.1

א.מטרת כריית המידע היא לתמחר את הרכבים לפי שאר התכונות.

ההנחות עליהן התבססה העבודה הן:

i. עדיפות למחיר גבוה ביחס למחיר נמוך יותר, מאחר ובמצב בו לא יהיה ביקוש לרכב מסוים , ניתן להוריד את מחירו. אפשרות זו עדיפה על תמחור רכב במחיר נמוך ובכך לוותר על אופציה להרוויח.

ii. ערכי הנתונים אמינים ,מאחר ואין אינדיקציה אחרת הסותרת את ההנחה.

.

ב.הנתונים:

תכונה	סוג	תחום ערכים	נתונים חסרים והערות
symboling	מספר בדיד אינטרוולי	טווח בפועל [-2,+3]	אין חוסרים, 2- הרכב בטוח יותר ממחירו ו3+ זה מסוכן.
normalized-losses	מספר רציף אינטרוולי	טווח בפועל [56,256]	41 ערכים חסרים המהווים 20% מכלל הנתונים. כמו כן רוב הערכים הם מתחת לחציון. מאחר ומגמת הירידה היא הדרגתית יחסית ניתן להניח שהערכים תקינים
make	קטגורי נומינלי	סוג הרכב[1]	אין חסרים. יש 6 סוגי רכבים שיש להם פחות מ-5 רשומות. כיוון שבעולם הרכב ,ליצרן יש משמעות גדולה לערך מחיר הרכב, נשאיר זאת כפי שזה.
fuel-type	קטגורי נומינלי	Gas/diesel	אין חסרים, 10% מהנתונים הם gas
aspiration	קטגורי נומינלי	Std/turbo	אין חסרים, בערך 18% הם tubo
num-of-doors	קטגורי אורדינלי	Two/four	2 רשומות חסרים
body-style	קטגורי נומינלי	Sedan, hatchback, wagon, hardtop, convertible	אין חסרים, ל2 ערכים יש פחות 10 רשומות.
drive-wheels	קטגורי נומינלי	Rwd,fwd,4wd	אין חסרים, יש פחות מ10 רשומות ל4wd
engine-location	קטגורי נומינלי	Front,rear	אין חסרים, יש רק 3 רשומות לrear
wheel-base	מספר רציף אינטרוולי	טווח בפועל [88.6,120.9]	אין חסרים, נראה שיש מגמה יורדת וככל שמתרחקים ממרכז המסה כך המרווחים גדלים, ניתן להסיק שההתפלגות תקנית
length	מספר רציף אינטרוולי	טווח בפועל [141.1,208.1]	אין חסרים, מגמת הנתונים נראית תקנית
width	מספר רציף אינטרוולי	טווח בפועל [60.3,72.3]	אין חסרים, מגמת הנתונים נראית תקנית
height	מספר רציף אינטרוולי	טווח בפועל [47.8,59.8]	אין חסרים, מגמת הנתונים נראית תקנית
curb-weight	מספר רציף אינטרוולי	טווח בפועל [1488,2555.566]	אין חסרים, נראה שיש מגמה יורדת וככל שמתרחקים ממרכז המסה כך המרווחים גדלים, ניתן להסיק שההתפלגות תקנית
engine-type	קטגורי נומינלי	Ohc, l, dohc, ohcf, ohcv, rotor,dohcv	אין חסרים, בערך 72% שייכים לערך ohc וב2 ערכים יש בכל אחד פחות מ5% מהרשומות
num-of-cylinders	קטגורי אורדינלי	Two, three, four, five, six, seve, eight	אין חוסרים, בערך 78% שייכים לערך Two וב4 ערכים יש בכל אחד פחות מ5% מהרשומות
engine-size	מספר רציף אינטרוולי	טווח בפועל [61,326]	אין חוסרים, נראה שיש מגמה יורדת וככל שמתרחקים ממרכז המסה כך המרווחים גדלים, ניתן להסיק שההתפלגות תקנית
fuel-system	קטגורי נומינלי	Mpfi, 1bbl, 2bbl,	אין חסרים,ב4 ערכים יש בכל אחד פחות מ5% מהרשומות

		spdi, idi, mfi, 4bbl,spfi	
bore	מספר רציף אינטרוולי	טווח בפועל [2.54,3.94]	4 רשומות חסרות, יש 2 ערכים שלא עומדים באותו קצב מגמת ירידה כמו שאר הנתונים ונמצאים בתחילת הטווח, לכן נתייחס אליהם כטעויות ונוריד אותם- כעת 6 רשומות חסרות והטווח הוא[2.91,3.94]
stroke	מספר רציף אינטרוולי	טווח בפועל [2.07,4.17]	4 רשומות חסרות, למרות שבערך 80% מהרשומות מתרכזות בחלק התחתון של המספר 3, לפי מגמת העלייה נסיק שהטווח תקין
compression-ratio	מספר רציף אינטרוולי	טווח בפועל [7,23]	אין חסרים.יש רווח בין ערך של11.5 ל21 ובין ערך 21 ל23 יש פחות מ10% מהרשומות,. מגמת הירידה בערכים אלו נראית דומה לשאר הערכים וניתן להסיק שהיא תקינה. כיוון שיש רווח נכיר בין 2 תחומי עליה אלו נזהר לא להשוות רשומות מ2 צדי הרווח עם תכונה זו.
horsepower	מספר רציף אינטרוולי	טווח בפועל [48,288]	2 רשומות חסרות, נראה שיש עלייה הדרגתית עד הטווח של 207 ואחריו קפיצה גדולה של בערך 60 ל2 ערכים. ניתן להסיק שאלו טעויות ,נמחק אותם ונשנה את הטווח – כעת 4 רשומות חסרות והטווח הוא[48,207]
peak-rpm	מספר רציף אינטרוולי	טווח בפועל [4150,6600]	2 רשומות חסרות, נראה שיש עלייה הדרגתית עד הטווח של 6000 ואחריו קפיצה גדולה ל2 ערכים עם ערך 6600. נסיק שאלו טעויות ,נמחק אותם ונשנה את הטווח – כעת 4 רשומות חסרות והטווח הוא[4150,6000]
City-mpg	מספר רציף אינטרוולי	טווח בפועל [13,49]	אין חוסרים. בקצוות יש רשומות עם ערכים שמופיעים מעט פעמים, אך לפי מגמות הנתונים נסיק שטווח תקין
highway-mpg	מספר רציף אינטרוולי	טווח בפועל [16,54]	אין חוסרים. בקצוות יש רשומות עם ערכים שמופיעים מעט פעמים, אך לפי מגמות הנתונים נסיק שטווח תקין
תכונת החיזוי: price	מספר בדיד אינטרוולי	טווח בפועל [5118,45400]	4 נתונים חסרים, אין חוסרים, נראה שיש מגמה יורדת וככל שמתרחקים ממרכז המסה כך המרווחים גדלים.נסיק שההתפלגות תקנית

