בס"ד

# למידת מכונה – פרוייקט גמר

מגישים:
בתאל ירושלמי 316320282
אלחי אגסי 205444748

## **נושא הפרוייקט: חיזוי מחירי יהלומים לפי ערכי Z,Y,X.**

**תקציר:**
במסגרת פרויקט זה, אנו מבקשים לחזות את מחירי היהלומים בהתבסס על ממדיהם: אורך (X), רוחב (Y), וגובה (Z). המחקר מתמקד בפיתוח מודל למידת מכונה המסוגל להעריך את המחיר בהתבסס על התכונות הפיזיות הללו.

**השאלות המנחות בפרוייקט זה הן:**

1. מה הקשר בין משקל היהלום (בקראטים) למחירו?
2. איך תכונות פיזיות נוספות של היהלום, כגון צבעו ובהירותו, משפיעות על מחירו?
3. האם ניתן לנבא את מחירו של היהלום בהתבסס על מאפיינים שונים כמו משקל, צבע, בהירות, ועוד?

**DataSet**: www.kaggle.com/datasets/shivam2503/diamonds

**Models:** רגרסיה לינארית (Linear Regression),
עץ החלטה (Decision Tree Regressor),
K-Nearest Neighbors (KNN),
SVM (מכונת וקטורי תמיכה).

כדי לתאר את אופן ביצוע הפרוייקט ושלבי העבודה נסביר ונפרט עבור כל שלב במבנה המחברת:

1. טעינת נתונים.
2. ניקוי נתונים.
3. ניתוח ויזואליזציה.
4. הכנת נתונים.
5. בניית מודל.
6. הערכת מודל.
7. שיפור המודל.
8. תוצאות ומסקנות.

**קישור לגיט:** https://github.com/BatelCohen7/Diamonds.git

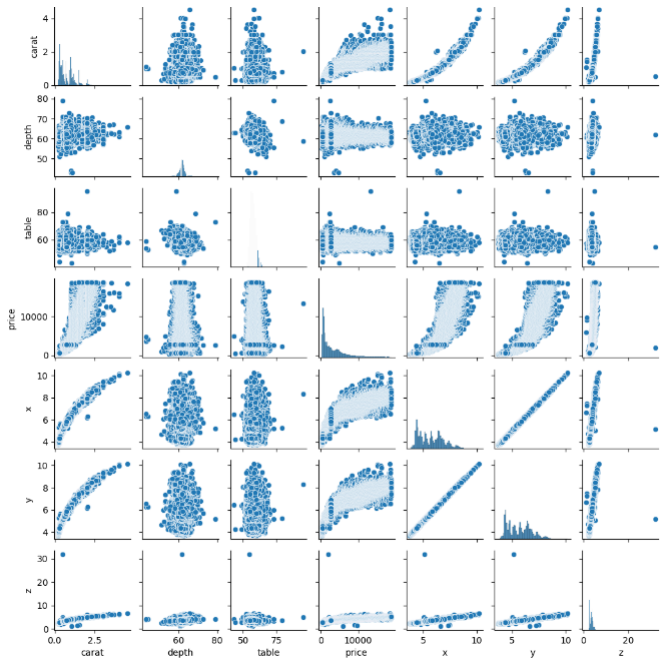
## 1. טעינת נתונים-

- **ייבוא וטעינת מערך הנתונים של היהלומים:**
בשלב טעינת הנתונים, אנו מתמקדים בכמה פעולות מרכזיות כדי להבטיח שנוכל לעבוד עם ה-data set של היהלומים בצורה יעילה. השלב הראשון הוא לבחור data set מתאים שמכיל את המידע על היהלומים שנרצה לנתח. בחרנו לעבוד בפייתון ולכן לפני שנטען את data, נתקין את ספריית Pandas, שמספקת כלים נוחים לניהול ועיבוד נתונים.
בשלב זה, נטען את ה-data set מהקובץ או מקור הנתונים אל תוך סביבת העבודה שלנו. לאחר שהנתונים נטענו, נבצע בדיקות ראשוניות כדי לוודא שהdata נטען כראוי ושאין בעיות בטעינה. לאחר שהשלב הזה הושלם בהצלחה, נעבור לשלב הבא של הפרויקט, שבו נבצע סקירה ראשונית של הdata ונתחיל בתהליך הניתוח והעיבוד שלו.

- **DataSet:**
הdataset של היהלומים שבו השתמשנו בפרויקט זה נלקח מאתר Kaggle , את data ניתן למצוא בכתובת הבאה: Diamonds Dataset. .
הdataset שבחרנו מכיל מידע על יותר מ-50,000 יהלומים, כולל פיצ'רים כמו משקל היהלום בקרטים, איכות החיתוך, צבע, ניקיון, עומק, שולחן (הפאה הרחבה ביותר שנמצאת בראש היהלום), ממדים פיזיים (X, Y, Z) ומחיר. הפיצ'רים הללו נאספו כדי לסייע בניתוח הקשר בין תכונות היהלום למחירו.

