פרויקט לימוד מכונה – חלק ב'

תביעת ביטוח רכב

מרצה: פרופ' בועז לרנר

תאריך הגשה: 23/06/2022

מגישים:

אופיר דוד 315695643

משה כהן 316161694

תוכן עניינים

המדד שעל בסיסו בחרנו לבדוק את מידת ההצלחה של המודלים שבנינו הוא מדד הדיוק (accuracy), אשר מתאים למחלקות מאוזנות.

Decision Trees

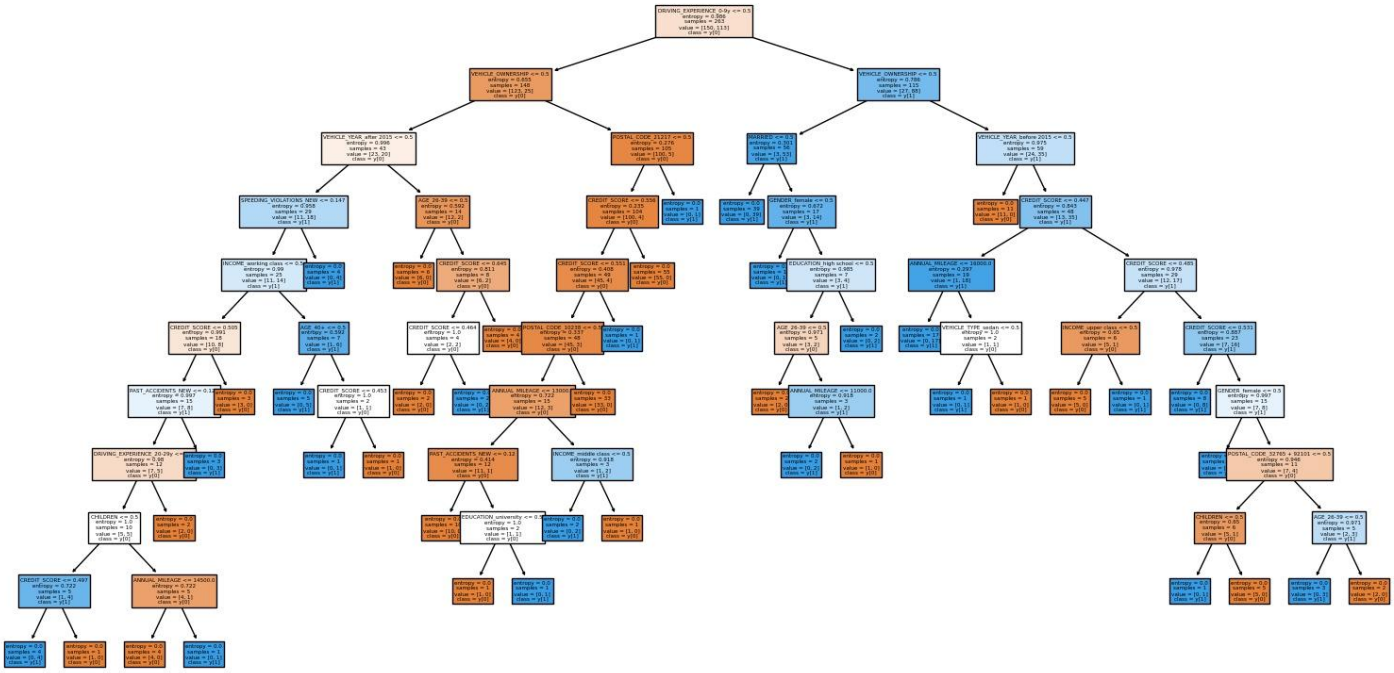
עץ החלטה מלא

נבנה שני עצי החלטה, האחד עבור הנתונים המקוריים כפי שהוצגו בחלק א' והשני לאחר איחוד הקטגוריות שהוצגו בסוף חלק א' (age, postal code). בנוסף, החלטנו שלא להסיר מבסיס הנתונים את המשתנים שבחרנו להסיר בחלק א' של הפרויקט, זאת מתוך הבנה שבמידה ומשתנים אלו אינם חשובים, אלגוריתם עץ ההחלטה לא יכניס אותם לתוך המודל.

עבור המודל עם הקטגוריות המקוריות מדד הדיוק בסט האימון הוא 1 ומדד הדיוק עבור סט הבחינה הוא 0.789. עבור המודל עם הקטגוריות המאוחדות מדד הדיוק בסט האימון הוא 1 ומדד הדיוק עבור סט הבחינה הוא 0.833.

כפי שניתן לראות, בשני העצים מדד הדיוק עבור סט האימון הינו 1 משום שעץ ההחלטה המלא נבנה על בסיס נתוני סט האימון ובכך הותאם העץ לאותם הנתונים - over fitting. כתוצאה מכך, כאשר העץ נבדק על בסיס סט הבחינה מדד הדיוק ירד.

כפי שצפינו ועל סמך מדדי הדיוק על סט הבחינה נבחר להמשיך ולפתח את המודל שעושה שימוש בקטגוריות המאוחדות מכיוון שאחוז הדיוק גבוה יותר. להלן העץ המלא שנבחר:



כוונון פרמטרים

נבחר לכוונן את הפרמטרים בשתי דרכים: הראשונה, בעזרת Grid Search הבוחן את כל הקונפיגורציות האפשריות בטווח ערכים נתון ובכך בהכרח נמצא את הקונפיגורציה הטובה ביותר מבין האפשרויות. השנייה, בעזרת אלגוריתם חמדני אשר בכל פעם ייבחר ערך להיפר-פרמטר ספציפי עד לבחירת ערכים עבור כל היפר-הפרמטרים הרצויים אשר יביא לידי ביטוי תיעדוף פנימי בין היפר-הפרמטרים. לבסוף, נרצה לבחור בכוונון אשר יוביל למיקסום מדד הדיוק על סט הבחינה.

