



מדען נתונים, לתקף ולבדוק מודלים אתה יודע?

כאשר נתונים משמשים לניבוי או לקביעת אסטרטגיית החלטה, באופן טבעי קיימת סכנה שמודל של למידת המכונה יעבוד טוב על הנתונים ששימשו לפיתוחו, אך לא יכליל (Generalize) בצורה טובה נתונים חדשים. מכאן עולה נקודה ברורה לפיה, חשוב מאוד שהנתונים אשר משמשים לפיתוח מודל של למידת מכונה ייצגו בצורה טובה את המצבים שעליהם ייושם המודל בעתיד. לדוגמא, שימוש בנתונים אודות מכירות, באזור מסוים שבו הלקוחות הינם בעלי הכנסה גבוהה, לצורך ניבוי המכירות עבור מוצר מסוים רק סביר שייתנו תוצאות מוטות.

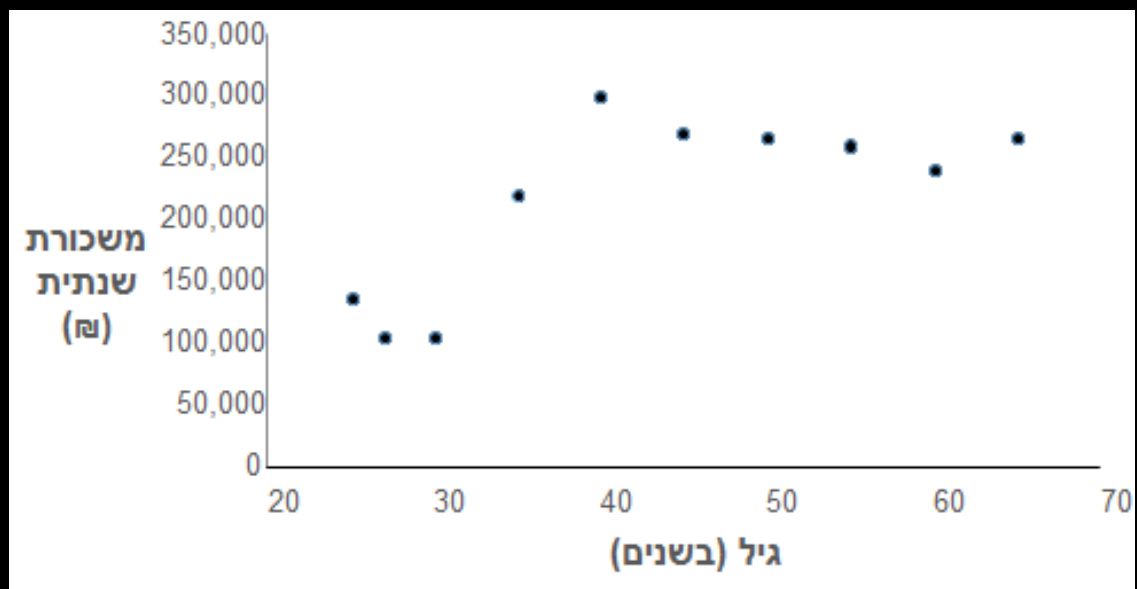
כפי שסטטיסטיקאים הבינו זה מכבר, חשוב ביותר גם לבדוק את המודל באמצעות מדגם אחר (Out-of-Sample). במילים אחרות, יש לבדוק את המודל על מדגם נתונים שונה מזה ששימש לקביעת הפרמטרים של המודל. מדעני נתונים מכנים את מדגם הנתונים שמשמש לפיתוח המודל בשם "סט אימון" (Training Set), בעוד שאת מדגם הנתונים המשמש לקביעת רמת הדיוק של המודל הם מכנים בשם "סט בדיקה" (Test Set), לעיתים קרובות יש גם עושים שימוש בסט תיקוף, (Validation Set).

במאמר זה נמחיש את השימוש בסט אימון ובסט בדיקה באמצעות דוגמא פשוטה. נניח שאנו רוצים לבנות מודל לניבוי/חיזוי המשכורות השנתיות של אנשים שעובדים באותו מקצוע באזור מסוים רק על סמך הגילאים שלהם. אז לקחנו מדגם נתונים של 10 אנשים (לשם המחשה בלבד לקחנו מדגם מאוד קטן. נעיר כי מערכי הנתונים המשמשים בלמידת מכונה גדולים פי עשרות מונים מהמדגם שלנו). נתוני מדגם זה אשר אותו נכנה "סט האימון", מוצגים בלוח שלהלן ומתוארים בתרשים שלהלן:

להלן נתוני סט האימון: משכורות מתוך מדגם מקרי של 10 אנשים שעובדים במקצוע מסוים באזור מסוים.