טבלה(1): התכונות השונות

1.סוגי הרכב: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu jaguar, mazda, mercedes-benz, mercury,mitsubishi, nissan, peugot, plymouth, porsche,renault, saab, subaru, toyota, volkswagen, Volvo

ג.שלבי הKDD(עפ"י הרצאתו הראשונה של פרופ' מרק לסט, שקפים 7-12):

i. <u>בחירת הנתונים:</u>

בשלב זה נצטרך לבחור את הנתונים עליהם נעבוד, הוא הבסיס לכל השלבים הבאים ויכול להשפיע על התוצאות שנקבל. למשל אם אנו לוקחים נתונים שנאספו מהשנה האחרונה או מהעשור האחרון. כמו כן נצטרך להתמקד על תת-קבוצה שתהיה רלוונטית למטרת כריית המידע שלנו. במקרה שלנו, קיבלנו את הנתונים ואלה כל הנתונים, אין חותמת זמן על הנתונים ומדובר על מעט רשומות,205, לכן ההתיחסות היא שכולם רלוונטיים.

ii. <u>ניקוי ועיבוד מקדים:</u>

אחרי שבחרנו לעבוד עם כל הרשומות, נתייחס לאיך לטפל ברשומות עם ערכים חסרים(למלא אותן או לא), ברשומות עם נתונים לא הגיונים,outliers,רעש ורשומות שהמידע שלהן אינו רלוונטי להמשך. נוכל

לנפות חלק מהערכים התקולים הללו ע"י השמטתם\עשיית ממוצע\הכנסת קבוע כלשהו\לפי התפלגות ועוד. כל העיבוד יפורט בסעיף ה'.

iii. <u>הורדת ממדיים וטרנספורמציה של הנתונים:</u>

בהורדת מיימדים, נקבל ייצוג מונמך של הנתונים שהוא קטן בנפח, אך עדיין משמר את היושרה של הנתונים האמיתיים.

למשל נבחן האם להמיר טווח של ערכים רציפים לקבוצות(למשל ב compression-ratio:בקטגוריה אחת טווח בין 7 ל-11.5 ובקטגוריה אחרת טווח של 21 עד 23). בטרנספורמציה, נהפוך או נגבש את הנתונים כך שתהליך כריית המידע יהיה יותר יעיל והדפוסים ימצאו יותר בקלות וקלים להבנה. כל העיבוד יפורט בסעיף ה'.

iv. <u>כריית המידע:</u>

נבחר את המטרה(במקרה זה היא חיזוי) ונבחר אלגוריתם, כאשר יש כמה משפחות של אלגוריתמים(DT, Clustering וכו') ולכל אלגוריתם יש נטייה לגבי כיוון מסוים ומדגיש תכונות אחדות. נחלק את סט הנתונים לנתוני בדיקה ואימון, למשל ע"י $k-fold\ cross-validation$. ראוי להזכיר שכדאי לבדוק את האלגוריתם עם פרמטרים שונים או להריץ מספר אלגוריתמים שונים ,כדי לקבל את התוצאה המהימנה ביותר או הברורה ביותר, למשל כדי להימנע מהתאמת יתר. יתואר בשלב ד.

v. <u>הערכת הדפוס:</u>

נבצע הדמייה של הנתונים ונעריך את הדפוסים שהתקבלו(כמה הם מעניינים, רלוונטים, אפשריים סטטיסטית וכו'), בנוסף נבחן האם כדאי לחזור חזרה לשלב הקודם כדי לבצע ניתוח מחודש של המידע. יתואר בשלב 2.

vi. <u>מסקנות:</u>

בסוף התהליך נקבל ידע, מודל או תחזית על הנתונים הנוכחים ו\או הבאים. במקרה שלנו נקבל מודל על מנת לסווג מחירי רכבים חדשים.

ד. חלופות אפשריות לכריית מידע

i. <u>רגרסיה לינארית מרובת ערכים:</u>

במודל זה נבנית נוסחה לפי מספר משתנים מספריים, שבבסיסה קשר לינארי בין המשתנים לבין חיזוי של התכונה הרצויה.

שיטה זו מתאימה למשתנים נומריים רציפים לינאריים במהותם. הקושי בשיטה זו היא להחליט כמה משקל לתת לכל תכונה, כיצד לטפל בערכים חסרים או במצב שהנתונים לא מתפלגים בצורה אחידה וכיצד לתת ערך למשתנים קטגורים. ננסה במקרה שלנו בו מתקיימים הקשיים הללו להתאים רגרסיה לינארית.

ii. <u>עץ החלטה ID3:</u>

האלגוריתם לבניית עץ החלטה, כאשר תכונת הפיצול נבחרת כל פעם, לפי זו הנותנת את ה $Information\,Gain$ המקסימלי(אנטרופיה מינימלית), ליצירת תת-קבוצה של נתונים תחת כל ערך של התכונה.