- **איסוף נתונים:**
למרות שאין פרטים מדויקים על אופן איסוף הdata באתר Kaggle , בדרך כלל נתונים כאלה נאספים ממאגרי נתונים של חברות יהלומים, אתרי מכירות, או מחקרים גמולוגיים. כל יהלום בדרך כלל נבדק ונמדד על ידי מומחים בתחום הגמולוגיה, שמקפידים על סטנדרטים גבוהים של דיוק ואמינות במדידות. בנוסף, כל תכונה של היהלום כמו החיתוך, הצבע והניקיון מדורגת לפי מערכות דירוג מקובלות בתעשייה, ומשקל היהלום נמדד בקרטים. הממדים הפיזיים (X, Y, Z) נמדדים במילימטרים ומייצגים את האורך, הרוחב והגובה של היהלום בהתאמה.

על מנת להציג באופן ויזואלי את הקשרים בין המשתנים השונים של נתוני יהלומים השתמשנו בתרשים הבא-




תרשים זוגות כזה עזר לנו לזהות מהירות טרנדים, קורלציות, וקשרים פוטנציאליים בין משתנים, וכן לגלות התפלגויות וחריגים. התרשימים הללו משמשים כלי חזק לניתוח ראשוני של data.

## 2. ניקוי הdata -

ראשית, זיהינו שישנם יהלומים בdataset שלנו שהממדים שלהם אינם פיזיית בעליל, כלומר, לפחות אחד מהממדים (אורך, רוחב או גובה) הוא אפס. מכיוון שיהלום בלי אחד מממדיו הוא לא יהלום בפועל, החלטנו להסיר את השורות הללו מהdata שלנו. זה נעשה על ידי סינון הdata ושמירה רק על השורות שבהן כל הממדים הם גדולים מאפס.

בנוסף ביצענו שני שלבים עיקריים כחלק מניקוי הdata כדי להפוך את מערך הנתונים שלנו לנקי יותר ולהכין אותו לניתוח מודל למידת מכונה:

I. **טיפול בהסרת חריגים:**
בחלק זה אנו מבצעים הסרת חריגים עבור כל עמודה נומרית ב-DataFrame שלנו. החריגים מוגדרים כערכים שנמצאים מחוץ לטווח של שלוש סטיות תקן מהממוצע. זו שיטה נפוצה ופשוטה לזיהוי חריגים.

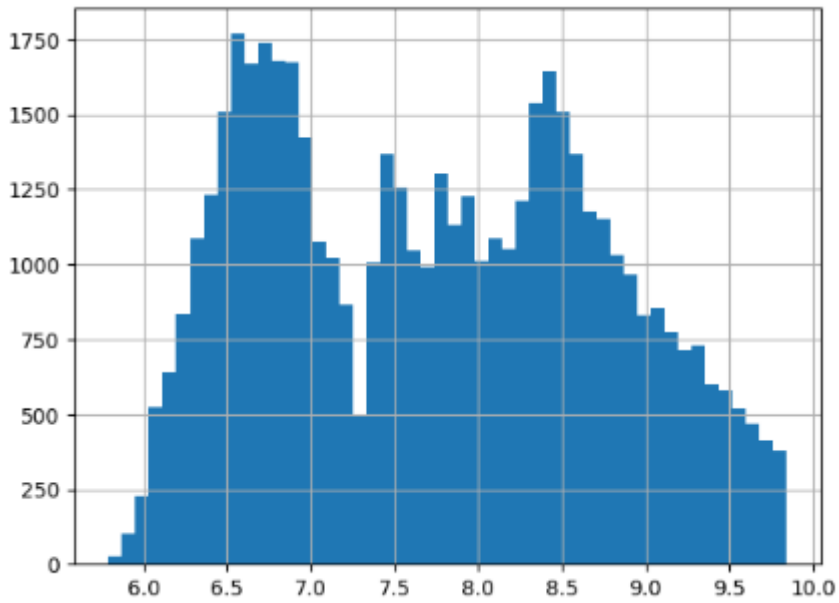
II. **טיפול בנתונים חריגים:**
זיהינו שישנם ערכי z (הגובה של היהלום) שאינם תואמים לצפי שלנו מהערכים של x ו-y (האורך והרוחב). בדרך כלל, ניתן לצפות שיהיה קשר ליניארי בין האורך והגובה של יהלום. לכן, השתמשנו בשיטת Linear regression כדי למצוא את הקשר המשוער בין x ו-z, ולאחר מכן הסרנו את הנתונים שחרגו באופן משמעותי מהקשר הזה. הסטייה המותרת נקבעה באמצעות פרמטר שקראנו לו ε, וכל נתון שהפרשו מהקו המשוער היה גדול מ-ε הוסר מהdata .

