 48 44 52 46 18 24 26  8 62 38 24
 18 48 48 64 19  1 17 63 10 62 46 30 61 24  9 63  8  8  6 38 62 24 10 10
  9 10  2  3 56 27 61  2  9 63  8  1 38 27 62 29 10 27 59 56 10  8 19 16
 23 32  2  9 18 55 18 38 23 62  6  9  9 48 18 18 44  8 52 58 59 10 10 17
 37 24 26 50 24 39 19 18 64 10  9 63  6 62 18 29  8 19  3 48 48 19 63  2
 29 10 62  2 63 29  8 58  1  8 37 37  8 46  3  2  6 18 38 11  9 63 63  3
  7 18 46 56 19 63 62  9  9 19  9 17 18 22  8 10 18 56 37 53 18 41 18  3
 24 38 38 37 59 10 37 44 18  7 19 16  2  3 32  9  8  5 53 38 61 18 17 41
 48 60 38  9  8 23 28 41]
0.5178571428571429
```

סיכום ומסקנות

-מטרת המחקר בפרוייקט הייתה לבדוק האם ניתן לחזות את חברת הרכב שתיצור את הרכב הבטיחותי ביותר על פי מאפיינים של רכיבי בטיחות שונים.

-ניתחנו את הבעיה במספר כיוונים וניסינו ליישם שיטות ומודלים שונים על מנת להגיע לתוצאות חיזוי הכי מדויקות.

-המסקנה שקיבלנו מתוצאות הפרויקט היא שלמרות שהשתמשנו במספר שיטות לחיזוי החברה הבטיחותית, ועם הנתונים שחילצנו מאתרים שונים, לא קיבלנו תוצאה גבוהה מספיק ויעילה עבור דרישות הבעיה.

-זה היה פרוייקט מאתגר בו למדנו המון דברים בלמידה עצמית החל משלב ההרכשה ב selenium , איסוף וסינון המידע ועד השלב האחרון של למידת המכונה.