אלגוריתם זה נוקט בגישה חמדנית ולכן לא מגיע בהכרח לאופימיזציה גלובלית. בנוסף האלגוריתם יכול ליצור התאמת יתר לנתוני האימון ויש לו קושי בנתונים רציפים כיוון שיש להם הרבה מקומות פיצול אפשריים.

iii. <u>עץ החלטה C4.5:</u>

האלגוריתם הוא שיפור של ID3, הוא עובד באותה צורה עם השיפורים המפורטים: טיפול בערכים רציפים ובבדידים, מבצע גיזום לעץ ובכך מקטין את האפשרות של התאמת יתר ומקטין את הפגיעה בסיווג על ידי חלוקה

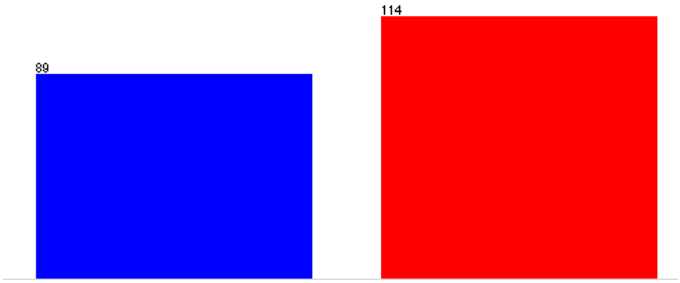
נוספת, בנוסף הוא מצליח להתמודד עם ערכים חסרים ותכונות עם עלויות שונות. נעדיף להשתמש באלגוריתם זה במקרה הנוכחי על פני אלגוריתם ID3.

iv. <u>אלגוריתם IBL מבוסס על K השכנים הכי קרובים</u>

אלגוריתם זה מצריך לבנות מבנה נתונים שיאפשר חיפוש במרחב הn מימדי, של k השכנים הקרובים ביותר גיאומטרית.
בבקשה לחזות את מחיר הרכב, נבדוק האם היא נמצאת במאגר הנתונים ע"י ביצוע חיפוש של k השכנים הקרובים גיאומטרית ונבדוק מי מהם יותר קרוב לרשומה שהגיעה.
הבעיה בשיטה שהיא אינה פועלת כמו שצריך במימדיים גבוהים, ואינה טובה בשילוב של משתנים בדידים ורציפים ואינה תומכת בערכים חסרים, לכן עדיף לא להשתמש בה.

ה.בדיקת ההתפלגויות של הנתונים, זיהוי הרעש והסרתו טופל בסעיף הראשון, כעת נתרכז בטיפול בנתונים החסרים. מאחר ויש רק 205 רשומות של נתונים, עדיף לנסות לתקן רשומות עם ערכים חסרים או עם רעש במקום להשמיטן.

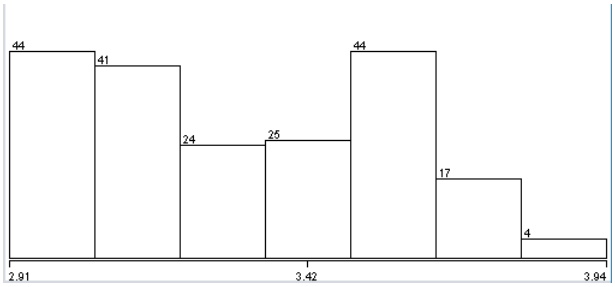
num-of-doors:



גרף(1):num-of-door - בכחול ערך two,באדום ערך four

- חסרים 2 ערכים, מאחר ויש כמות דומה של נתונים בכל אחד מהערכים, נוסיף 2 רשומות אחת עם ערך two והשנייה עם ערך four לכל רשומה חסרה.

Bore:

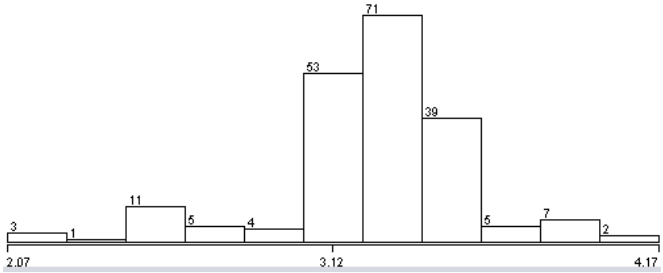


גרף(2): התפלגות הנתונים של ערך bore

ישנם 6 ערכים חסרים, מאחר ומדובר בכמות מעוטה של ערכים חסרים, כ 2.9% מכלל הנתונים, אז ניתן להם ערך של הנטייה המרכזית, שהיא 3.337. כיוון שסטיית התקן גדולה(0.265) יחסית לטווח הנתונים([3.94,2.91]), מצביע על כך שהנתונים מפוזרים יחסית, לפיכך ניתן ערכים באינטרוולים שווים במרחק של סטיית תקן אחת לכל אחת מהרשמות.

כלומר, בטווח של [3.072,3.602] ניתן את הערכים:3.16,3.249,3.337,3.425,3.514,3.602

:stroke

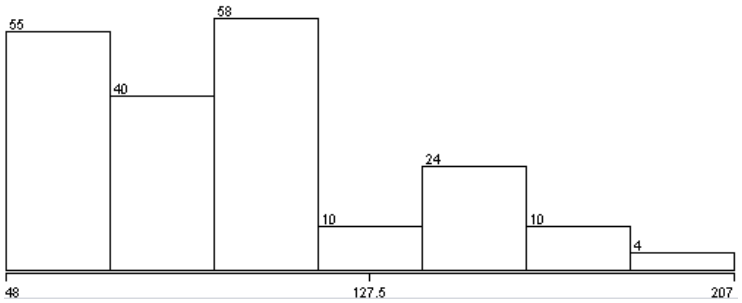


גרף(3): התפלגות הנתונים של ערך stroke

ישנם 4 ערכים חסרים, וכיוון שמדובר בכמות מעוטה של ערכים חסרים, כ 2% מכלל הנתונים, ניתן להם ערך של הנטייה המרכזית, שהיא 3.255.כיוון שסטיית התקן גדולה(0.317) יחסית לטווח הנתונים([4.17,2.07]), מצביע על כך שהנתונים מפוזרים יחסית, לפיכך ניתן ערכים באינטרוולים שווים במרחק של סטיית תקן אחת לכל אחת מהרשמות.