משכורת שנתית (₪)	גיל (בשנים)
135,000	25
260,000	55
105,000	27
220,000	35
240,000	60
265,000	65
270,000	45
300,000	40
265,000	50
105,000	30

להלן תרשים פיזור (Scatter Plot) של נתוני סט האימון:

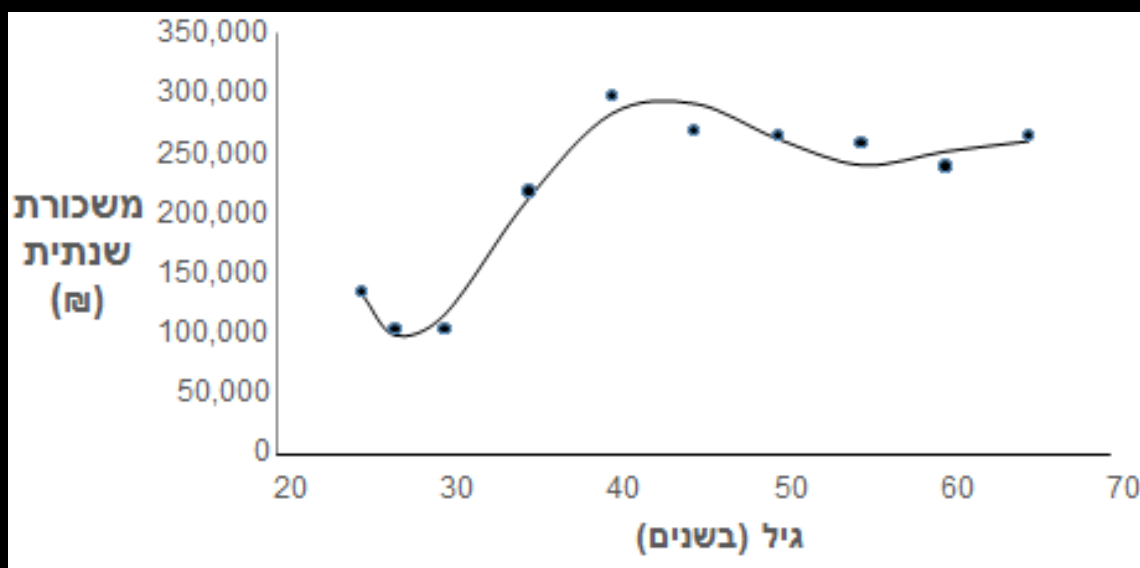


אין דבר מפתה יותר מלפתח מודל שיתאים את עצמו בצורה הטובה ביותר לנתונים הקיימים. מניסיוני כמדען נתונים מודל פולינום מסדר חמישי "עושה את העבודה" כמו שאומרים. למי שלא מכיר, ככה נראה מודל פולינום מסדר חמישי:

$$Y = a + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5$$

כאשר Y הוא המשכורת השנתית ו-X הוא הגיל של מקבל המשכורת. התוצאות עבור רמת ההתאמה של מודל הפולינום מסדר חמישי לנתוני סט האימון מוצגים בתרשים שלהלן.

להלן תוצאות טיב ההתאמה (Goodness of Fit) של מודל הפולינום מסדר חמישי לנתוני סט האימון:



כפי שניתן לראות, מודל הפולינום מסדר חמישי מספק התאמה טובה לנתוני סט האימון. סטיית התקן של הפרשים שבין המשכורת החזויה על בסיס המודל והמשכורת בפועל עבור עשרת האנשים שסט האימון, מכונה "שורש השגיאה הריבועית הממוצעת" (RMSE- Root)