היפר-פרמטרים שבחרנו הם:

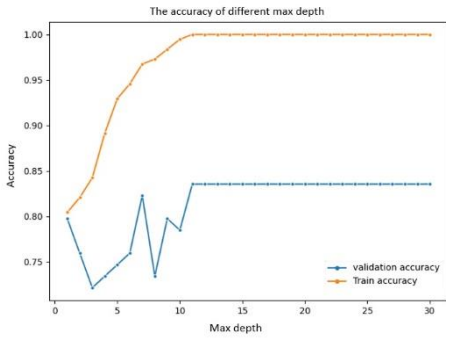
- Criterion: מדד שעל בסיסו נכריע איזה מאפיין יבחר בכל פיצול, כך שהמאפיינים המשמעותיים ביותר יבחרו ראשוניים. נבחן את המדדים entropy, gini, log loss. כל אחד מהם מחשב באופן שונה את חשיבותו של מאפיין, ולכן נרצה לבחור בקריטריון אשר יוביל לדיוק גבוה יותר.
- Max_depth: קריטריון המייצג את העומק המקסימלי של העץ במודל, כלומר מספר הצמתים המקסימלי בין שורש העץ לעלים. ככל שעומק העץ גדול יותר כך גדל הסיכוי ל-over fitting, ככל שיש פחות רמות גדל הסיכוי ל-under fitting ולכן בכדי למצוא עומק עץ שיסווג בצורה טובה גם את התצפיות של סט הבחינה עבור Grid search נבחן את עומקי העץ [3-10]. בעת ביצוע האלגוריתם החמדני נרצה לבחור מבין כלל האפשרויות כלומר [1-30].
- Max_features: קריטריון המייצג את מספר המאפיינים אשר מבניהם נבחר את הצומת הבאה בעץ. כלומר, לא נבחן את כלל המאפיינים שטרם נבחרו, אלא נבחן רק מספר מוגבל של מאפיינים רנדומליים בכל צומת. ככל שמספר המאפיינים יהיה קטן יותר ייתכן ולא נבחר את המאפיינים המשפיעים (under fitting) בעוד שבמידה ונבחר מתוך מספר גדול של מאפיינים אנו בוחרים אותם על סמך סט האימון וישנו חשש ל-over fitting. לכן, ב-Grid Search נרצה לבחון את טווח הערכים [5-10] משום שמדובר בגודל ביניים המאזן בין כמות המאפיינים הקיימים לכמות המאפיינים שיבחנו. בדומה לקריטריון mas depth, באלגוריתם החמדני נרצה לבחון את טווח הערכים [1-30].

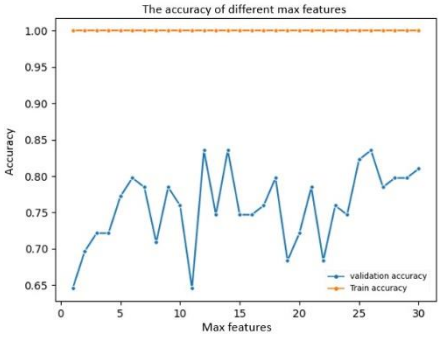
- כיוונון בעזרת אלגוריתם חמדני

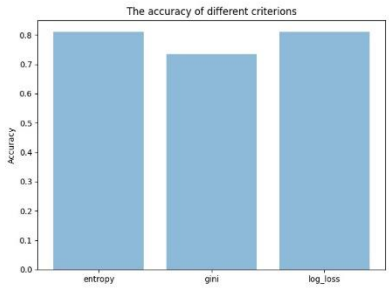
ראשית, נבחר לכוונן את הCriterion. נמצא כי הקריטריונים שממקסמים את מדד הדיוק הם Entropy ו-Log Loss.

לאחר מכן, נרצה לכוונן את הפרמטר Max Features עבור הקריטריונים שנבחרו בכוונון הראשון. נמצא כי ישנם מספר מאפיינים מקסימליים אשר ממקסמים את ערך הדיוק והם 12,14,26. כפי שניתן לראות בגרף, מספר המאפיינים המקסימלי אינו משפיע על הדיוק בסט האימון (שהינו 1) מכיוון שבכל המקרים האלגוריתם יסווג על בסיס עץ מלא. לכן, נעדיף את המודל הפשוט ביותר אשר יצמצם את הסבירות ל over fitting, נבחר 12 מאפיינים מקסימליים.

לבסוף, נרצה לכוונן את הפרמטר Max Depth עבור שני ההיפר-הפרמטרים שכוונו עד כה. נמצא כי עומקי העץ שממקסמים את ערך הדיוק על סט האימות הם 11 ומעלה. לכן בכדי לקחת את המודל הפשוט והכללי ביותר נבחר בערך 11.







לבסוף קיבלנו עץ אשר נבנה על סמך ההיפר פרמטרים: קריטריון הבחינה - entropy, עומק עץ -11, מספר מאפיינים מקסימלי - 12. לאחר אימון המודל על בסיס היפר פרמטרים אלו נמצא כי הדיוק על סט האימון המלא הוא 0.794 והדיוק על סט הבחינה הוא 0.781.