באופן כללי, התהליך של ניקוי הdata היה חיוני כדי לוודא שהdata שעליהם אנו מבצעים את הניתוח והמודלים שלנו יהיו נקיים ואמינים ככל האפשר. זה כולל הסרת נתונים שאינם משקפים יהלומים פיזיים והסרת חריגים שעלולים להטעות את המודלים שלנו.

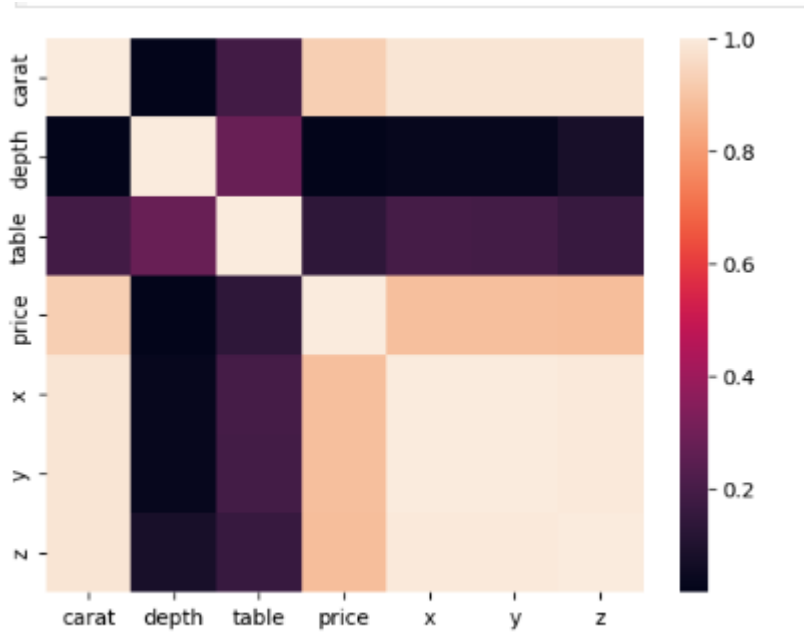
## 3. ניתוח וויזואליזציה: יצירת תרשימים לחקירת הקשר בין ממדי היהלום למחירו.

לאחר שנקינו את הdata והכנו אותם, התחלנו בפעילות חיונית של ניתוח וויזואליזציה כדי להבין טוב יותר את הנתונים שלנו ולזהות תבניות וקשרים בין הפיצ'רים השונים. כמה דוגמאות מרכזיות לוויזואליזציות שביצענו:

1. התפלגות מחירים: ציירנו היסטוגרמה של המחירים כדי להבין את התפלגות מחירי היהלומים. זה עזר לנו לראות את הטווח המחירי ולזהות אם ישנם חריגים או בעיות אחרות ב.data



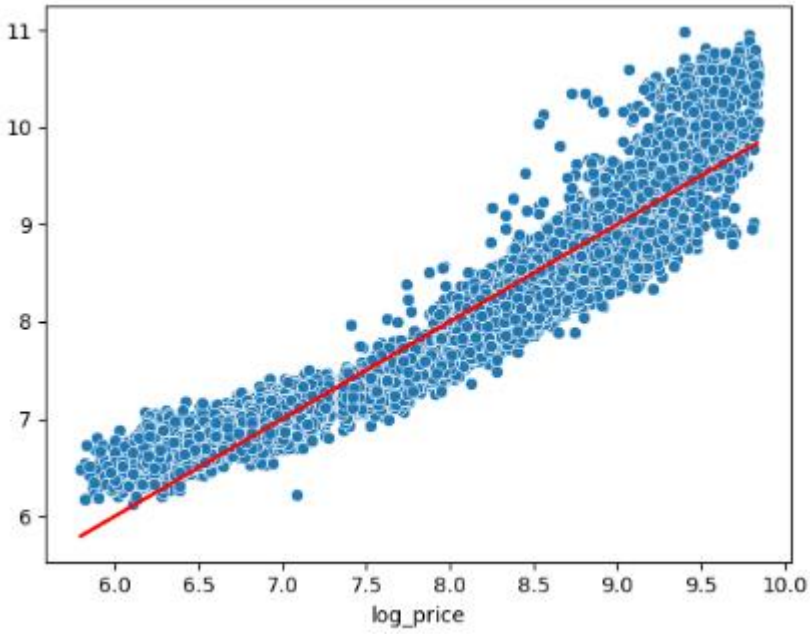

2. קורלציה בין פיצ'רים: באמצעות תרשים חום (heatmap), בחנו את מטריצת הקורלציה בין הפיצ'רים השונים של היהלום. זה עזר לנו לזהות אילו פיצ'רים קשורים בצורה חזקה זה לזה ועשויים להיות רלוונטיים לחיזוי מחיר היהלום.