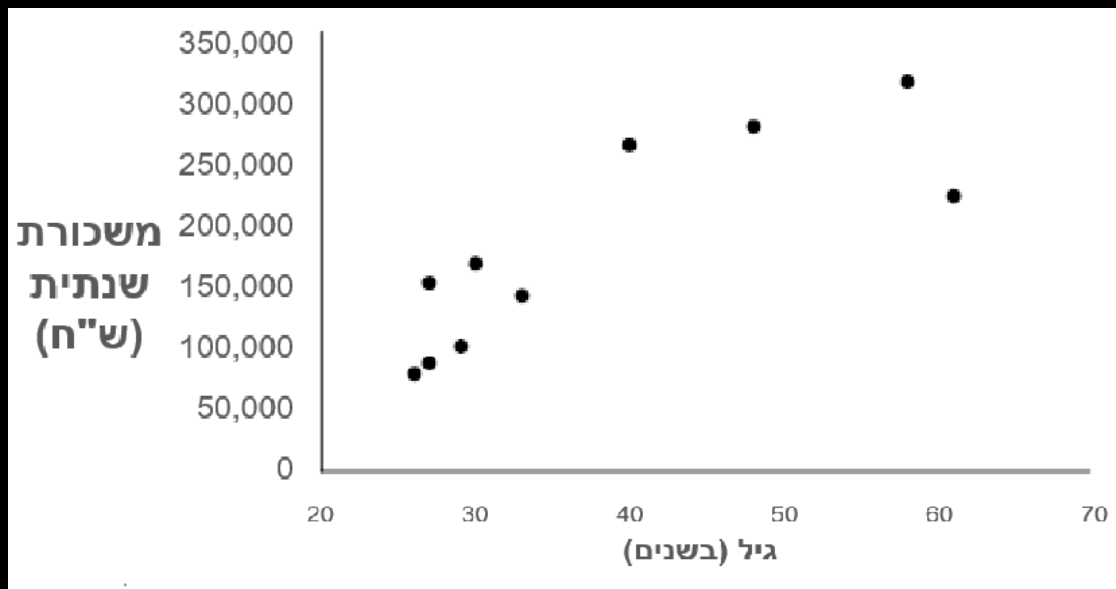
(Mean Square Error) והיא נאמדה על ידינו בכ- 12,902 ₪.

עם זאת, ההיגיון הבריא אומר לנו שיכול להיות שהמודל שלנו מתאים יתר על המידה (Over-Fits) לנתוני סט האימון וזאת בשל העובדה שהעקום שבתרשים לעיל לא נראה מציאותי בעליל: העקום יורד, עולה, יורד ואז עולה שוב ככל שהגיל עולה. למעשה עלינו לבדוק את המודל באמצעות נתוני מדגם אחר (Out-of-Sample). ואם נעבור משפת האקונומטריקה לשפת מדע הנתונים, אנו צריכים לקבוע האם המודל מכליל (Generalize) בצורה טובה נתונים חדשים ששונים מהנתונים אשר שימשו לפיתוח המודל.

להלן נתוני סט הבדיקה: משכורות מתוך מדגם מקרי של 10 אנשים נוספים שעובדים במקצוע מסוים באזור מסוים.

משכורת שנתית (₪)	גיל (בשנים)
166,000	30
78,000	26
310,000	58
100,000	29
260,000	40
150,000	27
140,000	33
220,000	61
86,000	27
276,000	48

להלן תרשים פיזור של נתוני סט הבדיקה:



כאשר אנו משתמשים במודל פולינום מסדר חמישי אנו מגלים ששורש השגיאה הריבועית הממוצעת (RMSE) שלו על נתוני סט הבדיקה הוא בערך 38,794 ₪, הרבה יותר גבוה משורש השגיאה הריבועית הממוצעת שלו על נתוני סט האימון (12,902 ₪).

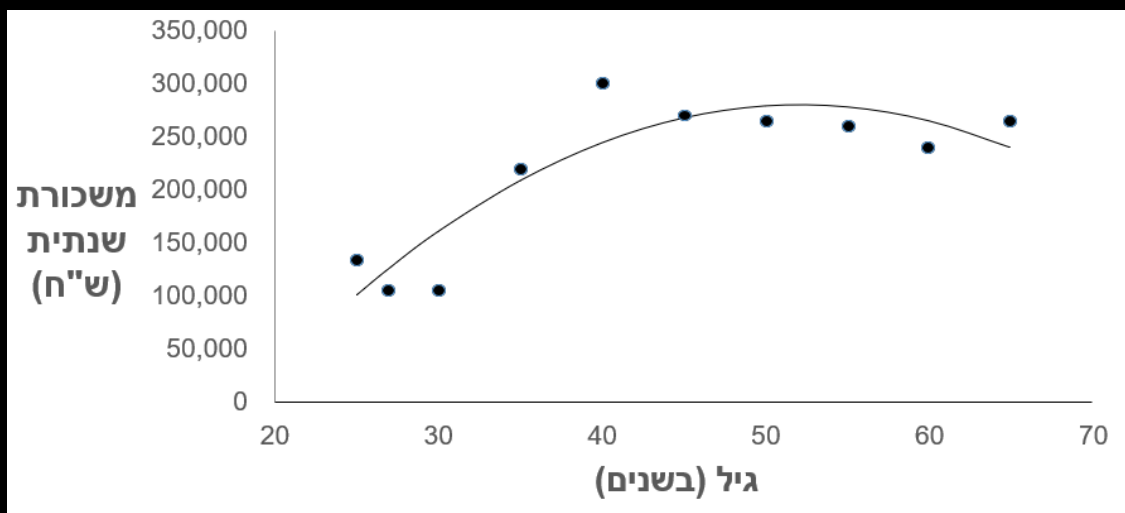
זוהי אינדיקציה ברורה לכך שמודל הפולינום מסדר חמישי מתאים יתר על המידה לנתוני סט האימון, משמע, הוא אינו מכליל בצורה טובה את נתוני סט הבדיקה.