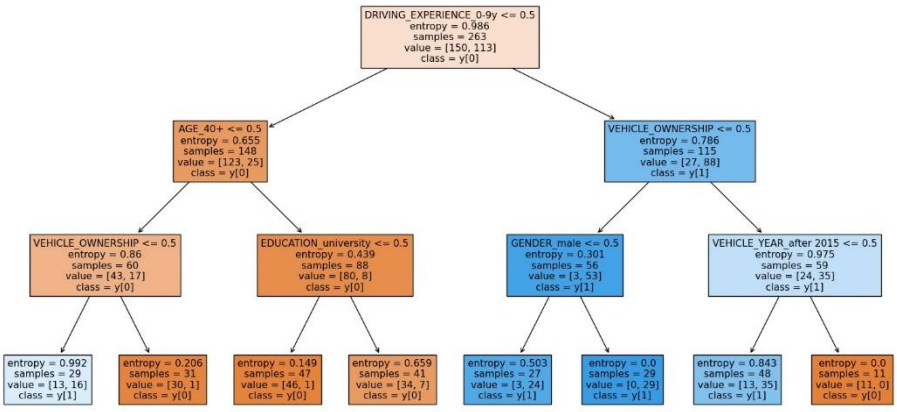
- <u>כיוונון בעזרת Grid Search (cv=10)</u>

עבור כל קונפיגורציה בחרנו לחלק את התצפיות ל-10 תתי קבוצות כאשר בכל פעם אחת מתתי הקבוצות תשמש כסט ולידציה. ערך זה מהווה איזון בין מספר רב של תתי קבוצות שמטרתן היא להקטין את התאמת היתר, אל מול קביעת גודל הקבוצות כך שיכילו מספר מספיק של תצפיות מהן נוכל ללמוד.

param_criterion	param_max_depth	param_max_features	mean_test_score	rank_test_score
entropy	3	8	0.84373	1
log_loss	3	8	0.84373	1
gini	4	10	0.83989	3
log_loss	4	8	0.83234	4
entropy	4	8	0.83234	4
log_loss	5	6	0.82906	6
entropy	5	6	0.82906	6
gini	4	8	0.82863	8
log_loss	4	6	0.82094	9
entropy	4	6	0.82094	9
gini	4	9	0.82094	11
log_loss	5	7	0.82080	12
entropy	5	7	0.82080	12
entropy	5	9	0.82080	12
log_loss	5	9	0.82080	12

לאחר ביצוע כיוונון היפר-הפרמטרים נמצא כי ישנן שתי קונפיגורציות שממקסמות את מדד הדיוק על סט הוולידציה (0.84373). משום שאין הבדל במדד הדיוק, אלא רק בקריטריון נבחר את הקונפיגורציה הבאה כמודל: קריטריון הבחינה - entropy, עומק עץ - 3, מספר מאפיינים מקסימלי - 8. לאחר אימון המודל על בסיס היפר-פרמטרים אלו נמצא כי דיוק המודל על סט האימון הוא 0.844, ודיוק המודל על סט הבחינה הוא 0.816.

מכיוון שכיוונון הפרמטרים בעזרת Grid search הניב דיוק גבוה יותר על סט הבחינה, נשתמש בעץ המבוסס על היפר הפרמטרים הללו ונקבל את המודל הבא:

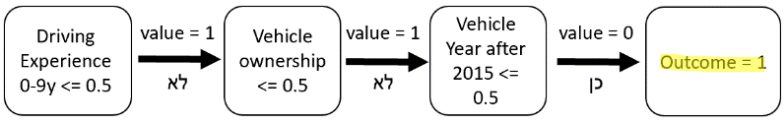


בהסתכלות על דיוק המודל עבור סט האימון הנוכחי למול דיוק סט האימון עבור עץ מלא, הדיוק אינו מקסימלי (100%) משום שכעת המודל מסווג בעזרת כמות מוגבלת של מאפיינים (עומק העץ לא מקסימלי ומספר המאפיינים אינו מקסימלי). בנוסף, בהסתכלות על סט הבחינה עבור שני מודלים אלו, נמצא כי במודל זה הדיוק הוא נמוך יותר. למרות זאת, נבחר במודל החלקי מכיוון ונרצה לבצע הכללה אשר תתאים את המודל לסט נתונים עתידי שאינו ידוע וייתכן שבעץ המלא התקבל מצב של over fitting מול סט האימון.

נבחן את סיווגה של רשומה ID=150 אשר נמצאת בסט הבחינה על סמך העץ שקיבלנו.

ID	AGE	GENDER	DRIVING_EXPERIENCE	EDUCATION	INCOME	CREDIT_SCORE	VEHICLE_OWNERSHIP	VEHICLE_YEAR	MARRIED	CHILDREN	POSTAL_CODE
150	40+	female	0-9y	university	upper class	0.766660479	1	before 2015	0	1	32765 + 92101

ANNUAL_MILEAGE	VEHICLE_TYPE	SPEEDING_VIOLATIONS_NEW	PAST_ACCIDENTS_NEW	AGE_BEFORE	POSTAL_CODE_BEFORE	OUTCOME
11000	sedan	0	0	40-64	32765	1

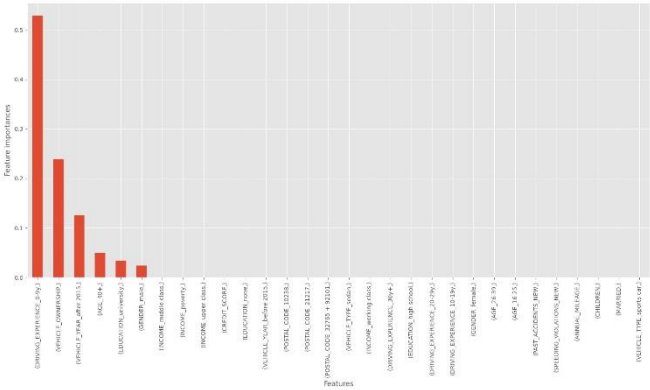


כפי שניתן לראות, הרשומה סווגה למחלקה 1 (הנהג יתבע את הביטוח) בהתאמה לסיווגה האמיתי.