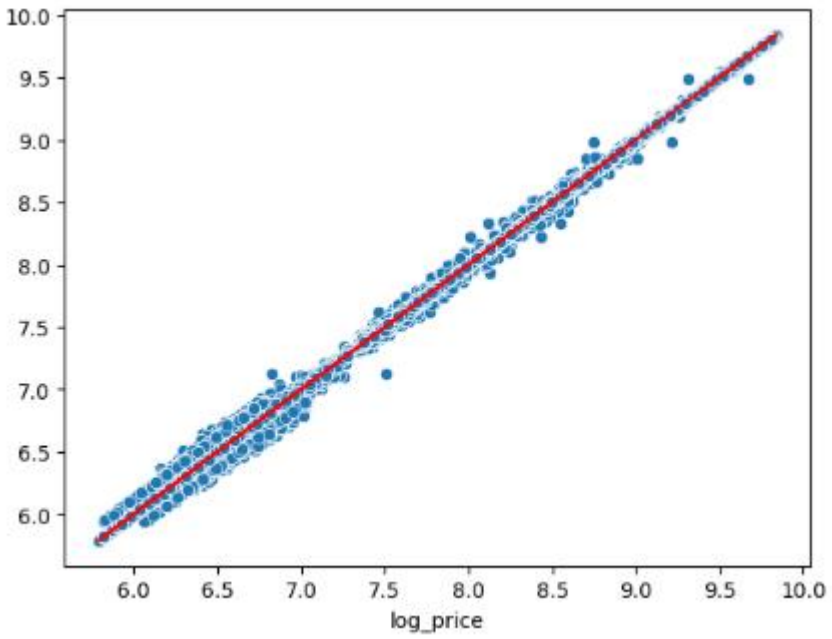

Due to the high correlation we decide to drop all dimensions besides the `carat`.

כאן ניתן לראות כי קיימת קורלזציה גבוה בין X,Y,Z ולבין ה caratולכן ניתן לבטא את X,Y,Z באמצעות ה carat. לכן ניתן להסיר את X,Y,Z מרשימת הפיצ'רים.

3. בדיקת השפעת הפיצ'רים על מחיר: עבור פיצ'רים קטגוריאלים כמו צבע וחיתוך, יצרנו תרשימי קופסה (box plots) כדי לראות איך הם משפיעים על מחיר היהלום. תרשימים אלה מספקים תובנות על ההבדלים במחיר בהתאם לקטגוריות שונות.



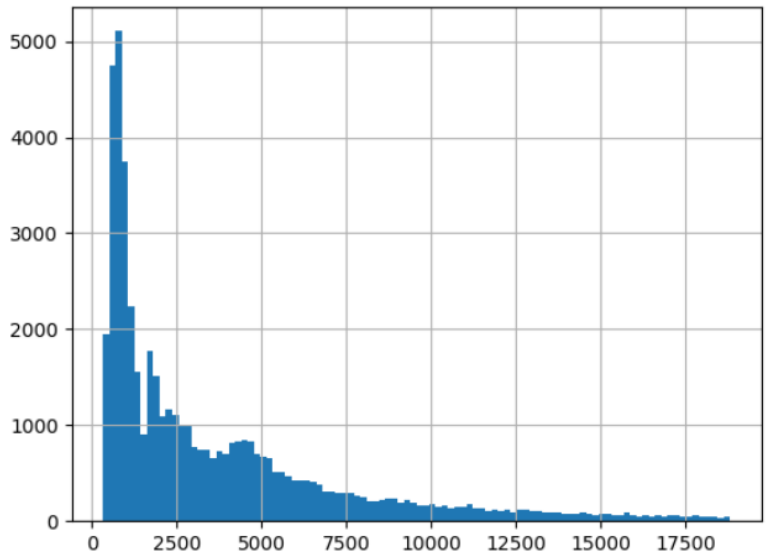

4. ניתוח פיצ'רים עיקריים :(PCA) במקרים בהם ישנם מספר רב של פיצ'רים, שימוש בניתוח פיצ'רים עיקריים (PCA) יכול לעזור להפחית את הממדיות ולזהות את הפיצ'רים העיקריים שמסבירים את השונות הגדולה ביותר ב data.




בסופו של דבר, פעילות ניתוח וויזואליזציה זו הייתה חיונית להבנת הdata שלנו ולהכנתם לבניית מודל מדויק ואפקטיבי יותר לחיזוי מחירי היהלומים.

## 4. הכנת הנתונים: עיבוד הdata והכנתו למודל למידת מכונה.

בחלק הזה ביצענו סקלינג ונורמליזציה לפיצ'רים הנומרים באמצעות טכניקת Min-Max Scaling, כדי להבטיח שכל הפיצ'רים יהיו באותו טווח ערכים וישפיעו באופן שווה על המודל.




בתמונה הבאה אנו רואים שלאחר ביצוע היסטוגרמה למחירים נוצר זנב ארוך שבו יש מעט נתונים עם מחירים מאוד גבוהים מה שיגרום לשיבוש המודל, לכן ניקח את עמודת המחירים ונבצע עליה log וכך נקבל התקבצות של נתונים בהתפלגות הרבה יותר הגיונית, כמובן שאת תוצאת הפרדיקציה של המודל נצטרך להחזיר מהlog הזה כדי לקבל את התוצאה האמיתי.