במקרה שכזה, הצעד הטבעי והמתבקש הוא לבחון מודל פשוט יותר, כמו למשל מודל ריבועי:

$$Y = a + b_1X + b_2X^2$$

כלומר, פולינום מסדר שני.

להלן תוצאות טיב ההתאמה של המודל הריבועי לנתוני סט האימון:



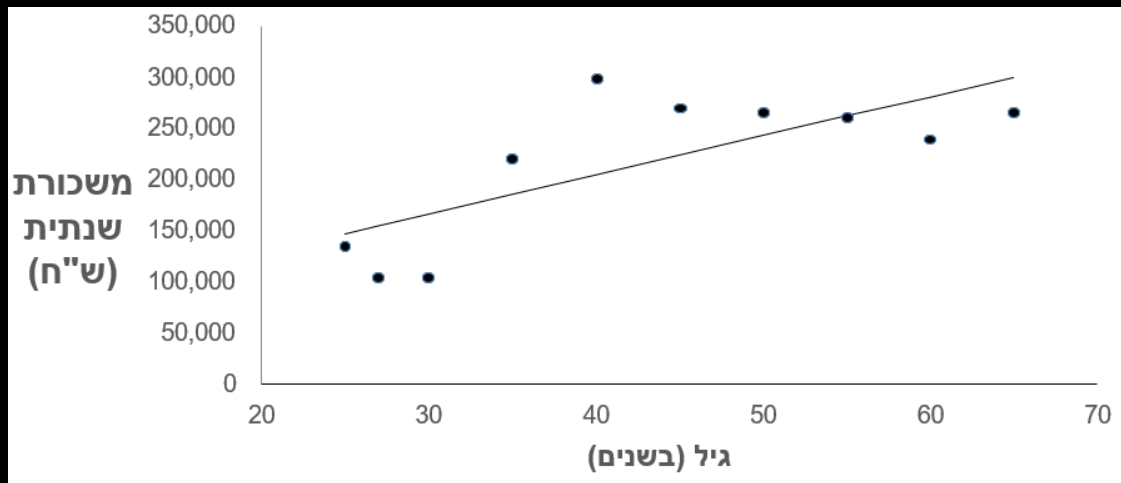
כמובן שההתאמה איננה טובה כמו זו של מודל הפולינום מסדר חמישי. שורש השגיאה הריבועית הממוצעת (RMSE) של המודל הריבועי על נתוני סט האימון הוא 32,932 ₪. עם זאת, המודל מכליל בצורה טובה נתונים חדשים. שורש השגיאה הריבועית הממוצעת (RMSE) של המודל הריבועי על נתוני סט הבדיקה הוא 33,554 ₪, קצת פחות גרוע משורש השגיאה הריבועית הממוצעת שלו על נתוני סט האימון (32,932 ₪). לפיכך, ניתן רק לצפות שהמודל הריבועי יספק תחזיות טובות יותר למשכורת השנתית עבור עובד בגיל מסוים מאשר מודל מורכב יותר כמו מודל הפולינום מסדר חמישי.

כאמור המודל הריבועי הינו פשוט יותר ממודל הפולינום מסדר חמישי וגם עובד טוב יותר ממנו ככלי ניבוי. עם זאת, אין זה אומר בהכרח שמודלים פשוטים יותר לעולם יהיו טובים יותר ממודלים מורכבים. לשם המחשה, הבה נבחן מודל עוד יותר פשוט יותר, כמו למשל מודל ליניארי:

$$Y = a + b_1X$$

כלומר, פולינום מסדר ראשון.

