# למידת מכונה ב-MATLAB - המדריך למשתמש המתחיל

## תוכן עניינים

**מבוא..... 2**

**סוגים של למידת מכונה..... 2**

**מאיפה מתחילים? ..... 3**

חיפוש אחר נתונים..... 3

עיבוד מקדים של הנתונים..... 4

מיצוי מאפיינים מהנתונים ..... 4

אימון המודל ..... 4

בחינת המודל..... 7

הטמעת המודל..... 9

**איך ממשיכים מכאן?..... 9**



סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il

## מבוא

איך הדואר האלקטרוני שלנו יודע "בעצמו" לסווג הודעות כחשובות/לא חשובות **('ספאם')** או כיצד המוצר שאנחנו בדיוק מחפשים מופיע לפתע ב**פרסומת** ברשתות החברתיות? ואיך **נטפליקס** יודע מה הטעם שלנו בסדרות?

התשובה, בשתי מילים – **בינה מלאכותית**.

תת תחום בעולם הבינה המלאכותית הוא למידת מכונה (Machine Learning). למידת מכונה היא שם כולל ליכולת של המחשב שלנו ללמוד מתוך סט של דוגמאות. האלגוריתם של למידת מכונה משתמש בשיטות חישוביות כדי "ללמוד" מידע ישירות מנתונים מבלי להסתמך על משוואה שנקבעה מראש. האלגוריתמים משפרים באופן אדפטיבי את ביצועיהם ככל שמספר הדגימות הזמינות ללימוד גדל.

אם תחום למידת המכונה מעניין אתכם אבל אתם צריכים קצה חוט שיכוון אתכם איך להתחיל בצורה פרקטית –הגעתם למקום הנכון!

## סוגים של למידת מכונה

נחלק את הבעיות שקיימות בעולמנו לשתי קטגוריות: למידה מונחית (Supervised Learning) ולמידה בלתי מונחית (Unsupervised Learning).

**למידה מונחית:** מתבססת על נתוני קלט ופלט ידועים (מידע מתוייג). מטרת למידה זו היא לסווג תצפיות חדשות על בסיס נתוני הקלט והפלט.

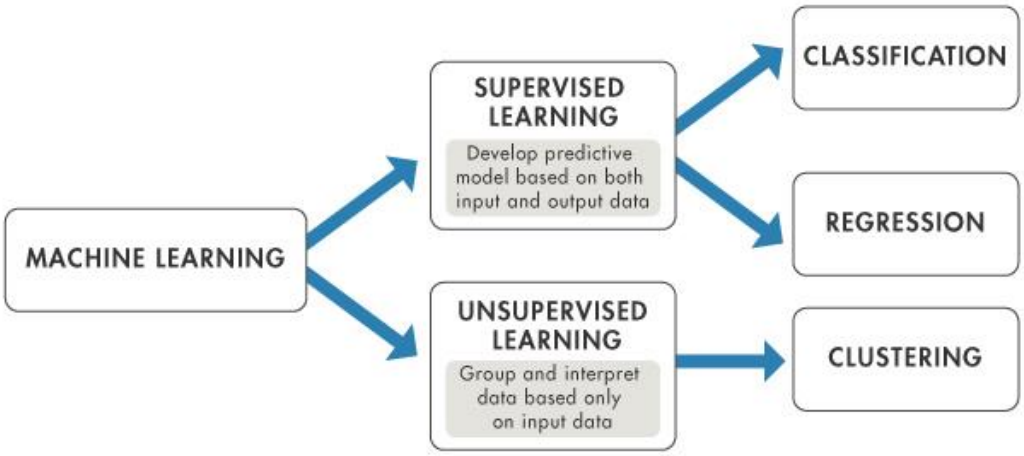
אלגוריתם למידה מונחית לוקח קבוצה ידועה של נתוני קלט ותיוגים מתאימים לנתונים (פלט) ומאמן מודל ליצירת תחזיות לנתונים חדשים. למידה זו משתמשת בטכניקות סיווג ורגרסיה כדי לפתח מודלי חיזוי.