כמו כן, קידדנו פיצ'רים קטגוריאלים לפורמט נומרי כדי שנוכל להשתמש בהן במודל. לבסוף, פצלנו את הנתונים לtrain וtest כדי להעריך את ביצועי המודל בצורה הוגנת.

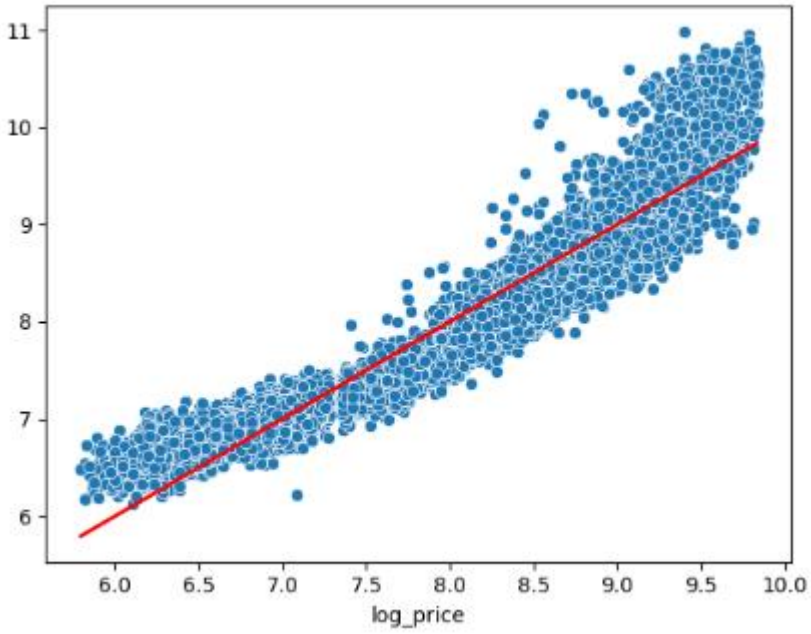
## 5. בניית מודל: פיתוח מודל לחיזוי מחירי היהלומים.

בפרוייקט שלנו, השתמשנו בארבעה מודלים עיקריים לחיזוי מחירי היהלומים: רגרסיה ליניארית, עץ החלטה, K-Nearest Neighbors (KNN), ו-SVM. כל מודל מבוסס על גישה שונה בלמידת מכונה, והוא נבחן על סמך יכולתו לחזות במדויק את מחירי היהלומים בסט הtest. להלן סקירה מפורטת של המודלים שבנינו:

### רגרסיה ליניארית:(Linear Regression)

רגרסיה ליניארית היא אחד המודלים הפשוטים והפופולריים ביותר בתחום למידת המכונה. המודל מנסה למצוא קו ישר (או מישור בממדים גבוהים יותר) שמתאר הכי טוב את הקשר בין הפיצ'רים למשתנה התגובה. במקרה שלנו, המשתנים המסבירים (הפיצ'רים) הם תכונות היהלום כמו משקל, צבע, וחיתוך, והמשתנה התגובה הוא מחיר היהלום.

השתמשנו ברגרסיה ליניארית כדי לקבוע קשר בין הפי'צרים של היהלום למחירו. כאשר בדקנו את המודל על ה data test, התקבלה רמת שגיאה מסוימת שמשקפת את הפער בין המחירים המתועדים למחירים שהמודל חיזה. המודל הליניארי נתן לנו יכולת להבין את הקשר הבסיסי בין הפי'צרים למחיר, אך גם הציג גבולות מסוימים ביכולת לתאר קשרים מורכבים יותר.




התמונה מציגה גרף שבו ניתן לראות את היחס בין ערכי המחירים שניבא המודל לבין המחירים האמיתיים של היהלומים. הנקודות הכחולות מייצגות את הנתונים של המחירים שנחזו על ידי המודל, ציר Y- לעומת המחירים האמיתיים, ציר X - בלוגריתם שלהם. כלומר, עבור כל נקודה במערך הבדיקה, הנקודה מציינת את המחיר האמיתי ואת המחיר שניבא המודל. הקו האדום מייצג את התוצאה האידיאלית שבה ערכי

החיזוי תואמים לחלוטין לערכים האמיתיים - כלומר, כל נקודה שעל הקו הזה משמעה שהמחיר שנחזה על ידי המודל הוא בדיוק המחיר האמיתי.

מתפלגות הנקודות מסביב לקו האדום מראות את השגיאה בין המחיר הניבוי למחיר האמיתי: ככל שנקודה קרובה יותר לקו, התחזית טובה יותר. אם נקודה נמצאת מתחת לקו, המודל ניבא מחיר נמוך מדי, אם היא מעל, המודל ניבא מחיר גבוה מדי. ניתן לראות שהנקודות מרוכזות סביב הקו האדום, והתוצאה שקיבלנו היא 0.31 מה שמראה על כך שהמודל ואלידי .