- <u>תובנות:</u>

1. המאפיין המפריד בצורה הטובה ביותר בין המחלקות הוא Driving experience 0-9y, כלומר האם לנהג יש 9 או פחות שנות ניסיון. ניתן לראות שבמידה ולנהג יש יותר מ9 שנות ניסיון, רוב הסיכויים שהוא לא יתבע את הביטוח, בניגוד לנהג אשר ניסיונו הוא 9 שנים או פחות – אשר רוב הסיכויים יתבע את הביטוח. זאת כמובן, כפי שצפינו בחלק א' של הפרויקט שמשתנה זה הינו בעל חשיבות גבוהה בהשפעה על הסיווג למחלקות וכי נהגים עם פחות ניסיון יתבעו את הביטוח בהסתברות גבוהה יותר.
2. המאפיין Vehicle ownership חוזר בשני ענפי העץ המרכזיים, ולכן ניתן להסיק מכך שהוא משתנה משמעותי בעת הסיווג בין המחלקות.
3. בניגוד לציפיותינו, לא נבחרו מאפיינים המעידים על אופי הנהיגה של הנהג. כלומר, העץ לא מסווג על בסיס היסטוריית הנהיגה של הנהג (מספר תאונות, מספר עבירות מהירות). המאפיינים שנבחרו מספקים פרטים כללים אודות האדם עצמו (גיל, מגדר, השכלה).
4. כפי שניתן לראות, לאחר הפיצול ברמה התחתונה של העץ קיימים מאפיינים אשר מסווגים לאותה המחלקה. לכן, בעת הגעה לצומת מסוג זה, בפועל לא יהיה צורך בבדיקת ערך המאפיין. מצב זה מתקיים מכיוון ועומק העץ הוגבל לאחר שהאלגוריתם בחר את המאפיין ללא ידיעה שזוהי הרמה האחרונה.

המטרה של ה-features importance היא לתת ציון לרמת החשיבות של כל מאפיין במודל. ערך ה-features importance של כל מאפיין מחושב על סמך ההפרש של ציון המודל עם הערך המקורי לבין הציון כאשר ערך המאפיין הוגרל. כאשר ערך features importance גבוה מציין מאפיין משמעותי יותר במודל.



כפי שניתן לראות בערכי ה-features importance שלושת המאפיינים שקיבלו את הציון הגבוה ביותר הם: driving experience (0-9y), vehicle ownership, vehicle year (after 2015). כלומר, הם המאפיינים בעלי החשיבות הגבוה ביותר. זאת, בהתאמה לכך ששניים ממאפיינים אלו נבחרו כצמתים בראשית העץ עובדה המעידה על היותם משמעותיים בתהליך הסיווג. בהתאמה, המאפיינים שקיבלו את הציון 0 (לדוגמה, המאפיין Income) לא נכנסו למודל.

כפי שניתן לראות, המאפיינים annual mileage, vehicle type קיבלו את הציון 0 ולא נכנס למודל. זאת בהתאמה להשערותינו במחקר המקדים בחלק א' אשר העלה חשד שקשה להפריד בערך משתנה המטרה על סמך מאפיין זה, ולכן הוסרו מבסיס הנתונים.

כמו כן, ציפינו שהמאפיינים speeding violations ו-past accidents יקבלו ציון גבוה ויכנסו אל המודל, זאת בניגוד לתוצאות שהתקבלו. סיבה אפשרית לפער זה נובעת מקשר חזק בין המשתנים שנבחרו להיכנס אל המודל, אשר מעידים על אופי הנהג כבן אדם. ייתכן ואופי האדם משפיע הן על אופי הנהיגה ועל סיכוייו לתבוע את הביטוח, יותר מהיסטוריית הנהיגה שלו.

Neural Networks

אלגוריתם זה עושה שימוש בכלל המאפיינים של הבעיה (בניגוד לעץ החלטה), ולכן נאמן את המודל רק על המשתנים הרלוונטיים על סמך חלק א' של הפרויקט, כלומר ללא המאפיינים שהוסרו. כשלב מקדים לרשת הנוירונים, נבצע נורמליזציה של הנתונים בכדי שלכלל המאפיינים יהיה את אותו טווח ערכים בין 0 ל-1.

מודל ברירת המחדל

תחילה, נאמן ונבחן את רשת הנוירונים עם ערכי ברירת המחדל: hidden layer size – (100,), learning rate init – 0.001, max iter – 200, relu=activation. בקונפיגורציה זו ישנם 30 נוירונים בשכבת הכניסה (אחד עבור כל מאפיין) ושכבה חבויה אחת עם 100 נוירונים. הדיוק שהתקבל על סט האימון הוא 0.951, והדיוק שהתקבל על סט הבחינה הם 0.842. ניתן להסיק מכך שהשגיאה עדיין אינה אפסית וייתכן שבעזרת איטרציות נוספות נוכל להמשיך ולמזער אותה.