- טכניקת **סיווג** (Classification) מסווגת ומחלקת את נתוני הקלט לקטגוריות –האם דוא"ל אמין או "ספאם", האם גידול סרטני הוא ממאיר או שפיר, מידע נוסף בקישור הבא.
- טכניקות **רגרסיה** (Regression) מנבאות תגובות רציפות לדוגמה - שינויים בטמפרטורה, מידע נוסף בקישור הבא.

**למידה בלתי מונחית:** כאן המידע לא מתוייג, מטרת האלגוריתם היא להסיק מסקנות לגבי הנתונים.

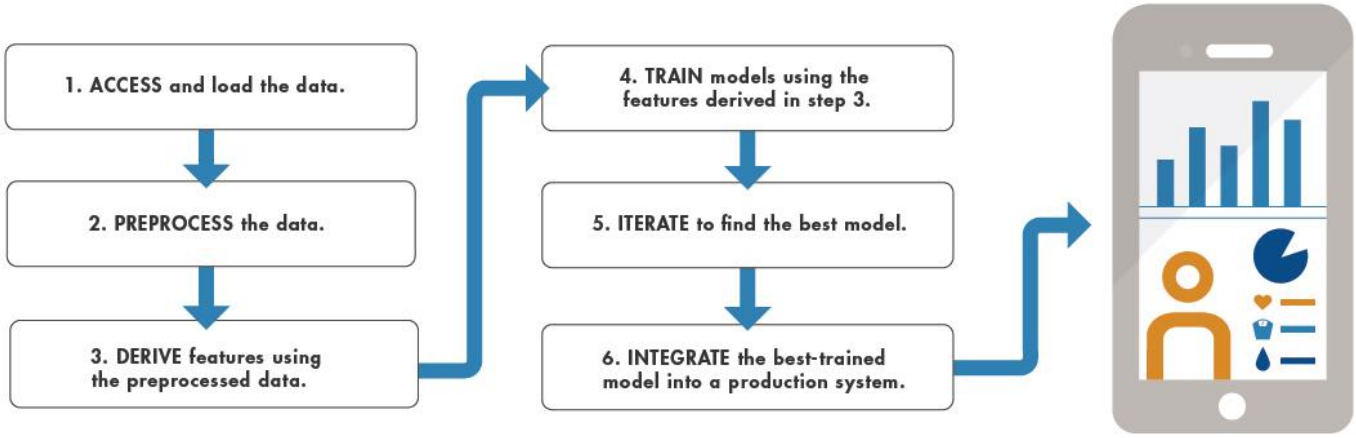
- טכניקת **אשכול** (Clustering) היא טכניקת הלמידה הנפוצה ביותר כאשר מדובר על למידה בלתי מונחית. בטכניקה זו, נרצה למצוא דפוסים נסתרים או מבנים מהותיים בנתונים. היישומים לאשכול כוללים לדוגמה ניתוח רצפי גנים וזיהוי אובייקטים, מידע נוסף בקישור הבא.


סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il




## מאיפה מתחילים?

נפרק את התרשים הבא שלב אחר שלב על מנת להבין איך להתחיל לבצע למידת מכונה ב-MATLAB.




### חיפוש אחר נתונים:

השלב הראשון כולל חיפוש אחר נתונים. למתחילים שביננו, אני ממליצה להוריד נתונים מאתרי האינטרנט השונים. ישנם אתרים רבים שמהם אנחנו יכולים להוריד נתונים במגוון רחב של תחומים:

1. Kaggle - אתר המכיל המון נתונים במגוון רחב של תחומים. עבור חלק מהנתונים ישנן מחברות בהן אנשים מתעדים את האלגוריתם שלהם ואת הדרך שלהם לפתרון. כך ניתן לעקוב ולהתחקות אחר אלגוריתמים בשלבי הלמידה הראשונים שלכם.
2. Google DataSet Search – מנוע חיפוש של גוגל, מאגר נתונים ענק המכיל בתוכו למעלה מ-25 מיליון מאגרי נתונים.
3. VisualData – אם תחום העיסוק שלכם הוא ראיה ממוחשבת, תהנו באתר זה מסט נתונים רחב של תמונות.



סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il

לאחר שבחרתם סט נתונים, הסתכלתם עליו ווידאתם שאתם מבינים אותו זה הזמן לטעון אותו ל-MATLAB. במידה ואינכם יודעים איך לטעון נתונים לתוך ה-MATLAB אתם מוזמנים לצפות בווביינר שהכנתי בנושא זה.

## עיבוד מקדים של הנתונים:

השלב השני כולל עיבוד מקדים של הנתונים וזאת על מנת להבטיח ניתוח מדויק, יעיל ומשמעותי. ניקוי נתונים מתייחס לשיטות למציאה, הסרה או החלפת נתונים חריגים או חסרים, איתור מקסימום מקומי, איתור שינויים פתאומיים היכולים לעזור בזיהוי מגמות משמעותיות והחלקת נתונים על מנת להסיר רעש. מידע נוסף על כל הפונקציות הרלוונטיות לתהליך העיבוד המקדים, נמצא בקישור הבא.
מוזמנים גם לצפות בווביינרים שלנו בנושא עיבוד מקדים – בקישור הבא ובנושא התמודדות עם נתונים חסרים, בקישור הבא.

## מיצוי מאפיינים מהנתונים:

השלב השלישי כולל מיצוי מאפיינים מהנתונים לאחר שלב העיבוד. בשלב זה נרצה למצוא תבניות בנתונים שלנו על מנת להצליח לסווג אותם בצורה הטובה ביותר. הבנה טובה של הנתונים תקל עליכם בבחירת מאפיינים משמעותיים.
מהם המאפיינים בהם אנשים מרבים להשתמש? הרבה משתמשים במאפיינים מתחום הסטטיסטיקה כמו: מינימום, חציון והערך השכיח וזאת על מנת למצוא מגמות שיסווגו את הנתונים בצורה טובה.

בהקשר של תמונות, מאפיינים נפוצים יהיו – היסטוגרמה של צבעים או פילוג של נגזרות בכיוונים שונים.

אציין כי הדרך היחידה לדעת אם מאפיין נחשב לאיכותי, היא להוסיף אותו למודל ולבדוק אם הוא משפר את התוצאות.

כמובן שעבור סוג קלט שונה (אותות, תמונות, טקסט) המאפיינים יהיו שונים בתכלית. לכן ריכזתי לכם תיעוד של מציאת מאפיינים עבור כל אחד מסוגי הקלטים:

- מיצוי מאפיינים עבור אותות – בקישור הבא
- מיצוי מאפיינים עבור תמונות – בקישור הבא
- מיצוי מאפיינים עבור טקסט – בקישור הבא
- מיצוי מאפיינים מעולם הסטטיסטיקה – בקישור הבא

דוגמאות לבחירת מאפיינים עבור אותות, תוכלו למצוא בווביינר שלנו על מיצוי מאפיינים מאותות ב-MATLAB, בקישור הבא.

## אימון המודל:

השלב הרביעי הוא אימון המודל שלנו, המאפיינים שמיצינו בשלב השלישי מהווים את הקלט למודל. כאשר אנו עוסקים בלמידה מונחית, נחלק את הנתונים שלנו ל-2 קבוצות: סט לאימון וסט ולידציה. על סט האימון נאמן את המודל שלנו ועל סט הולידציה נבחן את דיוק המודל.