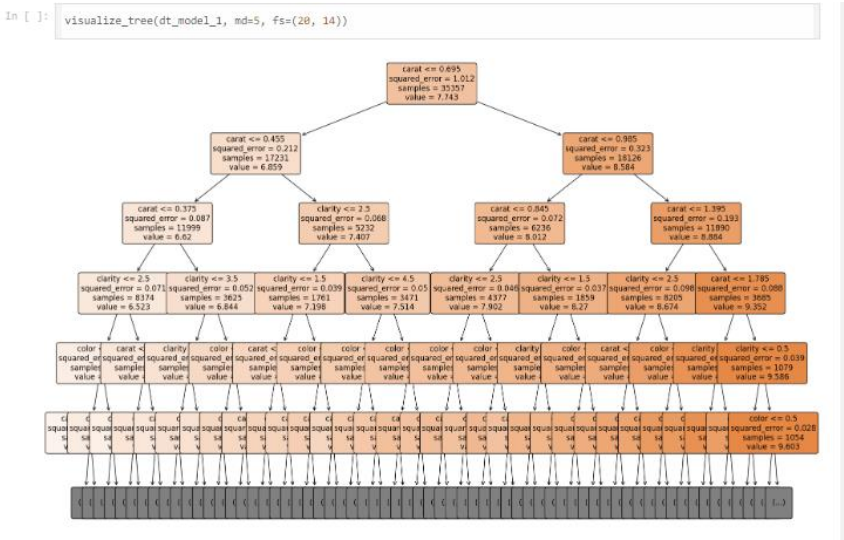
## עץ החלטה:(Decision Tree Regressor)

עץ החלטה הוא מודל שמשתמש בסדרת שאלות ותשובות כדי להגיע להחלטה או פרדיקציה. בהקשר של רגרסיה, עץ החלטה ינסה לקבוע ערך כלשהו עבור משתנה התגובה, בהתבסס על התשובות לשאלות שנשאלות לגבי המשתנים המסבירים.

בפרוייקט שלנו בנינו עץ החלטה כדי לחזות את מחירי היהלומים. העץ מסייע להבין אילו תכונות הן הכי משמעותיות בקביעת מחיר היהלום, וכיצד הן משפיעות על המחיר. עם זאת, עצי החלטה נוטים להיות רגישים לנתוני אימון ויכולים להתאים יתר על המידה (Overfitting) לנתונים הקיימים, מה שמצריך טיפול זהיר כדי להבטיח שהמודל יכול לספק תוצאה זהה עבור נתונים חדשים.

שלבי העבודה:

1. <u>הכנת הנתונים:</u> המודל מתחיל עם data הכוללים פיצ'רים כמו משקל היהלום ,(carat) סוג החיתוך ,(cut) צבע ,(color) ניקיון ,(clarity) עומק ,(depth)וגודל הטבלה .(table) המחיר הוא המשתנה שאנו מנסים לחזות, והוא מומר לסקאלה לוגריתמית כדי לטפל בפיזור הארוך של ערכי המחיר. הנתונים מתחלקים לסט train וסט test כדי לאמוד את יכולת החיזוי של המודל.
2. <u>התאמת המודל:</u> מודל עץ החלטה מותאם לסט הtrain. המודל עובד על ידי יצירת עץ שבו כל קודקוד מייצג פיצול בהתבסס על ערכי הפיצ'ר, וכל ענף מייצג תוצאת חיזוי למחיר היהלום.
במודל הראשוני שיצרנו ההיפר-פרמטרים המוגדרים כברירת מחדל הביאו לעץ חלטה גדול מדי, מה שיצר סיכון ל- Overfitting. הפיצולים בעץ נעשו בהתבסס על ערכים שונים של הפיצ'רים המהווים תכונות של היהלומים.