כוונון פרמטרים

היפר הפרמטרים אותם נרצה לכוונן הם:

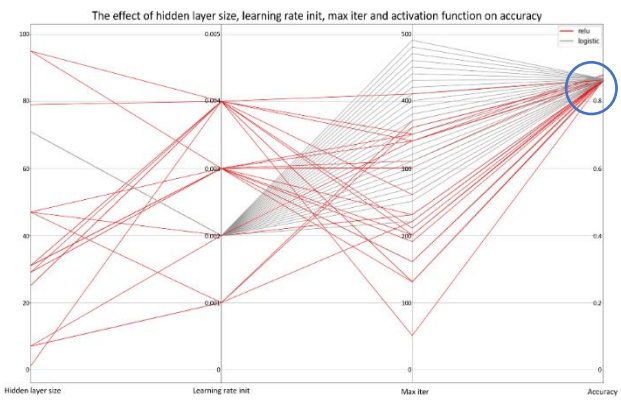
- Hidden layer size: היפר פרמטר זה מאגד בתוכו את מספר השכבות החבויות ומספר הנוירונים בכל שכבה חבויה. ככל שמספר השכבות יהיה גבוה יותר המודל יאפשר למידה מעמיקה יותר ופתרון של בעיות מורכבות יותר, אך במקביל ישנו חשש להתאמת יתר של המודל. לכן, נרצה לבחון שני במצבים:

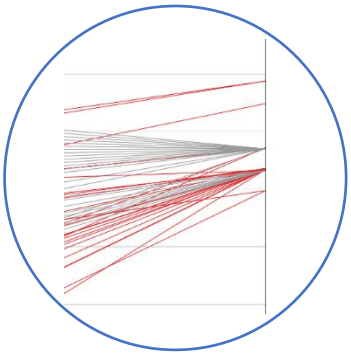
הראשון עם שכבה חבויה אחת וטווח הנוירונים הוא [1,100] והשנייה עם שתי שכבות חבויות כאשר מספר הנוירונים בכל שכבה הינו זהה ובטווח ערכים [1,100].

- Max iter: היפר פרמטר זה מגביל את מספר האיטרציות שנבצע במידה ולא הגענו ל- learning rate הרצוי. בחרנו לבחון את טווח הערכים [1,500] בכדי לוודא שנקבל מודל אשר מצליח להתכנס ובכך לצמצם את השגיאה.
- Activation: היפר פרמטר זה מכיל בתוכו מספר פונקציות אקטיבציה אפשריות של הנוירון המבצעת טרנספורמציה בין הקלט של הנוירון לפלט שלו. פונקציה זו משפיעה על תהליך הלמידה על סמך השגיאות שהתקבלו במהלך האימון, ולכן נרצה לבחון את הפונקציות relu, logistic.
- Learning rate init: קצב הלמידה הקובע את מהירות השינוי של ערכי המשקולות כתגובה לאמידת הטעות בעת תהליך העדכון שלהן. קצב קטן מידי יאריך את תהליך הלמידה בעוד קצב גדול מידי עלול לגרום להתכנסות לנקודת מקסימום מקומית. לכן, נרצה לכוונן ולבחון את הפרמטר בכדי למצוא איזון ביניהן. נבחן ערכים בטווח [0.001,0.005].

- שכבה חבויה אחת

עבור כיוונון הפרמטרים עם שכבה חבויה אחת נשתמש ב-Grid search (cv=5, בחרנו בcv קטן יותר ביחס לעץ ההחלטה משיקולי זמן ריצה). לאחר ביצוע כיוונון ההיפר-פרמטרים נמצא כי הקונפיגורציה שממקסמות את מדד הדיוק על סט הוולידציה היא hidden layer size - 31, learning rate init - 0.003, max iter - 341, activation - relu.



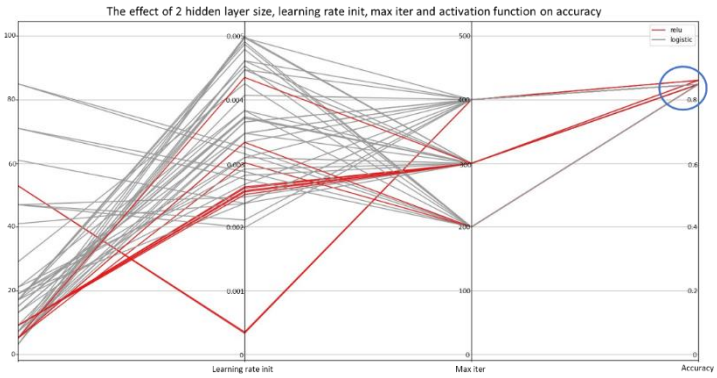


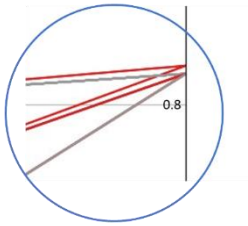
בגרף מוצגות 50 הקונפיגורציות הטובות ביותר שנמצאו ע"י ה-Grid search. כפי שניתן לראות, שלושת הקונפיגורציות הטובות ביותר, הן עבור פונקציית האקטיבציה relu. בנוסף, טווח הערכים בהם היפר-הפרמטר max iter מניב דיוק גבוה על סט הבחינה הוא רחב. לאחר אימון הרשת על בסיס היפר-פרמטרים אלו נמצא כי דיוק המודל על סט האימון הוא 0.977, ודיוק המודל על סט הבחינה הוא 0.842.

- שתי שכבות חבויות

עבור כיוונון הפרמטרים עם שתי שכבות חבויות נשתמש בלולאות מקוננות, אשר יבחנו את כלל הקומבינציות. לאחר ביצוע כיוונון ההיפר-פרמטרים נמצאו מספר קונפיגורציות אשר מדד הדיוק שלהן הוא הגבוה ביותר. נבחר בקונפיגורציה הפשוטה ביותר (מספר איטרציות ומספר נוירונים קטן) והיא hidden layer size- (9,9), learning rate init - 0.003, max iter - 261, activation - relu.

בגרף מוצגות 50 הקונפיגורציות הטובות ביותר שנמצאו ע"י הלולאות המקוננות. כפי שניתן לראות, גם כעת הקונפיגורציות הטובות ביותר הן עבור פונקציית האקטיבציה relu. בנוסף, בניגוד לתוצאות של הקונפיגורציות עם שכבה חבויה אחת קיימים שלושה ערכי היפר פרמטר max iter הממקסמים את מדד הדיוק. כמו כן, קיימים רק שתי ערכי דיוק על סט הבחינה עבור קונפיגורציות אלה. לאחר אימון הרשת על בסיס היפר-פרמטרים אלו נמצא כי דיוק המודל על סט האימון הוא 0.977, ודיוק המודל על סט הבחינה הוא 0.86.