כלומר, בטווח של [3.572,2.938] ניתן את הערכים:3.097,3.255,3.414,3.572

:Horsepower

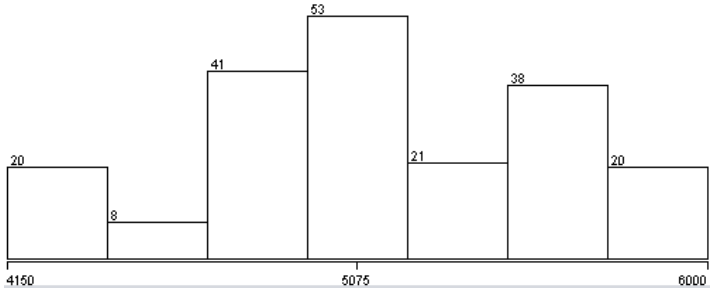


גרף(4): התפלגות הנתונים של ערך Horsepower

ישנם 4 ערכים חסרים, וכיוון שמדובר בכמות מעוטה של ערכים חסרים, כ 2% מכלל הנתונים, ניתן להם ערך של הנטייה המרכזית, שהיא 102.557, כיוון שסטיית התקן גדולה(36.012) יחסית לטווח הנתונים([48 ,207]), מצביע על כך שהנתונים מפוזרים יחסית, לפיכך ניתן ערכים באינטרוולים שווים במרחק של סטיית תקן אחת לכל אחת מהרשמות.

כלומר, בטווח של [138.569,66.545] ניתן את הערכים:84.551,102.557,120.563,138.569

peak-rpm:

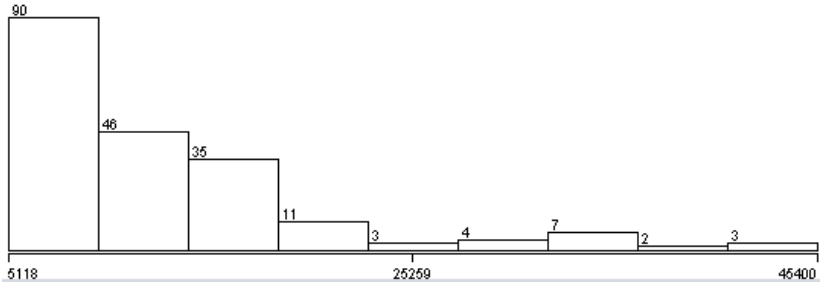


גרף(5): התפלגות הנתונים של ערך peak-rpm

ישנם 4 ערכים חסרים, וכיוון שמדובר בכמות מעוטה של ערכים חסרים, כ 2% מכלל הנתונים, ניתן להם ערך של הנטייה המרכזית, שהיא 5110.697, כיוון שסטיית התקן גדולה(458.364) יחסית לטווח הנתונים([6000, 4150]), שמצביע על כך שהנתונים מפוזרים יחסית, לפיכך ניתן ערכים באינטרוולים שווים במרחק של סטיית תקן אחת לכל אחת מהרשמות.

כלומר, בטווח של [5569.061,4652.333] ניתן את הערכים:4881.515,5110.697,5339.879,5569.061

price:



גרף(6): התפלגות הנתונים של ערך price

ישנם 4 ערכים חסרים, וכיוון שמדובר בכמות מעוטה של ערכים חסרים, כ 2% מכלל הנתונים, ניתן להם ערך של הנטייה המרכזית, שהיא13207.129, כיוון שסטיית התקן גדולה(7947.066) יחסית לטווח הנתונים([45400, 5118]), שמצביע על כך שהנתונים מפוזרים יחסית, לפיכך ניתן ערכים באינטרוולים שווים במרחק של סטיית תקן אחת לכל אחת מהרשמות.
כלומר, בטווח של [21154.195,5260.063] ניתן את
הערכים:9233.596,13207.129,17180.662,21154.195

normalized-losses:
בתכונת normalized-losses ישנם הרבה חוסרים(20%), ונרצה למלא אותם במקום להשמיט את הרשומות, כדי לטפל בחוסרים, בתיאור התכונה כפי שתועד במסמך ניתן לראות שהוא מנורמל לפי גודל קלסיפיקציית המכונית(body-style).
מיון של המכוניות לפי body-style , השלמת החוסרים לפי המרחק הכי קצר, מבין השורות עם אותו ערך של body-style שיש להן ערך ב normalized-losses,(כמו בk-nn כאשר k זה מספר השורות בnormalized-losses שיש להם ערך והם מהווים את מרכזי הclusters הראשוניים, כאשר יש איטרציה אחת) המרחק נקבע להיות אוקלידי, כאשר השינויים הושמו לתכונות כמפורט:

תכונה	שינוי
symboling	נרמול לסקלה של [0,1]
make	החישוב בוצע לפי נספח 1
fuel-type	יש 2 ערכים אז לgas ניתן הערך 0 ולdiesel 1
aspiration	יש שני ערכים, לstd ניתן הערך 0 ולturbo הערך 1
num-of-doors	יש שני ערכים, לtwo ניתן הערך 0 ולfour הערך 1
body-style	חוסר ידיעה איך לתת ערך יחסי לכל סגנון-גוף התעלמות מתכונה זו
drive-wheels	ברור של 4wd יש את הערך המקסימלי 1, הנעה אחורית יותר נפוצה לכן תקבל את הערך 0.5 ומעט רכבים משתמשים בהנעה קדמית שהיא פחות יעילה 0
engine-location	יש שני ערכים, לfront ניתן הערך 0 ולrear הערך 1
wheel-base	נרמול לסקלה של [0,1]
length	נרמול לסקלה של [0,1]
width	נרמול לסקלה של [0,1]
height	נרמול לסקלה של [0,1]

curb-weight	נרמול לסקלה של [0,1]
engine-type	חוסר ידיעה איך לתת ערך יחסי לכל סוג מנוע , התעלמות מתכונה זו
num-of-cylinders	מספר הצינילנדרים קיבלנו ערך מספרי לפי שמם ונורמלו
engine-size	נרמול לסקלה של [0,1]
fuel-system	חוסר ידיעה איך לתת ערך יחסי לכל סוג מערכת דלק, התעלמות מתכונה זו
bore	נרמול לסקלה של [0,1]
stroke	נרמול לסקלה של [0,1]
compression-ratio	נרמול לסקלה של [0,1]
horsepower	נרמול לסקלה של [0,1]
peak-rpm	נרמול לסקלה של [0,1]
City-mpg	נרמול לסקלה של [0,1]
highway-mpg	נרמול לסקלה של [0,1]