להלן תוצאות טיב ההתאמה של המודל הלינארי לנתוני סט האימון:



מבחינה ויזואלית ניתן לראות כי המודל הלינארי אינו מצליח "לתפוס" את שיעור הדעיכה (או שיעור הצמיחה היוורד) של המשכורות עבור אנשים מעל לגיל 50. ממצא זה מקבל אישור באמצעות העובדה ששורש השגיאה הריבועית הממוצעת (RMSE) של המודל הלינארי על נתוני סט האימון הוא 49,731 ₪, הרבה יותר גרוע משורש השגיאה הריבועית הממוצעת שלו על נתוני סט האימון (32,932 ₪).

להלן סיכום תוצאות שורש השגיאה הריבועית הממוצעת של שלושת המודלים עבור שני סטים של הנתונים:

מודל לינארי	מודל ריבועי	מודל פולינום מסדר חמישי	
49,731	32,932	12,902	סט האימון
49,990	33,554	38,794	סט הבדיקה



מהטבלה לעיל עולה כי הן המודל הלינארי והן המודל הריבועי מכלילים בצורה טובה נתונים חדשים, בעוד שמודל הפולינום מסדר חמישי איננו מכליל בצורה טובה נתונים חדשים. לאמור- מודל הפולינום מסדר חמישי מתאים יתר על המידה לנתוני סט האימון, בעוד שהמודל הלינארי מתאים חסר על המידה לנתוני סט האימון.

כיצד אם כך ניתן להשיג את האיזון בין התאמת יתר להתאמת חסר? זוהי סוגיה חשובה בלמידת מכונה. חלק מהאלגוריתמים של למידת מכונה דורשים מספר גדול מאוד של פרמטרים ולכן קל מאוד להגיע למצב של התאמת יתר כאשר סט האימון הוא גדול. בהתבסס על המקרה הפשוט שניתחנו במאמר זה, ניתן לנסח את כלל האצבע הבא:

יש להמשיך ולהעלות את רמת המורכבות של המודל עד אשר נתוני סט הבדיקה (קרי, נתוני מדגם אחר) מצביעים על כך שהמודל כבר איננו מכליל בצורה טובה נתונים חדשים.

בדוגמת המשכורות שלנו השתמשנו בשני סטים של נתונים. האחד (סט האימון) שימש לפיתוח המודל והאחר (סט הבדיקה) שימש לבדיקת המודל על נתוני מדגם אחר.

באקדמיה, כלכלנים אמפיריים (קרי, אקונומטריקאים וסטטיסטיקאים) מפתחים תמיד רק מודל אחד ולאחר מכן מחליטים האם הוא שימושי דיו על לביצוע תחזיות, אם לאו. בפרקטיקה, מדעני נתונים (קרי, אקטוארים ומנהלי סיכונים) מפתחים מספר מודלים ומשווים את התוצאות (ממש כפי שאנו עשינו במאמר זה). לפיכך, הפרקטיקה המיטבית הינה לחלק את הנתונים הזמינים לשלושה סטים: (1) סט אימון; (2) סט תיקוף; ו- (3) סט בדיקה.

סט האימון משמש לפיתוח מודלים אלטרנטיביים, בעוד שסט התיקוף נועד פעם אחת לחקר באיזו מידה המודלים שפותחו באמצעות סט האימון מכלילים נתונים חדשים ופעם שניה

לבחור מבין המודלים הללו את המודל שמכליל בצורה הטובה ביותר נתונים חדשים. סט הבדיקה "נשמר בצד" כגיבוי ומשמש בתור נתוני מדגם אחר לצורך בדיקת מידת הדיוק של המודל הנבחר בתום התהליך.

חלוקה טיפוסית של הנתונים הינה כדלקמן: 60% לסט האימון, 30% לסט התיקוף ו-10% לסט הבדיקה. עם זאת, זה תלוי כמובן במודלים ובכמות הנתונים הזמינים.

חשוב ביותר להדגיש שסט האימון שעשינו בו שימוש בדוגמא זו איננו מייצג נורמטיבי לסט אימון בלמידת מכונה (לכל בר בי רב ברור הלא שמעשר תצפיות לא ממש ניתן ללמוד בצורה מהימנה על קשר מסוים). על כן מטרת מאמר זה הייתה להמחיש את בעיית התאמת היתר/התאמת החסר באמצעות דוגמא פשוטה למדי.