הרשת בעלת שתי השכבות קיבלה מדד דיוק גבוה יותר עובדה שנובעת מכך שככל שמספר השכבות עולה כך גם יכולת הלמידה על סמך הקשרים בין המאפיינים. בנוסף, כפי שניתן לראות מספר האיטרציות גדל לעומת המודל הראשון – וכתוצאה מכך קטנה השגיאה ומדד הדיוק עלה. ולכן המודל שייבחר הוא רשת נוירונית עם שתי שכבות, כאשר בכל שכבה חבויה ישנם 9 נוירונים וישנם 30 נוירונים בשכבת הכניסה.

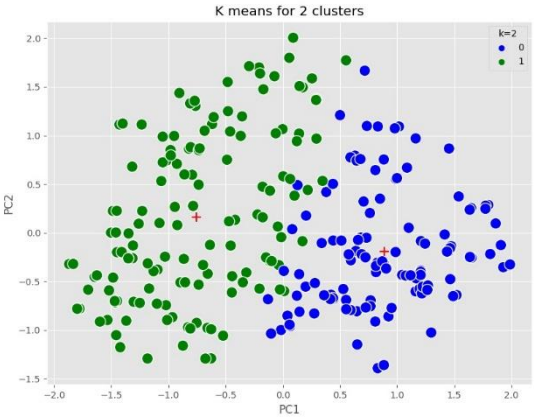
K-means

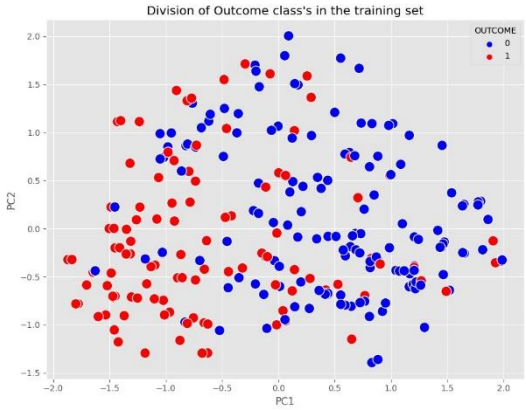
כעת, ניישם את אלגוריתם k-means שמטרתו לחלק את התצפיות לאשכולות בתהליך למידה לא מונחת. הגדרנו את מספר האשכולות להיות שניים בהתאם לבעיית הסיווג אותה אנו ממדלים. האלגוריתם חילק את התצפיות לשתי אשכולות, מכיוון ולא ניתן לדעת איזה מחלקה כל אשכול מייצג, נבחן את הסנטרואיד של כל אשכול. מבדיקה שביצענו על הסנטרואידים מצאנו כי האשכול הראשון (0) מאופיין בנהגים בקבוצת גיל מבוגרת עם הרבה שנות ניסיון, נשואים ובעלי משפחות, השכלה גבוהה ועמידות כלכלית הכוללת דירוג אשראי גבוה ורכב בבעלותם.

מנגד, האשכול השני (1) מאופיין בנהגים צעירים ללא שנות ניסיון רבות, רמת השכלה נמוכה יותר ועמידות כלכלית נמוכה הכוללת דירוג אשראי נמוך וללא רכב בבעלותם. ממחקר מקדים שביצענו בחלק א' מצאנו כי נהגים צעירים ללא ניסיון מעורבים יותר בתאונות דרכים, ונהגים המבוססים יותר כלכלית נוטים פחות לתבוע את הביטוח. לכן, עקב ההבדלים בין האשכולות בגילאים, במספר שנות הניסיון וברמה הסוציו- אשר לא יתבעו את הביטוח בשנה הקרובה, לעומת האשכול השני המייצג נהגים שיתבעו את הביטוח.

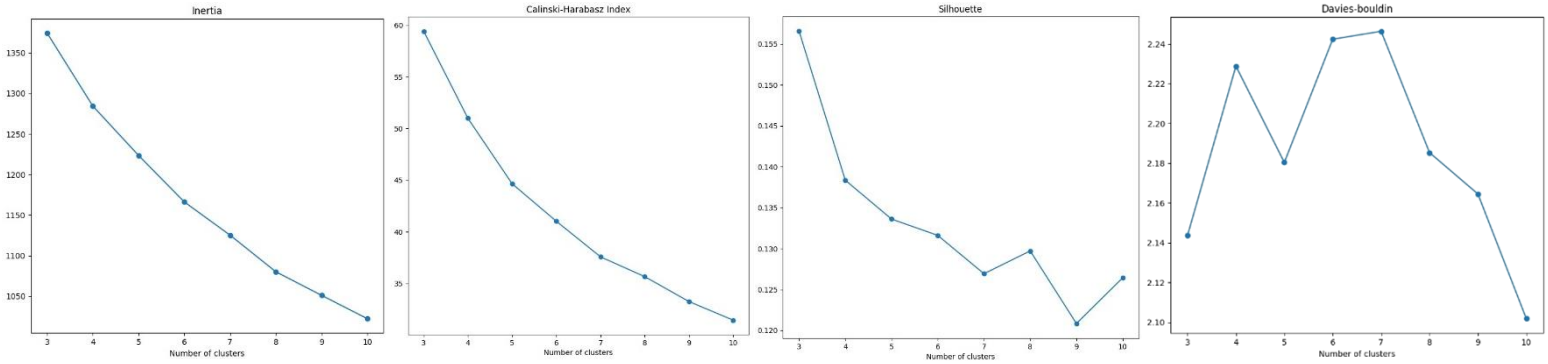
בתרשים הימני ניתן לראות את החלוקה למחלקות על בסיס הסיווג האמיתי של משתנה המטרה, בעוד שבתרשים השמאלי ניתן לראות את החלוקה לאשכולות לאחר יישום האלגוריתם. בניגוד ללמידה לא מונחת, במצב זה ידוע הסיווג למחלקות ולכן ניתן לבחון את אחוז הדיוק של החלוקה לאשכולות. בהשוואה למול סט הבחינה נמצא כי המודל דייק ב-76.3%.