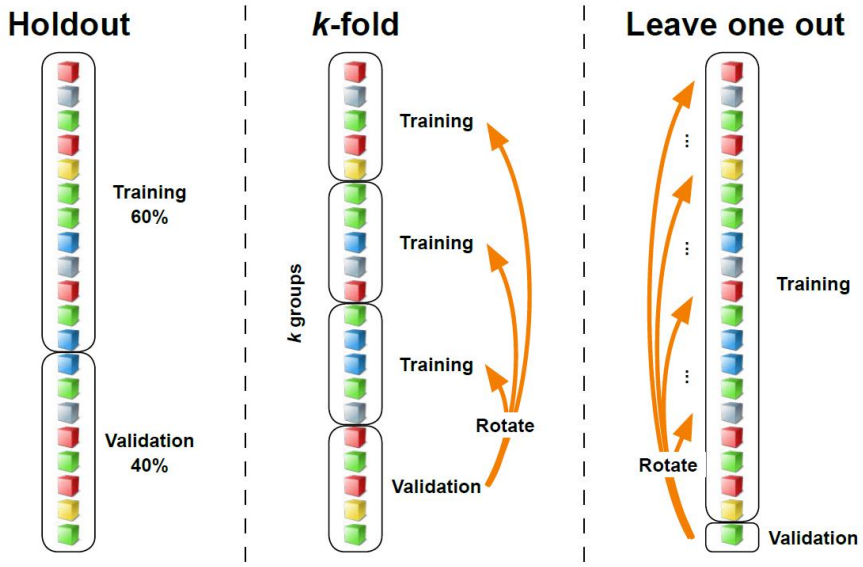

סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il

ישנן שלוש טכניקות לביצוע החלוקה:

1. **Holdout –** במידה ויש לכם סט נתונים גדול, תוכלו להשתמש בשיטה זו שמחלקת את הנתונים באופן אקראי לשניים, בהתאם לאחוזים שתגדירו מראש. בדרך כלל, סט האימון גדול יותר מסט הולידציה.

עבור סט נתונים מצומצם מומלץ להשתמש באחת משתי השיטות הבאות:

2. **K-fold –** מחלק את סט הנתונים ל-K תתי דגימות הנבחרות בצורה אקראית. מקבץ אחד משמש לולידציה כשהאחרים לאימון. חוזרים על המודל כך שקבוצת הולידציה משתנה עבור כל איטרציה.
3. **Leave one out –** מקרה מיוחד של K-fold , בו מספר המקבצים שווה למספר התצפיות, כאשר תצפית יחידה משמשת לולידציה עבור כל איטרציה.

