**פרק 2 : המודל הראשון שלי**

פרק זה ירגיש לחלקכם כמו hello world ל-machine learning, וכך הוא צריך להרגיש. ניקח בתור התחלה מודל פשוט (ומעצם כך ניתן להגיד עליו דברים יפים אנליטית שלא ניתן להגיד על מודלים יותר מורכבים בכזו אלגנטיות) וננסה להשתמש בו על בעיה אמיתית.

ההתחככות הזו של להפעיל מודל במציאות מעוררת הרבה שאלות, כיצד נכון לבחון את המודל, איך משווים מודלים שונים, ואילו בורות מחכים לי - איש הדאטא סיינס המתחיל.

**פרק זה יתעסק בבעיות רגרסיה - שהן אולי הצד היותר אינטואיטיבי להגדרת loss והבנת הframework הסטנדרטי.**

**חלק תיאורטי:**

1. קרא את פרק 3 בISLR - רגרסיה לינארית - Linear Regression.
2. קרא את פרק 5 בISLR - שיטות חלוקת דאטא - Resampling Methods.
3. קרא את פרקים 6.1 ו6.2 בISLR - השוואת מודלים - Linear Model Selection.

**חלק מעשי:**

בקהילה של הדאטא סיינס יש מספר דאטאסטים מוכרים אשר מהווים בנצ'מארק לכלל הקהילה. מאגר המידע על מחירי שכונות בבוסטון הוא אחד מהם.

המאגר מכיל לא הרבה רשומות, על שכונות בבוסטון ומחירי הדיור הממוצע בהם. על כל שכונה ישנם מספר פריטי מידע אשר יכולים לעזור לחזות את המחיר הממוצע בשכונה.

המאגר הוא בין הדאטאסטים הראשונים שנוצרו בתחום, הוא אינו גדול ומשום כך שאלת התאמת היתר עולה בצורה יותר חדה.

היעזר בsklearn על מנת לטעון את המידע למחברת.

**לאחר מכן בצע את שלבי המחקר הקלאסי (כמובן שזו רק שבלונה, ויש לעשות איטרציות תכופות עם החונך בשלב זה, ולחזור על שלבים קודמיםלאחר הערות):**

1. בצע אקספלורציה: לדוגמא -
   1. בחן התפלגויות וקורלציות
   2. וודא שאתה מבין את משמעות העמודות - מתיאורן ומהערכים שהם מכילים
2. נקה את המידע: לדוגמא -
   1. השלם ערכים חסרים
   2. הורד נקודות מידע אשר מהווים אנומליות (לשלב הזה לא מדובר בהפעלת אלגוריתמי anomaly detection).
3. הרץ את המודל: לדוגמא -
   1. בניית מודל על כלל המידע.
   2. בחינת המודל הנוצר (משקולות וכו)
   3. בחינת טעויות או סטיות גדולות מהערך הרצוי.
   4. בחינת אימון עם ובלי פיצ'רים קיימים.
   5. חילוץ מאפיינים משמעותיים ואימון מחודש לשיפור המדד.