על כן, ניתן להסיק כי החלוקה לאשכולות שביצענו באמצעות מחקר הסטרואידים התבצע כראוי.





כעת, נאמן שמונה מודלים כך שבכל מודל מספר אשכולות שונה (3-10 אשכולות). בכדי לבחור את כמות האשכולות המועדפת נרצה שהשונות בתוך האשכול (homogenetiy) תהיה מינימלית, בעוד שהשונות בין האשכולות (seperation) תהיה מקסימלית. בגרפים מוצגים מדדים המתחשבים בפרמטרים אלו. לדוגמה, מדד האינרציה מחושב באמצעות סכום ריבועי המרחקים של כל תצפיות לנקודת המרכז שלהן, לכן מודל טוב מאופיין עם מדד אינרציה נמוך. על מנת למצוא את כמות האשכולות האופטימלית נמצא את הנקודה שבה האינרציה מתחילה להאט (תמורה שולית פוחתת).

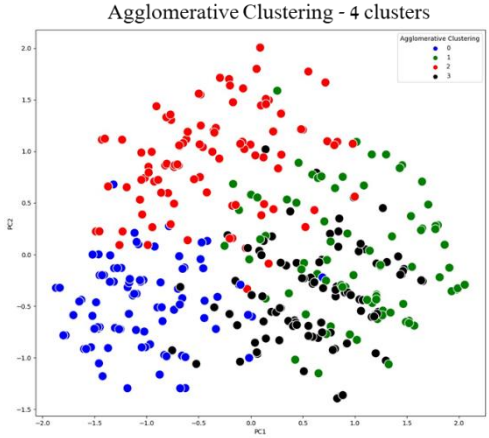


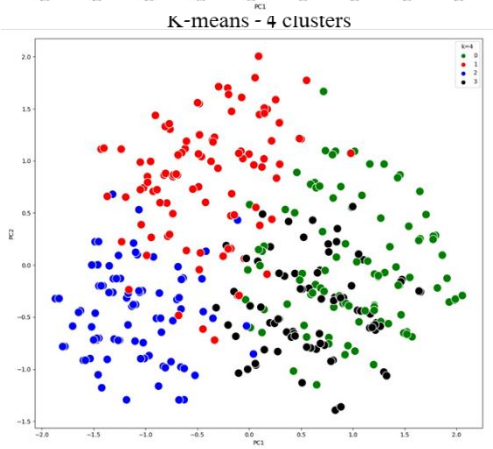
את המדד davies-bouldin נרצה למקסם בעוד שאת שאר המדדים נרצה למזער. ניתן לראות כי עבור המדדים inertia, calinski harabasz index, shihouette השינוי המשמעותי ביותר הינו 4/5, 4/5, 4 בהתאמה. כמו כן, ערך מדד ה-davies-bouldin המקסימלי הוא שבע, אבל השיפור המשמעותי ביותר גם הוא עבור הערך ארבע. לכן, על סמך כלל המדדים נבחר k=4.

חלוקה לארבע אשכולות לא מייצגת את בעיית הסיווג שלנו בה קיימת חלוקה לשתי מחלקות בלבד. עם זאת, עקב כמות מאפיינים רבה הלקוחים מעולמות תוכן שונים אודות הנהגים (מאפיינים חברתיים, כלכליים, משפחתיים) ייתכן וניתן לחלק את התצפיות ליותר משתי קבוצות, מלבד הסיווג לתביעת הביטוח, על בסיס קווי הדימיון והקשרים בין התצפיות עבור עולמות תוכן שונים.

נבחן את החלוקה לארבעה אשכולות גם באמצעות האלגוריתם **Agglomerative Clustering** (AC) המחלק לאשכולות באופן שונה מאלגוריתם K-means. אלגוריתם זה מוצא את הנקודות שהכי קרובות אחת לשנייה ומאחד אותם לאשכול. במהלך התהליך הוא משייך את התצפיות הנותרות גם אל האשכולות

שנמצאו בהתאם לקרבתם. האלגוריתם מסתיים כאשר כל התצפיות סווגו לאחת מ-K האשכולות הרצויות. אלגוריתם זה אינו עושה שימוש בסנטרואידים של האשכולות, ולכן יכולת ההפרדה שלו בין המחלקות מוגבלת יותר. עם זאת, זמן הריצה של שני האלגוריתמים דומה – Kmeans:O(n^2), AC:O(kn^2).





בהשוואה בין האלגוריתמים עבור ארבעה אשכולות ניתן לראות כי שניהם חילקו באופן דומה את התצפיות לאשכולות. נציין כי אין משמעות למספר האשכול ולכן נבחין בינהם באמצעות הצבעים שלהם. מבדיקה שביצענו על הסטרואידים של האשכולות שהתקבלו נמצא כי קיים דימיון רב במאפיניים של כל אחד מהאשכולות. לדוגמה, עבור שני האלגוריתמים האשכול השחור מאופיין בנהגים מבוגרים, מבוססים כלכלית עם הרבה שנות ניסיון, והאשכול הכחול מאופיין בנהגים צעירים, ללא ניסיון שלרוב אינם נשואים וללא ילדים. עם זאת, השונות בין מאפייני האשכולות עבור אלגוריתם K-means גבוהה באופן משמעותי מאלגוריתם AC. כלומר, עבור אלגוריתם זה התקבלו תוצאות יותר חד-משמעיות עבור מאפיינים מסוימים (לדוגמה, כל הנהגים שסווגו לאשכול השחור יש רכב ישן, לעומת כל הנהגים שסווגו לאשכול האדום להם רכב חדש). זאת לעומת אלגוריתם AC במאפיניים מסוימים הפיזור אחיד בין האשכולות (לדוגמה, דירוג אשראי וחלוקה לאזורי מגורים).