טבלה(2): טרנספורמציה של הנתונים כדי למצוא את הערכים החסרים של normalized-losses

הטיפול בנתונים בוצע ב"אקסל" של "הכנת הנתונים" ,מילוי החוסרים ב normalized-losses התבצע לפי הקוד בקובץ המטלב "nearest_neighbour".

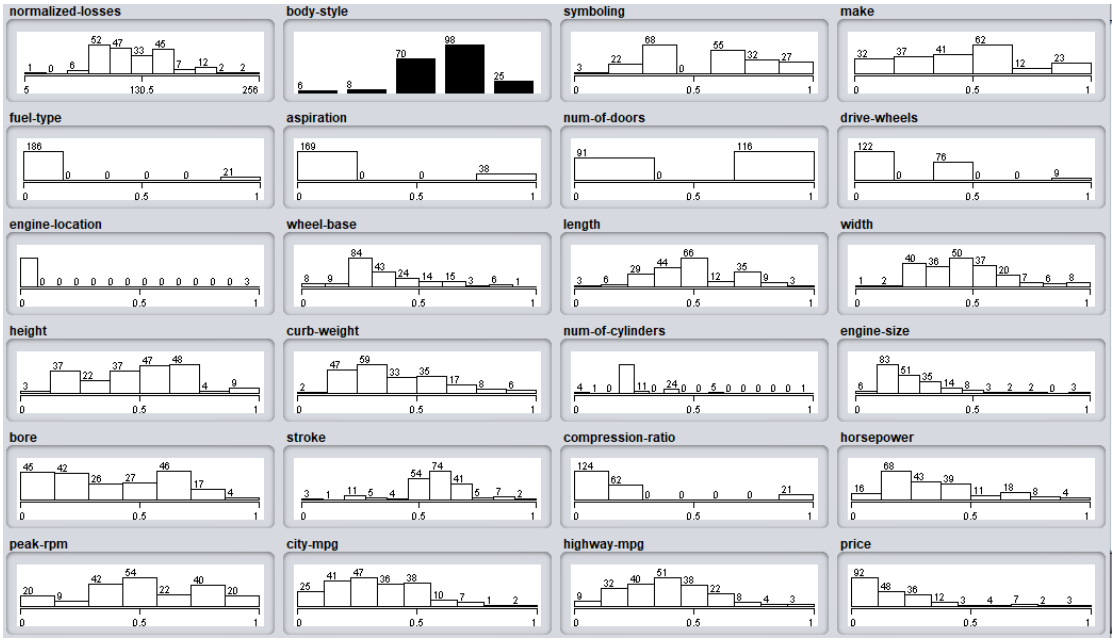
בשיטה של עץ החלטה, הושארו כל התכונות כמו שהן (ללא נרמול ולא מעבר מערך מחרוזתי למספרי), וחלוקה של טווח המחירים ל12 תחומים, ובחירה בתחומים שווי עומק עקב התפלגות לא שווה של הנתונים(מופיע בגיליון "הכנת הנתונים->מילוי סופי-עץ החלטה").

Price range	Number of records	סיווג
-∞ -5000	-	A
5000-6600	20	B
6600-7300	20	C
7300-8000	21	D
8000-9100	21	E
9100-10300	21	F
10300-12500	21	G
12500-15500	21	H
15500-17200	21	I
17200-22500	21	J
22500-45400	20	K
45400-+∞	-	L

טבלה(3): חלוקה לתחומים של המחיר

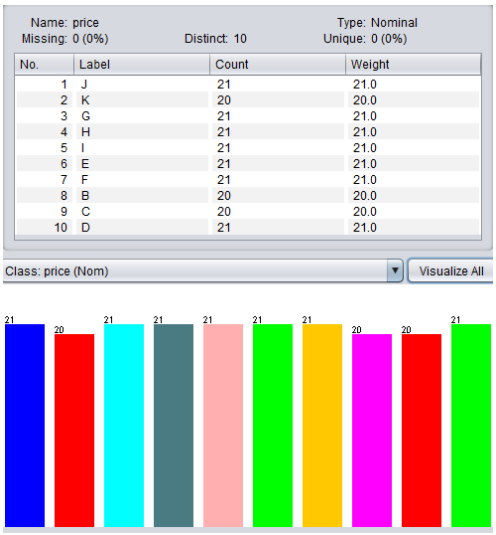
תצוגה גרפית של כל התכונות:

עבור רגרסיה לינארית-



תמונה(1): גרפים של קבוצות המחירים השונים

עבור עץ החלטה-



Name: price
Missing: 0 (0%)
Distinct: 10
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	J	21	21.0
2	K	20	20.0
3	G	21	21.0
4	H	21	21.0
5	I	21	21.0
6	E	21	21.0
7	F	21	21.0
8	B	20	20.0
9	C	20	20.0
10	D	21	21.0

Class: price (Nom)
Visualize All

תמונה(2): גרפים של קבוצות המחירים השונים

תמונה(3): גרפים של כל הערכים לפי סיווג לקבוצות השונת

.2

א.שתי שיטות החיזוי, שנבחרו הן:

i. עץ החלטה C4.5 – שיטה זו מתמודדת עם ערכים בדידים ורציפים, כמו בנתונים שלנו. הוא עדיף מאלגוריתם ID3 ולשם תרגול עצי החלטה(על פני סוג אחר של רגרסיה שנותנת חיזוי רציף).