פרטים אודות כותב המאמר: האקטואר רועי פולניצר, FRM

רועי בעל תואר שני במימון (התמחות בניהול סיכונים ואקטואריה) ותואר ראשון בכלכלה (התמחות במימון), שניהם מאוניברסיטת בן-גוריון בנגב, בעל דיפלומה בניהול סיכונים פיננסיים (FRM®) מאוניברסיטת אריאל בשומרון ולמד בתוכנית ללימודי תעודה באקטואריה באוניברסיטת חיפה. כמו כן, רועי אקטואר מלא



(Fellow) בלשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (F.I.L.A.V.F.A.), מוסמך כמעריך שווי מימון תאגידי (CFV) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA), מוסמך כמנהל סיכונים פיננסיים (FRM) מטעם האיגוד העולמי למומחי סיכונים (GARP) ומוסמך כמומחה לניהול סיכונים (CRM) מטעם האיגוד הישראלי למנהלי סיכונים (IARM).



לרועי ניסיון של מעל ל- 15 שנה בביצוע ניתוחים כמותיים במכשירים פיננסיים, בהערכת שווי תאגידים ונכסים בלתי מוחשיים, באמידה וכימות סיכונים כמו תמותה, אריכות ימים, תחלואה, ביטולים והחלמה מנכות, ובמידול ומדידת סיכוני שוק, אשראי, תפעוליים, מודל, מזילות והשקעות לצורכי יישום הוראות רגולטוריות ותקינה חשבונאית, פיתוח, יישום ותיקוף מודלים בתחומים של הערכות שווי, ניהול סיכונים, אקטואריה והנדסה פיננסית, קביעת תעריפי ביטוח חיים, הערכת פרמיות סיכון והערכת עתודות ביטוח, קביעת עלות תנאי פנסיות (צוברות ותקציביות) והכנת מאזנים אקטואריים לקרנות פנסיה, ניתוח וחיזוי מצבים פיננסיים מורכבים וכן העברת סמינרי הדרכה והשתלמויות בתחומי התמחות: מימון, אקטואריה, הערכות שווי, בנקאות, ניהול סיכונים, אופציות והנדסה פיננסית.

ניסיונו של רועי בתחום ה- Data Analysis, כולל: עבודה עם מאגרי מידע גדולים Big Data תוך שימוש ב- Statistical Learning (כגון: סטטיסטיקה תיאורית, הסתברות, הסקה סטטיסטית, סטטיסטיקה א-פרמטרית, חלוקת נתונים, נרמול נתונים, Fitting ו- Bayes Theorem) ובאלגוריתמים מסוג Unsupervised Learning (כגון: Hierarchical Clustering, k-means Clustering, Density-based Clustering, Distribution-based Clustering ו- Principle Components Analysis) למציאת דפוסים וזיהוי מגמות ואנומליות בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, פיתוח תשתית לצורך ניתוח נתונים, שילוב והטמעת כלים לצורך גישה ושליפה עצמאית של נתונים ממאגרי מידע, פיתוח דוחות, ממשקים ומסכים באמצעות כלי ויזואליזציה.

ניסיונו של רועי בתחום ה- Data Science, כולל: עבודה עם מסדי נתונים גדולים Big Data תוך שימוש באלגוריתמים מסוג Supervised Learning (כגון: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, Logistic Regression, Maximum Likelihood Estimation, k-Nearest Neighbors, Decision Tree, Random Forest, Ensemble, Bagging,



Boosting, Naïve Bayes Classifier, Linear Separation, Support Vector Machine, Non-Linear Separation, SVM Regression, Artificial Neural Network, Convolutional Neural Network (Recurrent Neural Network) לניבוי וסיווג בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה ובמודלים מסוג Reinforcement Learning (כגון: Q-learning, Monte Carlo Simulation, Temporal Difference Learning ו-n-Step Bootstrapping) לקבלת החלטות מרובות שלבים בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, זיהוי אתגרים עסקיים שבהם DATA יכול להוות גורם מכריע בשיפור קבלת החלטות, איתור ואיסוף מקורות מידע, הגדרה ואיפיון של שימושי המידע, בניית מסד המידע, אפיון והגדרת הצגת המידע ותוצריו, פיתוח כלים, מודלים, תהליכים ומערכות בתחום האנליזה, תוך שימוש בכלי אנליזה מתקדמים (EXCEL, VBA ו-R).