<u>השוואה בין מודלים</u>

לאחר שבחרנו את הקונפיגורציות הטובות ביותר עבור כל אחד מהאלגוריתמים, נשווה ובניהם ונבחר את המודל שלדעתנו יחזה בצורה הטובה ביותר תצפיות חדשות שהמודלים לא נחשפו אליהם. נציין כי בעולם האמיתי עבור מודל ה-Kmeans שאינו מונחה לא נוכל לחשב את אחוז הדיוק משום שאין סיווג למחלקות. מכיוון ובבעיה זו קיימים הסיווגים למחלקות בסט בחינה נוכל להשתמש בהם ובעזרתם לחשב את אחוז הדיוק של מודל זה. להלן confusion matrix עבור שלושת המודלים:

K-means

	0	1
0	45	14
1	13	42

רשת נוירונית

	0	1
0	55	4
1	12	43

עץ החלטה

	0	1
0	52	7
1	14	41

<u>בחרנו להשוות בין המודלים על בסיס שני מדדים:</u>

1. **אחוז הדיוק על סט הבחינה** (accuracy) - מדד זה משקף את דיוק המודל על סט נתונים חדש. לכן, נרצה למקסם את ערך מדד זה.
2. **מספר הטעויות מסוג ראשון** - סיווג התצפית כנהג שלא יתבע את הביטוח למרות שבפועל הוא אכן יתבע אותו. להערכתנו, טעות מסוג זה מזיקה יותר מטעות מסוג שני משום שסיווג זה עלולה להוביל להפסדים כספים לחברת הביטוח. לכן, נרצה לצמצם את ערך מדד זה.

מודל	אחוז דיוק	מספר הטעויות מסוג ראשון
עץ החלטה	0.781	14
רשת נוירונית	0.86	12
K-means	0.763	13

ניתן לראות על סמך מדד הדיוק כי מודל ה-Kmeans חוזה בצורה הכי פחות טובה תצפיות חדשות. סיבה אפשרית לכך היא שמודל זה אינו בוחר את המאפיינים המשפיעים ביותר, אלא מקבץ את התצפיות לאשכולות על בסיס הדימיון ביניהם. בנוסף, בבעיה זו קיימים מאפיינים בעולמות תוכן שונים (מאפיינים חברתיים, כלכליים, משפחתיים) וייתכן והמודל יסווג את התצפיות לאשכולות כך שמהות המחלקה היא שונה (משתנה המטרה שונה). כמו כן, בניגוד לציפיותינו כי הדיוק של עץ ההחלטה יהיה הגבוה ביותר עקב מספר גבוה של מאפיינים קטגוריאליים, הדיוק של הרשת הנוירונית גבוה יותר. ניתן להסביר זאת משום שבמודל הרשת למידת הקשרים בין המאפיינים עמוקה יותר (צמצום השגיאה).

על פי מדד מספר הטעויות מסוג ראשון, למרות שכלל הערכים קרובים מאוד אחד לשני גם כאן המדד של הרשת הנוירונית הוא הטוב ביותר.

על סמך שני המדדים נעדיף את הרשת הנוירונית כמודל המסווג במידת ההצלחה הטובה ביותר.

<u>המודל הנבחר</u>

המודל שבחרנו כמסווג הטוב ביותר הינו רשת נוירונית בעלת ההיפר-פרמטרים הבאים:
hidden layer size - (9,9), learning rate init - 0.003, max iter - 261, activation - relu. בחרנו בהיפר-פרמטרים אלה משום שהינם המשפיעים ביותר על אופן למידת המודל ודיוק מתן המשקולות בין הנוירונים. פונקציית האקטיבציה מבצעת טרנספורמציה בין מה שנכנס לנוירון לבין מה שיוצא ממנו, כך שהשכבה האחרונה נותנת אינדיקציה לסיווג למחלקות ולכן משמעותית בדיוק המודל. כמו כן, קיימת התאמה בין מספר השכבות החבויות וכמות הנוירונים בכל שכבה לאופן למידת הקשרים בין המאפיינים, לכן בחרנו בהיפר פרמטר זה על מנת שתהליך הלמידה יהיה מדויק ככל הניתן. בנוסף, למספר האיטרציות המקסימלי ולקצב הלמידה של המודל ישנה השפעה על הצלחת המודל, כך שככל שנבצע כמות רבה של איטרציות נצמצם את השגיאה, אך עם נגדיל את קצב הלמידה ייתכן ונגיע בזמן קצר לנקודת מקסימום מקומית ובכך למודל שאינו מדויק.

כפי שהצגנו קודם, בניתוח התוצאות של המודל על סט הבחינה המודל מדייק ב-86%. כפי שניתן לראות, למרות שסט הנתונים הינו מאוזן, הרשת מסווגת יותר תצפיות למחלקה 0. בהתאמה, קיימות יותר טעויות מסוג ראשון, 14 טעויות מסוג ראשון ו-4 טעויות מסוג שני. משום שטעות מסוג ראשון היא משפיעה יותר על היבטים הכלכליים של חברת הביטוח – זוהי חולשה של המודל.

רשת נוירונית

	0	1
0	55	4
1	12	43