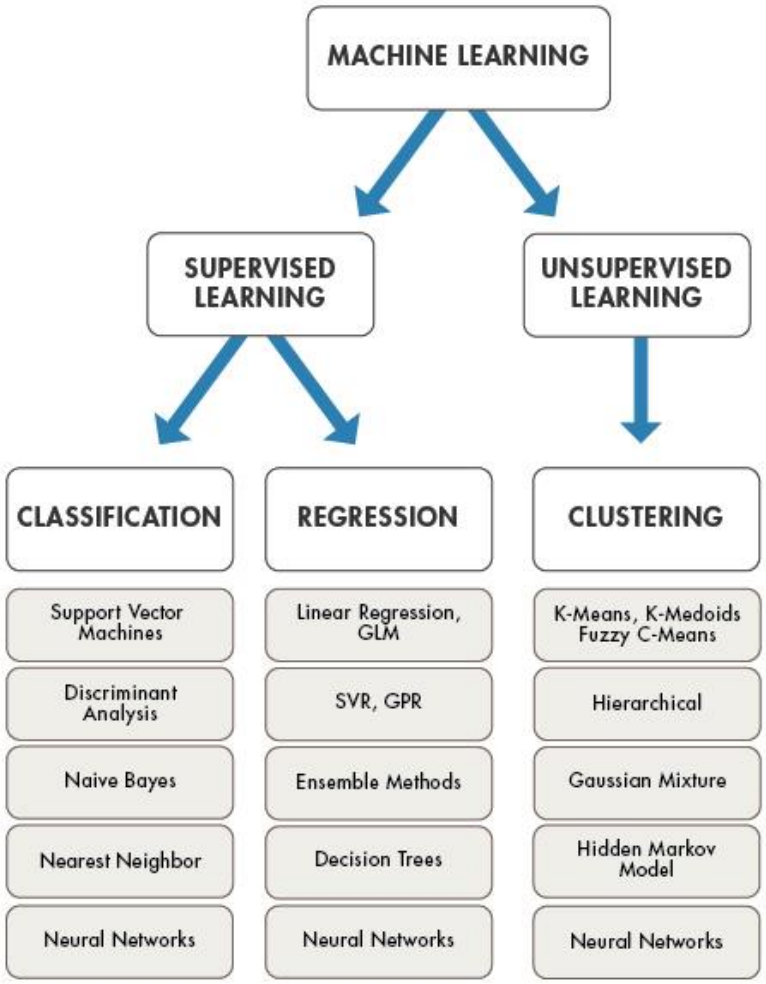



החלוקה תיעשה ב- MATLAB על ידי פונקציה cvpartition, כאשר בעזרתה נוכל להגדיר באיזה שיטה נרצה לבצע את החלוקה (Cross Validation).

אחרי שהבנו איך אנחנו מבצעים את החלוקה לסט אימון וסט ולידציה ננסה להבין מהו האלגוריתם שיניב לנו את התוצאות הטובות ביותר. בחירת האלגוריתם זהו שלב מכריע בתהליך למידת המכונה. ישנם עשרות אלגוריתמים עבור למידת מכונה (מונחית ובלתי מונחית) וכל אחד מהם נוקט בגישה שונה ללמידה.

להלן תרשים של האלגוריתמים השונים עבור למידה מונחית ובלתי מונחית:


סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il




עבור כל אלגוריתם קיימת פונקציית אימון המתאימה לו. פונקציות האימון של רגרסיה ב-MATLAB מתחילות ב-fitr ושל קלסיפיקציה מתחילות ב-fitc כאשר הסיומת מתאימה לשם המודל. לדוגמה fit**nb** מתייחס לאימון **N**aive **B**ayes model.

- עבור מודלים בתחום הלמידה המונחית, תוכלו למצוא את הפונקציות המתאימות בקישור הבא, תחת הכותרת Fit a Model.
- עבור מודלים בתחום הלמידה הבלתי מונחית, תוכלו למצוא מידע אודות המודלים השונים בקישור הבא.

נניח וניסינו מודל מסוים, איך נדע שזה המודל האופטימלי עבורנו? זה מביא אותנו לשלב החמישי!

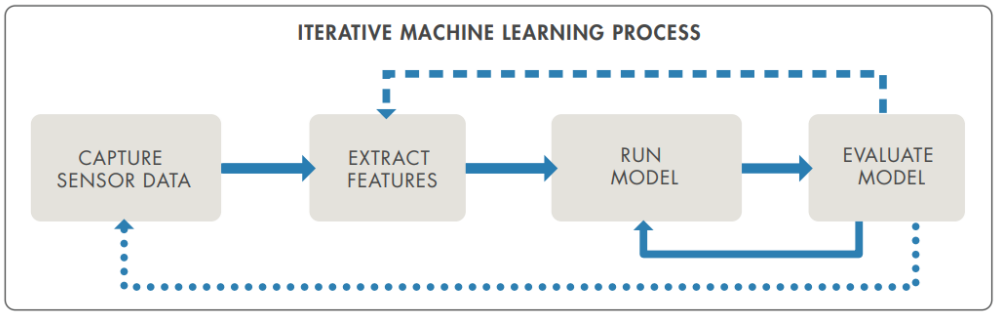


סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il

## בחינת המודל :

השלב החמישי מתייחס לבחינת המודל והרצת איטרציות עבור מודלים שונים וזאת על מנת לבחור את המודל המתאים עבור הנתונים שלכם.
בחינת המודל תיעשה על ידי הרצת קבוצת הולידציה שיצרנו קודם לכן וזאת על מנת לבחון את הצלחת המודל, כלומר האם הוא מצליח לעשות קלסיפיקציה או רגרסיה עבור נתונים "חדשים". כאן אנחנו ננסה להעריך האם המודל שלנו טוב על ידי השגיאה המתקבלת.

הבחינה תיעשה על ידי הרצת איטרציות ומציאת האלגוריתם שמניב את התוצאות הטובות ביותר עבורנו, שלב זה מבוסס בחלקו על ניסוי וטעייה.