ii. רגרסיה לינארית מרובת ערכים – בהינתן שקבענו נכונה את הערכים ושהקשר הוא לינארי במהותו, החיזוי שנקבל יהיה מדוייק ,הרבה מן התכונות הן רציפות, עם זאת קיימים מספר קשיים:

1. להחליט כמה משקל לתת לכל תכונה.
2. כיצד לטפל בערכים חסרים
3. מצב שהנתונים לא מתפלגים בצורה אחידה ולינאריים
4. כיצד לתת ערך למשתנים קטגורים

על מנת להתמודד עם אותם קשיים , שימוש בהכנת הנתונים שבגיליון "הכנת הנתונים", עזר למלא את ערכי הnormalized-loss באמצעותו טופלו קשיים:1,2,4.

ההתייחסות עבור קושי מספר 3. כיוון שלכל הנתונים, כפי שנסקרו בחלק 1 סעיף א, יש קו מגמה יורד\עולה ,הונח ,בקירוב גס, שהם מתפלגים בצורה אחידה ולינאריים, בנוסף הנרמול של כל אחד מהתכונות מסייע לנפות מתן משקל לא מאוזן לחלק מהתכונות.

ב.<u>אלגוריתם C4.5:</u>

לאלגוריתם מספר מקרי בסיס:

- כל הדגימות ברשימה שייכות לאותה המחלקה. כאשר מתקיים מצב זה , נוצר עלה עבור עץ ההחלטה המורה לבחור את המחלקה הזו.
- אף אחת מהתכונות לא מספקת שום רווח אינפורמטיבי. במקרה זה, C4.5 נוצר צומת החלטה במעלה העץ באמצעות הערך הצפוי של המחלקה.
- מופע של מחלקה שלא נראתה קודם מגיע. C4.5 יוצר צומת החלטה במעלה העץ באמצעות הערך הצפוי של המחלקה.

שלבי אלגוריתם בניית העץ:

1. בדוק עבור מקרי הבסיס לעיל.
2. עבור כל תכונה x, מצא את ה information gain ratio המנורמל שלמחלוקה על תכונה x.
3. X' יהיה התכונה עם הinformation gain ratio הגבוה ביותר
4. נוצר צומת החלטה שמחלקת על פי X' .
5. חזור שוב על תתי-הרשימות המקבלות מחלוקה על X' והוסף את הקודקודים כ"ילדים" של צומת ההחלטה מסעיף 4.

כמו שנאמר אלגוריתם זה, הוא שיפור של אלגוריתם ID3, השיפורים הם:

- גיזום העץ.
- טיפול בערכים רציפים ע"י יצירת סף וחלוקה לערכים מעל ולערכים מתחת.
- טיפול בערכים חסרים – ערכים חסרים לא ילקחו בחשבון בחישוב הgain.

<u>רגרסיה לינארית מרובת ערכים:</u>

צירוף לינארי של ערכים מספריים רציפים לכדי משוואה לינארית, הערכת מקדמי המשוואה תתבצע על ידי OLS(ordinary least (squares שנקבעים על ידי הערכת סכום הריבועים הקטן ביותר.

ל OLS יש מינימום גלובלי[נספח2] ותיתן תוצאה אחת לפי הערכת סכום הריבועים המינימלי.

הערכת סכום הריבועים המינימלי מתבצעת על ידי מנימיזציית סכום ריבועי השאריות, ניתן לתת לזה ביטוי סגור:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \left(\sum \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum \mathbf{x}_i y_i\right)$$

כאשר $\hat{\beta}$ זה וקטור המקדמים.

<u>מדדים שונים ברגרסיה:</u>

- שימוש באינדיקציה של R^2 ,שהיא יחס השונות במשתנה התלוי (המחיר) שנחזה מהמשתנים הבלתי תלויים(שאר התכונות), על מנת לבדוק את הקירבה שלה המגמה למגמה לינארית.

- שיערוך השגיאה במדידה תתבצע על ידי סטיית תקן, שחישובה הוא:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

כאשר N זה מספר התוצאות,x_i זה הערך של תוצאה ו $\bar{x}$ זה הממוצע

ג.

שתי השיטות התבצעו בעזרת 10-fold-cross-validation,המחלק את סט הנתונים כך ש$\frac{1}{10}$ מסט יהיה בדיקה והשאר הנתונים האימון, זה מתבצע 10 פעמים כאשר נתוני בדיקה כל פעם נבחרים להיות $\frac{1}{10}$ אחרים.

<u>תוצאות הריגרסיה הלינארית:</u>

הרצה של קוד המטלב המצורף("linear_regression"), הקוד מבצע 10-fold-cross-validation, כאשר כל פעם לוקח 21 רשומות להיות ה validation (זה 10% מסט הנתונים) ושאר הרשומות הן הtest.

הקוד פועל כלהלן:

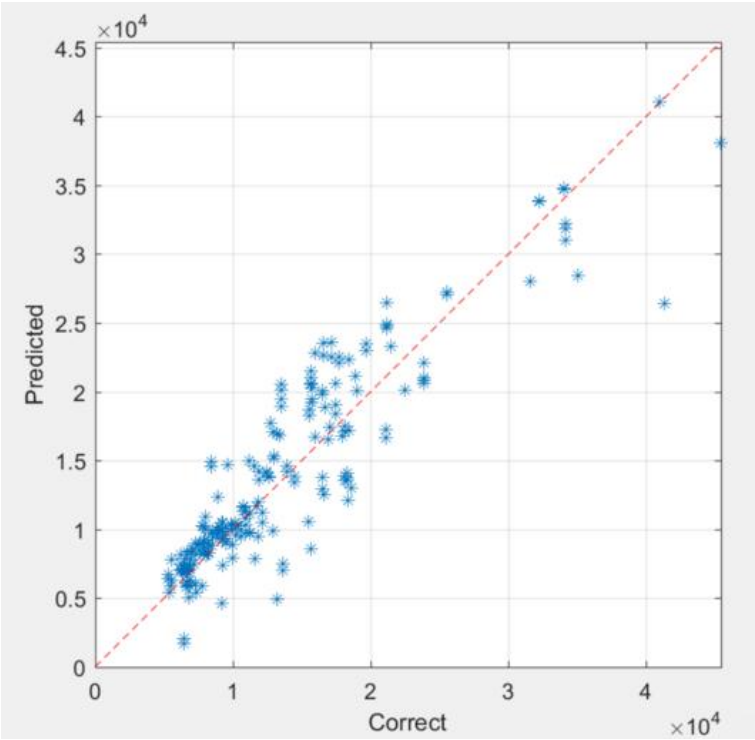
- מריץ רגרסיה לינארית על הרשמות הtest ובעזרתה יוצר ערך חזוי על המחיר של נתוני הvalidation.
- ממיין את נתונים הvalidation לתחומים לפי טבלה מספר 3.
- ממיין את המחירים החזויים מהרגרסיה הלינארית לתחומים לפי טבלה מספר 3.
- בודק את מספר אי התאמות בין המחירים בפועל לבין המחירים החזויים, כשאי התאמה מוגדרת להיות מחיר שנמצא בתחום אחד כשבפועל הוא אמור להיות בתחום אחר.

טבלת סיכום הריצות:

Run number	Mismatches number	R^2	STD	RMSE
1	15	0.671	2860.199	2794.238
2	12	0.697	3110.831	3189.766
3	16	0.436	2810.293	3017.084
4	10	0.849	2955.876	2994.531
5	13	0.848	3085.401	3028.471
6	11	0.914	2513.366	2469.887
7	17	0.870	3032.718	2959.797
8	13	0.418	3200.895	3480.316
9	15	0.935	2671.393	2669.534
10	19	0.664	4522.086	4418.05

*טבלה(4): עבור כל רצה של רגרסיה לינארית ב*10-fold-cross-validation
מספר האי התאמות בה(מתוך 21),r^2,std וה rmse

כמו כן התקבל $R^2 \approx 0.73$ (ממוצע על כולם), מידת ההתאמה של ההתאמה הלינארית למדידות בפועל גבוה(בין אפס לאחד), כדי להמחיש זאת נסתכל על גרף של המחירים בפועל לעומת המחירים החזויים:



גרף(7): ערכי המחירים בפועל לעומת המחירים החזויים

ככל שהנקודות בגרף יותר קרובים לערכים שעל הקו המקווקו(x=y), כך ההתאמה טובה יותר. נראה שיש התאמה טובה.

<u>תוצאות עץ ההחלטה C4.5:</u>

להלן טבלה של הפרמטרים שהוזנו לריצה כולל התוצאות:

Results								parameters			
Number of Leaves	ROC Area	Specificity	Accuracy	Recall	Precision	Wrong allocation	True allocation	Pruning	Confidence Factor	Min number of objects per leaf	Number of run
31	78.50%	92.80%	35.27%	35.30%	34.70%	134	73	Y	0.15	5	1
13	80.30%	92.60%	34.30%	34.30%	33.50%	136	71	Y	0.15	15	2
37	77.70%	92.90%	36.23%	36.20%	36.10%	132	75	Y	0.25	5	3
13	80.30%	92.60%	34.30%	34.30%	33.50%	136	71	Y	0.25	15	4
40	77.60%	93.20%	39.13%	39.10%	39.00%	126	81	Y	0.5	5	5
14	80.10%	92.70%	34.78%	34.80%	34.30%	135	72	Y	0.5	15	6
67	76.60%	93.40%	40.58%	40.60%	40.10%	123	84	Y	0.75	5	7
14	81.10%	92.70%	34.78%	34.80%	34.40%	135	72	Y	0.75	15	8
97	76.20%	93.40%	40.58%	40.60%	40.20%	123	84	N	X	5	9
14	79.80%	92.70%	34.78%	34.80%	34.30%	135	72	N	X	15	10

טבלה(5): הפרמטרים והתוצאות שהתקבלו עבור סדרה של 10 הרצות שונות על נתוני העץ החלטה

הפרמטרים השונים נלקחו מהריצה של עץ ההחלטה, בנוסף לערכים שחושבו בצורה הבאה:

specificity- חושב ע"י $1 - \text{FP Rate}$.

true allocation- חושב ע"י סכימת הערכים שבאלכסון הconfusion matrix.

false allocation- חושב ע"י $\text{total records} - \text{true allocation}$.

ד.+ה.

עבור רגרסיה לינארית:
לפי טבלה 4, המספר הממוצע של אי ההתאמות הוא 14.1, כלומר הדיוק הוא:

$$accuracy = \frac{validation\,records\ per\ run - average\,mismatches\,number}{validation\,records\ per\ run} = \frac{21-14.1}{21} \approx 32.86\%$$

עבור עץ ההחלטה:
נראה שהדיוק נע בין טווח של 33.5% לבין40.2% בהתאם לפרמטרים שהוזנו.