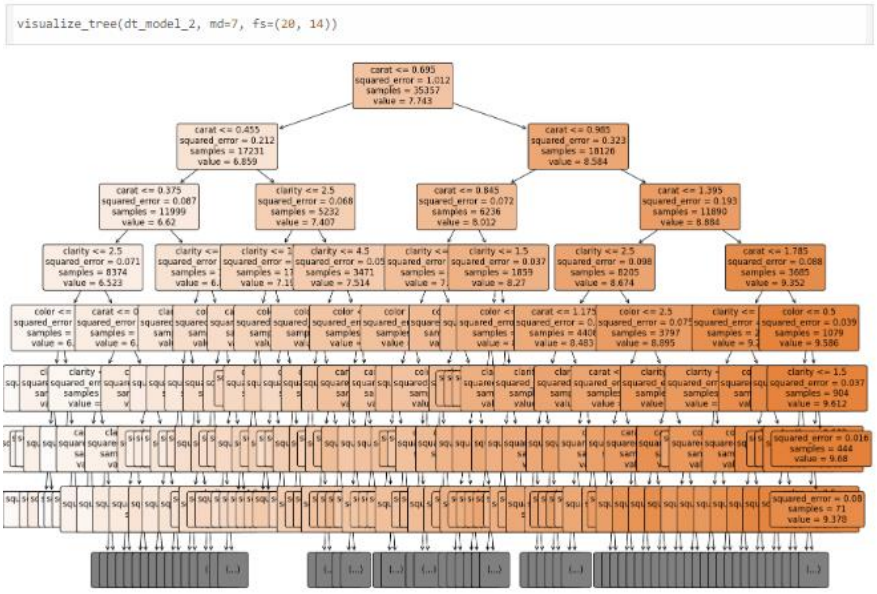



התאמת המודל נמדדה באמצעות חיזוי המחירים לנתוני האימון והשוואתם למחירים האמיתיים. התוצאה הייתה טובה מאוד בסט האימון, אך התברר כי המודל ב Overfitting כאשר הוא נבדק בסט הבדיקה.

3. <u>טיפול ב-Overfitting :</u> ביצענו כמה גישות שונות כדי להפחית את Overfitting, הגבלת גודל העץ והגבלת מספר העלים בעץ.

במודל השני והמתוקן נלקחו צעדים כדי להפחית את הOverfitting על ידי הגדרת מספר מקסימלי של עלים ומספר מינימלי של דוגמאות לכל עלה, מה שהביא למודל יותר כללי עם יכולת כללה טובה יותר.

המודל השני מאפשר להעריך את חשיבות הפיצ'רים השונים על ידי השפעתן על השגיאה הממוצעת בחיזויים.




המודל תרם לחיזוי מחירי יהלומים על ידי למידת הקשר בין הפיצ'רים השונים של יהלומים למחיריהם ומתן חיזוי מבוסס על נתונים קיימים, תוך התמודדות עם אתגרי Overfitting ובחירת פיצ'רים רלוונטים.

## K-Nearest Neighbors (KNN) :

KNN הוא מודל למידת מכונה שמבוסס על הרעיון ש"פריטים דומים נמצאים קרוב זה לזה". בהקשר של רגרסיה KNN ,יחפש את K היהלומים הקרובים ביותר (לפי מדד מרחק מסוים) ליהלום שעליו אנו רוצים לערוך תחזית, ויחשב את המחיר הממוצע שלהם כתחזית למחיר היהלום הנתון.

במחברת שלנו, השתמשנו ב KNN-כדי לחזות את מחירי היהלומים, בנינו מודל KNN ראשון בו ביצענו חיזויים על סט האימון, הערכת הביצועים של המודל על ידי השוואת התוצאות לערכים האמיתיים וחישוב שגיאת ה-RMSE, עשינו זאת תוך שימוש ב10 שכנים קרובים שונים כדי לראות כיצד זה משפיע על דיוק התחזיות.
בנוסף בנינו מודל שני עם נתונים מנורמלים באמצעות MaxAbsScaler לשיפור תוצאות החיזוי, ראינו כי עץ החלטה עם היפר-פרמטרים מותאמים (min_samples_leaf, max_leaf_nodes) נבנה כדי למנוע Overfitting ולהבטיח יכולת כללה טובה יותר.
מתוך הנתונים האלה, ניתן לראות שהמודל השני היה יעיל יותר בחיזוי מחירי היהלומים מאשר המודל הראשון. זה מראה שנרמול הנתונים עזר לשפר את ביצועי המודל. מצאנו ש KNN-יכול להיות יעיל במקרים מסוימים, אך גם הוא דורש עיבוד נתונים זהיר כדי להבטיח שהמרחקים בין הנקודות מחושבים בצורה משמעותית.

## SVM (Support Vector Machine):

SVM הוא מודל למידת מכונה המתמקד במציאת המרחק הגדול ביותר (המרווח) בין הdata של שתי הקבוצות (במקרה של רגרסיה, המרחק מהווקטור התומך). במקרה זה, אנו משתמשים ב SVM למשימת ריגרסיה (SVR - Support Vector - Regression)כדי לחזות את מחירי היהלומים.

שלבי עבודת המודל:
הכנת הנתונים: הנתונים מתחלקים למאפיינים (X) ולתגובה ,(y) כאשר 'log_price' הוא הפיצ'ר המוסבר (מחיר היהלום בלוגריתם) ושאר העמודות משמשות כמאפיינים.
חלוקה לסטים: הנתונים מחולקים לסט train וסט test ביחס של 70:30.
אתחול המודל: מודל SVR מאותחל עם גרעין ,RBF (Radial Basis Function) ערך Cשל 1.0 שמגדיר את עוצמת העונש על שגיאות, וערך epsilon של 0.1 שמגדיר את הרווח שמסביב לקו הרגרסיה.
המודל מתאמן על סט האימון ומבצע חיזויים על סט האימון והבדיקה ומחשבים את שגיאת ה (Root Mean Square Error) RMSE-כדי להעריך את ביצועי המודל.