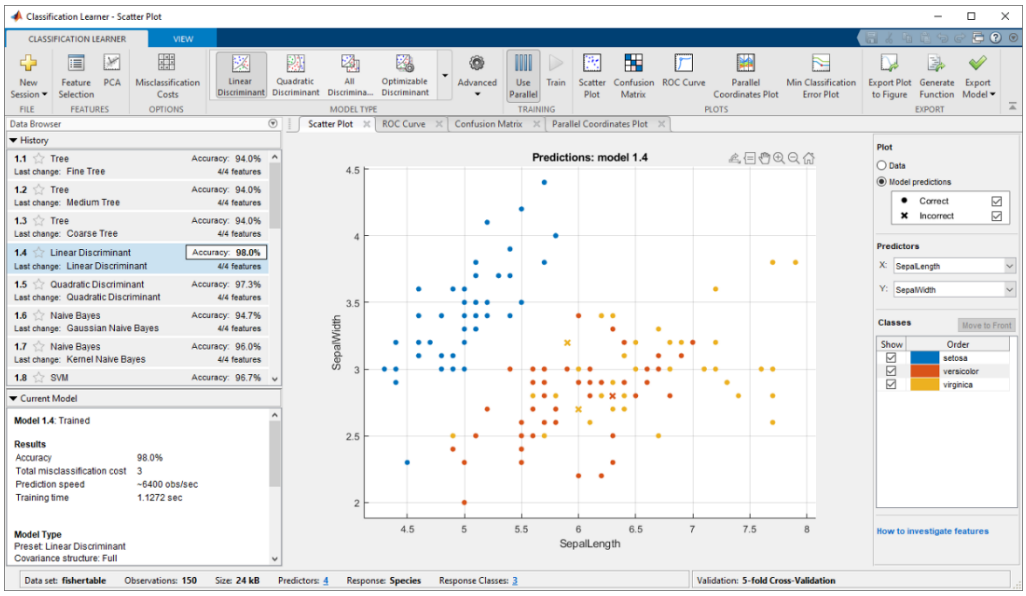



**MATLAB מספק כלים שיעזרו לכם להתנסות באלגוריתמים שונים בקלות רבה וכך להתנסות ולבחור את האלגוריתם הטוב ביותר לכם.**

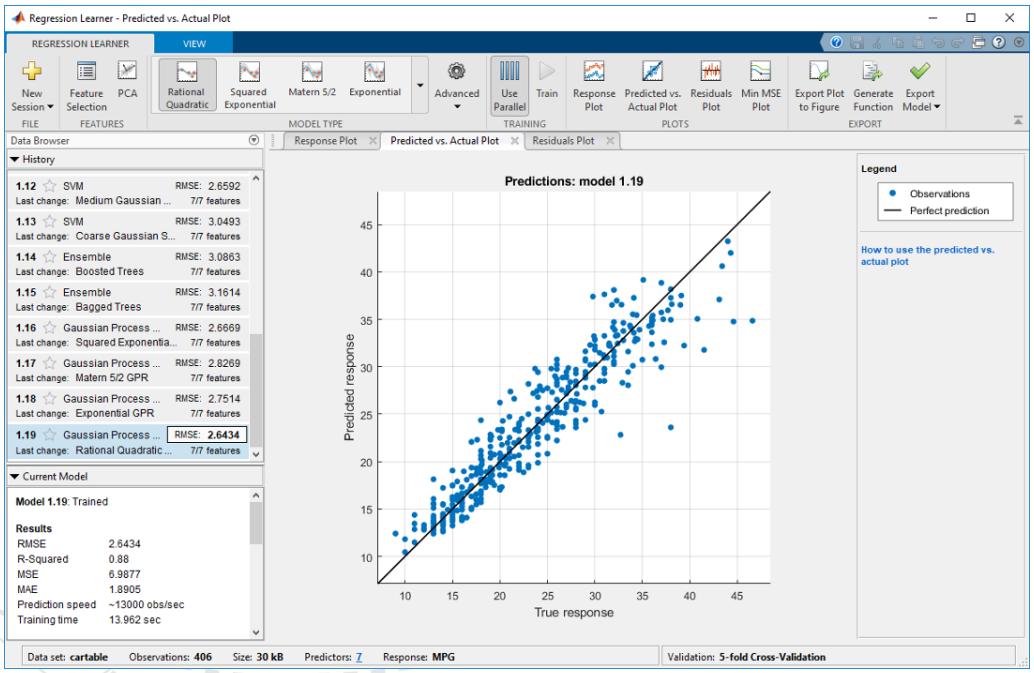
ב- **Statistics and Machine Learning Toolbox** נמצאות בין היתר שתי אפליקציות שימושיות היכולות "לעשות לכם חיים קלים" בכל הנוגע ללמידה מונחית. בעזרת האפליקציות, תוכלו לבחור תכונות עבור הנתונים שלכם, לייצר סכימות ושרטוטים שיעזור לכם לנתח את התוצאות, לאמן את המודל ולהעריך את התוצאות המתקבלות. לאחר השוואה בין המודלים השונים (כאשר הכל נעשה על ידי לחיצת כפתור), ניתן לייצא את המודל הטוב ביותר לתוך סביבת ה- MATLAB או לייצר קוד MATLAB שיבצע אוטומציה לתהליך בניית המודל. ייצוא המודל בתור קוד מומלץ במיוחד לאלו מכם שנמצאים בתחילת התהליך ויכולים ללמוד באמצעות האפליקציה על הפונקציות השונות שעומדות מאחורי האלגוריתמים.


סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il

1. **Classification Learner App –** אפליקציה אינטראקטיבית המתאימה מודל סיווג לנתונים. בעזרת האפליקציה נוכל לבצע אימונים אוטומטיים לחיפוש אחר מודל הסיווג הטוב ביותר. מידע על האפליקציה, ניתן למצוא בקישור הבא.




2. **Regression Learner App –** אפליקציה אינטראקטיבית המבצעת אימון עבור מודלים של רגרסיה. נוכל לבחור בין אלגוריתמים שונים, להשוות בין השגיאה המתקבלת עבור כל אחד מהמודלים וכך נוכל לבחור את המודל הטוב ביותר. מידע על האפליקציה, ניתן למצוא בקישור הבא.




אחרי שבחרנו את המודל הטוב ביותר עבור הנתונים שלנו אפשר להגיד שאנחנו כמעט סיימנו.



סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il

**הטמעת המודל** :

השלב האחרון זהו שלב הטמעת המודל. באמצעות כלי ה- Deployment ש-MATLAB מציעה אפשר לפרוס את האלגוריתם שלנו על מגוון רחב של פלטפורמות:

1. MATLAB Compiler
2. MATLAB Compiler SDK
3. MATLAB Production Server
4. MATLAB Coder
5. GPU Coder
6. HDL Coder

## איך ממשיכים מכאן?

אני מצרפת מספר קישורים שימושים לאלה מכם שרוצים להשלים קצת ידע תיאורטי או להתנסות בצורה מעשית:

1. קורס מומלץ מאוד ב- **Coursera** המספק מידע נרחב אודות למידת מכונה – בקישור הבא.
2. **Machine Learning Onramp**, קורס מעשי **חינמי** שאורכו שעתיים, בו תתנסו בכתיבת קוד של למידת מכונה באמצעות ה-MATLAB - בקישור הבא.
3. ולאלה מכם שרוצים קורס יותר מעמיק בנושא, מוזמנים להירשם לקורסים במרכז ההדרכה של סיסטמטיקס לפרטים נוספים היכנסו לקישור הבא.
4. להורדת גרסת ניסיון היכנסו לקישור הבא. מדריך להורדת גרסת ניסיון תוכלו למצוא בסרטון הבא.
5. לשאלות נוספות לרשותכם מוקד התמיכה הטכנית של סיסטמטיקס.

בהצלחה!



סיסטמטיקס בע"מ
רחוב ראול ולנברג 4, תל אביב 69719
טלפון: 03-7660111
פקס: 03-7660120
מוקד שירות: 03-7660101
דוא"ל: info@systematics.co.il
www.systematics.co.il