<u>מסקנות מכריית המידע:</u>

1. כריית מידע היא תהליך דינמי. הרבה פעמים כתבתי את שיטת החיזוי ועיבוד הנתונים בצורה שונה עד לבחירה של 2 השיטות שצוינו. כששיטות אחרות הראו על נטייה מסוימת או קושי בחיזוי הצלחתי להבין יותר טוב את התמונה הרב-מיימדית שהסט מייצג וללכת לגישה אחרת.
2. עיבוד הנתונים עבור גישות שונות מצריך עיבוד שונה. עובדה זו חשובה בעיבוד התוצאות שמתקבלות , מאחר וצריך להתייחס לאופן עיבוד הנתונים ההתחלתי בהשוואתן.
3. ניתן לראות שהדיוק של הרגרסיה הלינארית נמוך בהרבה גם מהדיוק הנמוך ביותר בעץ ההחלטה בצורה מובהקת. העובדה שבחיזוי באמצעות רגרסיה לינארית השתמשנו רק ב22 תכונות לעומת 25 תכונות בעץ החלטה(כפי שמצוין בטבלה 2) כנראה תרם לזה במידה מה. בנוסף, ניתן לראות שסטיית התקן הממוצעת של הריצות ברגרסיה(עבור סיווג המחיר), לפי טבלה 4, הייתה בערך- 3076, בעוד שטווחי הסיווג בפועל של המחיר לפי טבלה 3 היו בממוצע- 4040, וזו שגיאה גדולה- שגיאה בסדר גודל של טווח הסיווג. ניתן להסיק מכך, שזו ההשפעה הגדולה על כך שהדיוק היה נמוך ברגרסיה הלינארית וכנראה נבע מההתפלגויות הלא אחיד של הנתונים ולא מבחירת המודל הלינארי בגלל שהתקבל R^2 גבוה.
4. בשיטת עץ ההחלטה, ניתן לראות שעבור מינימום של 5 אובייקטים לעלה נקבל דיוק יותר גבוהה מאשר 15. המסקנה המתבקשת שעבור כמות גדולה יותר של ערכים בעלים נקבל עץ קטן יותר אשר לא יסווג היטב. כנראה שהסיבה נעוצה במיעוט רשומות.
5. הרצה מספר 9, בעץ החלטה, ללא גיזום, התקבל הדיוק הגבוה ביותר. זו חשובה כיוון שהיא מראה שגיזום לא בהכרח משפר את עץ ההחלטה.
6. הרצת עץ החלטה עם פרמטרים שונים נותן תוצאות שונות ושופך אור על אופי הנתונים שיש לנו. לכן כדאי לבחון את התוצאות במתן ערכים שונים לפרמטרים ולא להסתמך על ברירת מחדל.

הצעות לשיפורים:

1. סט הנתונים קטן מדי, מכיל רק 205 רשומות. לקבלת תוצאות טובות יותר כדאי להגדילו.
2. לא נבדקה תלות בין התכונות השונות. במידה והייתה נמצאת קורלציה חזקה בין תכונות, היה ראוי לוותר על אחת מאותן התכונות מאחר והיא לא מוסיפה מידע.
3. ראינו שבחירת פרמטרים משפיעה על התוצאות. קיימים נתונים נוספים שהיה אפשר לבדוק בבחירת העץ וברגרסיה הלינארית. יתכן והיה אפשר לשפר את התוצאה עבור משחק עם פרמטרים נוספים והאפשרויות השונות, אך סביר להניח שהשיפור לא היה משמעותי.
4. ברגרסיה הלינארית, לא כל התפלגויות התכונות היו אחידות. ניתן היה להתאים סוג אחר של התפלגות לכל תכונה, להביא את זה בחשבון ברגרסיה ובכך לשפר את הדיוק .

נספחים

נספח 1

מתן ערך מספרי לדגם המכונית בוצע לפי מתן ערך מספרי עולה ככל שמחיר הדגם בממוצע יותר יקר, הנתונים למחירי הדגם נלקחו מתוך מחירים ממוצעים של מכוניות משומשות באותו דגם בשנה האחרונה בארה"ב, מקור . מספר דגמים לא הופיעו ברשימה, להם בוצע חיפוש של הרכבים המשומשים המוצעים למכירה, באותו מודל מכל השנים וחושב חציון המחירים. מחיר הרכבים הומר לדולר.

הרכבים שלא הופיעו וחציון מחיריהם:

alfa-romero, מקור- יש 467 תוצאות, החציון שנמצא(278 מהתוצאות מתחתיו) היה במחיר 14,852$

renault, מקור- יש 6722 תוצאות, החציון שנמצא(3704 מהתוצאות מתחתיו) היה במחיר 12,376$

peugot, מקור- יש 8731 תוצאות, החציון שנמצא(3964 מהתוצאות מתחתיו) היה במחיר 9,901$

isuzu - לא מצאתי מקור מספיק טוב אז הענקתי לו שווי כשווי של מכונית רגילה יפנית אחרת מהרשימה- Nissan.

Mercury- לא מצאתי טווח מחירים, מתאים כיוון שהמותג כבר לא קיים, אך לפי ויקיפדיה, זה מותג חטיבתי של חברת פורד אז הענקתי לו שווי כשווי של פורד זול לפי מקור.

plymouth- לא מצאתי טווח מחירים, מתאים כיוון שהמותג כבר לא קיים, אך לפי ויקיפדיה, הוא התחרה עם פורד ושברולט כמותג יותר זול אז הענקתי לו שווי כשווי של שברולט זול.

בסה"כ:

מחיר ממוצע משוער בדולרים	Maker
56,000	porsche
30,000	mercedes-benz
29,000	audi
28,000	jaguar
25,000	BMW
19,800	mercury
19,500	chevrolet
19,499	plymouth
18,000	volvo
17,400	toyota
17,300	subaru
16,000	Dodge
15,500	Honda
15,000	nissan
15,000	isuzu
14,852	alfa-romero
13,500	mazda
12,400	volkswagen
12,376	renault
11,700	mitsubishi
9,901	peugot
5,800	saab

טבלה(6): מחיר ממוצע משוער בדולרים לייצרני הרכבים

נמספר את הערכים מ1 עד 21 לmaker הכי יקר וננרמל

מחיר ממוצע משוער בדולרים	maker
1	porsche
20/21	mercedes-benz
19/21	audi
18/21	jaguar
17/21	BMW
16/21	mercury
15/21	chevrolet
14/21	plymouth
13/21	volvo
12/21	toyota
11/21	subaru
10/21	Dodge
9/21	Honda
8/21	nissan
7/21	isuzu
6/21	alfa-romero
5/21	mazda
4/21	volkswagen
3/21	renault
2/21	mitsubishi
1/21	peugot
0	saab

טבלה(7): סידור יחסי מנורמל של ייצרני הרכבים

נספח 2

מקור