עבודת גמר בקורס גישות חישוביות במדעי המוח

1. מאגר הנתונים נבחר מתוך האתר UCI machine learning repository .   
   הנתונים מכילים פרטים אודות פוסטים שפורסמו במהלך שנת 2014 בדפי הפייסבוק של חברות מותגי קוסמטיקה מפורסמים.   
   סט הנתונים מכיל 500 דוגמאות.   
   תכונות (פיצ'רים): 7 פיצ'רים ידועים לפני פרסום הפוסט, ו 12 נאספו לאחר הפרסום.  
   הנתונים כללו 4 סוגים: זיהוי הפוסט, סיווג פוסט, תוכן, ביצועים.   
   ואלו הם הפיצ'רים:

* **Page total likes**  - כמות "לייקים" לעמוד, משתנה מספרי בין 81,370 ל 139441
* **Type** – סוג ה"פוסט", משתנה קטגוריאלי, מכיל אחת מהאפשרויות: Link, Photo, Status, Video (קישור, תמונה, סטטוס, ווידיאו).
* **Category –** קטגוריה של הפוסט, משתנה קטגוריאלי, מכיל אחת האפשרויות: action, product, inspiration   
  (הצעות רכישה, פרסום המוצר, תוכן שלא ניתן לקשר מפורשות לעמוד).  
  action = 1, product = 2, inspiration = 3
* **Post Month** – חודש פרסום הפוסט, משתנה קטגוריאלי מספרי 1-12.
* **Post Weekday** – יום בשבוע בו פורסם הפוסט, משתנה קטגוריאלי מספרי 1-7 (ראשון-שבת).
* **Post Hour** – שעת פרסום הפוסט, משתנה קטגוריאלי מספרי 1-23 (שעות היממה).
* **Paid** – האם שולם, משתנה קטגוריאלי בינארי : 0 לא שולם, 1 שולם.
* **Lifetime Post Total Reach** - כמות האנשים שראו את הפוסט (unique user), משתנה מספרי בין 238 ל 180480.
* **Lifetime Post Total Impressions** – כמות החשיפות לפוסט. משתנה מספרי, בין 570 ל 1110282
* **Lifetime Engaged Users –** כמות האנשים שלחצו על הפוסט במקום כלשהו (unique user), משתנה מספרי בין 9 ל 11452 .
* **~~Lifetime Post Consumers~~** ~~– כמות האנשים שלחצו על הפוסט במקום מסוים. משתנה מספרי בין 9 ל 11328~~
* **~~Lifetime Post Consumptions~~** ~~– כמות לחיצות שנעשו בפוסט במקום כלשהו. משתנה מספרי בין 9 ל 19779 .~~
* **~~Lifetime Post Impressions by people who have liked your Page~~** ~~– כמות החשיפות לפוסט מתוך משתמשים שלחצו לייק על העמוד. משתנה מספרי בין 567 ל 1107833~~
* **Lifetime Post reach by people who like your Page** - כמות החשיפות לפוסט מתוך משתמשים שלחצו לייק על העמוד (unique user), משתנה מספרי בין 236 ל 51456.
* **~~Lifetime People who have liked your Page and engaged with your post~~** ~~– כמות אנשים שלחצו במקום כלשהו על הפוסט וגם עשו לייק לעמוד (unique user), משתנה מספרי בין 9 ל 4376~~
* **Comment** – כמות תגובות על הפוסט, משתנה מספרי בין 0 ל 372
* **Like** – כמות לייקים לפוסט, משתנה מספרי בין 0 ל 5172
* **Share –** כמות השיתופים שעשו לפוסט, משתנה מספרי בין 0 ל 790
* **~~Total Interactions~~** ~~– משתנה מחושב מספרי שמכיל סכום של כמות תגובות + כמות לייקים + כמות שיתופים, משתנה מספרי בין 0 ל 6334~~

**\*\*** unique user – הכוונה ל session, כלומר אם משתמש נכנס פעמיים מכתובות IP אחרות (לדוגמה מהטלפון ומהמחשב, על אף שזהו אותו משתמש פייסבוק, ייספרו כשתי כניסות שונות).

המשתנה שנרצה לנבא הוא: כמות לייקים עבור פוסט.  
  
המשתנים שנראים לנו הכי רלוונטיים:   
  
לדעתנו, האתגר/מכשול המרכזי במשימה שנבחרה הוא :   
העבודה מול הדאטא: כיוון שנדרשנו לבעיית קלסיפיקציה, חילקנו את המשתנה שנרצה לחזות על פי טווחים, קיימות אינספור חלוקות, ואין דרך לדעת מהי החלוקה שהניבוי עבורה יצליח בצורה הטובה ביותר.  
בנוסף, בחירת התכונות שאיתן נעבוד בבניית המודל – גם אלו נעשו על פי שיקול דעת אישי שלנו.  
בפרויקטים בתחום למידת המכונה נהוג לנסות מודלים שונים, שימוש בפיצ'רים שונים וטווחים שונים עד שיימצא השילוב הטוב ביותר, בפרויקט הנוכחי הסתמכנו על בחירה אחת בכל נושא (לפי ההוראות ועצת המרצה).   
  
הורדנו את דוגמה 113, כיוון שלא היה עליה את הנתון של הניבוי like   
ולדוגמה האחרונה שחסר הנתון של paid שמנו מה שיש לרוב

1. בחלק זה בקוד, בפונקציה \_\_\_ טענו את הנתונים מהקובץ, ביצענו רנדומיזציה של סט הנתונים (על מנת שלא יהיה bias שנובע מסדר הופעתם בקובץ). לאחר מכן, סט הנתונים חולק ל:  
    80% - training set, 20% - test set.   
   הנתונים יישמרו לצורך נוחות בשני קבצים נפרדים data.train, data.test .