המודל מנסה לחזות את מחירי היהלומים בהתבסס על פיצ'רים שונים כמו קראט, חיתוך, צבע ועוד. באמצעות הפרדה מרחבית של הdata, ה SVR-מנסה למצוא את הפונקציה הטובה ביותר שמתארת את הקשר בין הפיצ'רים למחיר היהלום.
האמינות של המודל תלויה במספר גורמים כמו איכות הdata, הבחירה הנכונה של פרמטרים (כמו C ו-epsilon), והתאמת המודל למבנה הנתונים. במקרה זה, ערכי ה-RMSE גבוהים יחסית (0.4507 עבור סט הtrain ו-0.4461 עבור סט הtest), מה שמעיד על כך שיכולת החיזוי של המודל אינה מדויקת מאוד.

6. **הערכת המודל: בחינת ביצועי המודל באמצעות נתוני בדיקה.**
לאחר שבנינו את המודלים, עברנו לשלב ההערכה כדי לבדוק את יכולת החיזוי שלהם. השתמשנו במדדים סטטיסטיים כמו שגיאת הריבוע הממוצעת (RMSE) ומדד R-מרובע (R-squared) כדי להעריך את דיוק החיזויים של כל מודל. מודלים עם ערכי RMSE נמוכים וערכי R-מרובע גבוהים נחשבו למודלים עם ביצועים טובים יותר. בנוסף, ביצענו תהליך של ולידציה חוזרת (Cross-validation) כדי לוודא את אמינות התוצאות ולמנוע Overfitting.

7. **שיפור המודל: טיונינג פרמטרים וניסיון לשפר את ביצועי המודל.**
בהתבסס על הערכת המודלים, נקטנו בצעדים לשיפור המודלים שהציגו ביצועים פחות טובים. שיפורים אלו כללו טיונינג של היפר-פרמטרים, כלומר, התאמת פרמטרים כמו עומק העץ במודל עץ החלטה או מספר השכנים ב-KNN. כמו כן, שימשנו טכניקות כמו פיצ'ר אינג'ינירינג כדי ליצור מאפיינים חדשים שעשויים לשפר את הביצועים של המודל. לבסוף, בדקנו את האפשרות של שימוש במודלים אנסמבל כדי לשלב את התחזיות ממספר מודלים שונים ולקבל תוצאה מדויקת יותר.

8. **תוצאות ומסקנות:**
בשלב זה, אנו סוקרים את התוצאות שהושגו על ידי המודלים שפיתחנו ומנתחים את השפעתם על חיזוי מחירי היהלומים.

<u>חשיבות התכונות:</u> הניתוח שלנו הראה שפיצ'רים מסוימים היו בעלי השפעה משמעותית על המחיר של היהלום. לדוגמא, משקל היהלום (carat) היה גורם משמעותי ביותר לחיזוי המחיר. פיצ'רים נוספים כמו צבע, חיתוך וניקיון גם היו בעלי השפעה, אך במידה פחותה. הנתונים הללו מאפשרים לנו להבין אילו פיצ'רים חשובים יותר בקביעת מחיר היהלום ולהתמקד בהם במודלים עתידיים.

<u>ניתוח שגיאות המודל:</u> השגיאות שהתקבלו במודלים נמדדו באמצעות RMSE ו-R-squared. בעוד שהמודל SVM הראה יכולת חיזוי טובה עם ערכי RMSE נמוכים, עץ החלטה הראה נטייה ל-Overfitting, מה שהוביל לשגיאות גבוהות יותר בסט הבדיקה. ניתוח השגיאות מסייע לנו לזהות את המודלים שדורשים שיפור ולמקד את המאמצים לכיוון זה.

<u>הצעות לשיפורים עתידיים:</u>
* ניסיון לטיונינג יותר מדויק של היפר-פרמטרים במודלים, במיוחד בעץ החלטה כדי להפחית את הנטייה ל-Overfitting.
* שילוב של מודלים בטכניקות אנסמבל כדי לנסות לשפר את יכולת החיזוי על ידי הפחתת וריאביליות.
* הוספת פיצ'רים נוספים שעשויים לשפר את החיזוי, כמו היסטוריית מכירות או פרטים על המקור של היהלום.
* בחינה של שימוש בטכניקות למידה עמוקה כדי לקבל מודלים מתקדמים יותר שיכולים לתפוס מורכבויות בנתונים.

לסיכום, הפרויקט סיפק תובנות חשובות לגבי חיזוי מחירי יהלומים והציע דרכים לשיפור המודלים בעתיד. המחקר פתח דלת למחקרים נוספים בתחום זה והדגיש את הפוטנציאל של למידת מכונה בתחום הגמולוגיה.