לימוד מכונה 364-1-1811

**פרויקט - חלק א'**

זהו החלק הראשון מבין שני חלקים.

**הנחיות הגשה:** **דו"ח התרגיל הראשון וקוד ה- Python שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה-2.12.21 בשעה 23:55. מספיקה הגשה של אחד מבני הזוג.**

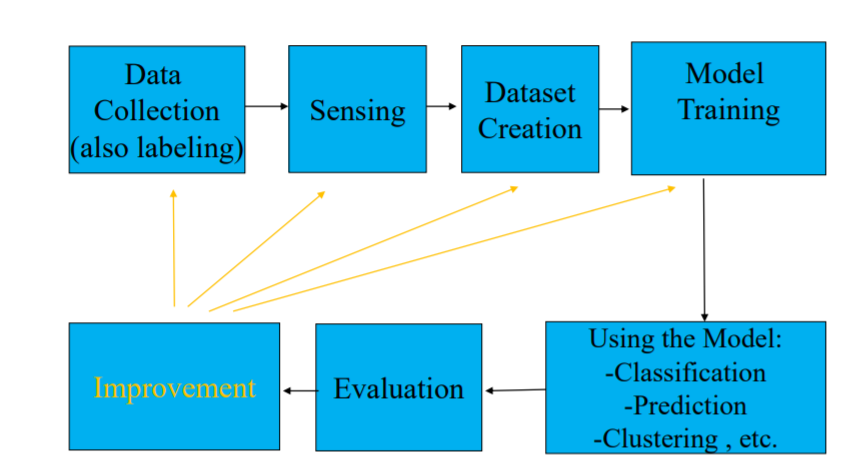
מטרת התרגיל: בתרגיל זה נתרגל את שלושת השלבים הראשונים בתהליך יצירת מכונה לומדת. למעשה השלב הראשון והשני (Data Collection, Sensing) כבר ניתן לכם, ולכן מרבית העבודה תהיה בשלב השלישי (Dataset Creation). שלבים אלה ישמשו אותנו בהבנת והכנת הנתונים לקראת התרגיל השני בו נשתמש בנתונים לאימון ובחינה של מערכות לומדות והשוואה ביניהן.

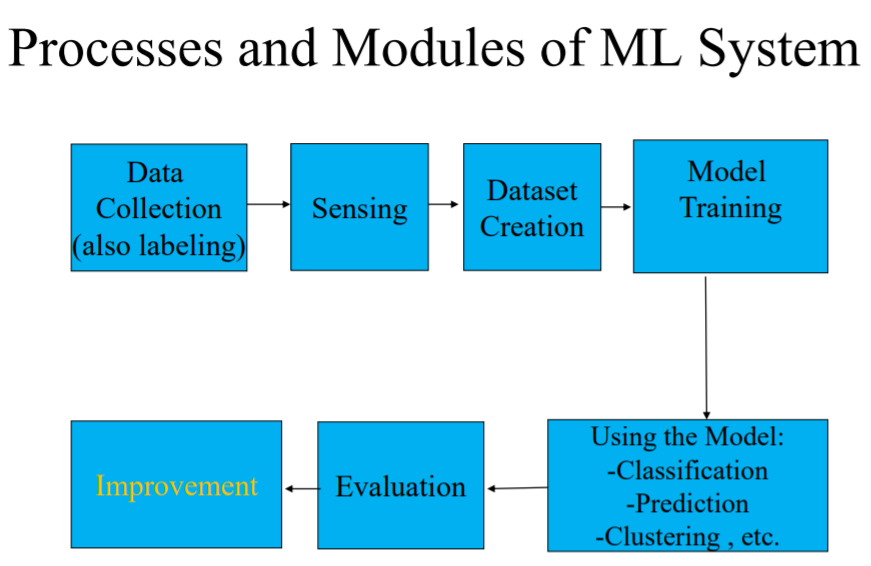
צוותי הגשה: הגשת התרגיל הינה בזוגות, בהתאם לקבוצות המוגדרות במודל.

דגשים לדו"ח: **אורך הדו"ח לא יעלה על 10 עמודים** (לא כלל עמודים נלווים כמו שער ותוכן עניינים), בגודל כתב 12, פונט Arial ורווח של שורה וחצי. **חריגה מהגדרות אלו תגרור הורדת נק'**. יש לשמור על תמציתיות ולהתמקד בתובנות המרכזיות שלכם בכל סעיף. יש להגיש קובץ Word המאפשר השארת הערות על גביו, אין להגיש קובץ PDF!

שפת תכנות: Python.

שאלותבנוגע לתרגיל, יש לפרסם בפורום הייעודי שייפתח במודל בלבד, פניות למייל לא יענו!





איור 1: תזכורת מההרצאה - שלבי פרויקט machine learning

# הסבר אודות בסיס הנתונים לפרויקט

את קבצי הנתונים של הפרויקט ניתן למצוא ב Moodle.

בסיס הנתונים מכיל מידע אודות אנשים הנמצאים בתכנית הכשרה לDS בחברה כלשהי.

מטרתכם תהיה לחזות מי מבין האנשים ירצה להמשיך לעבוד בחברה ומי יצא לחפש עבודה במקום אחר בסיום התכנית.

X\_test.csv – משמש לטובת הגשת החיזוי הסופי. מכיל את המשתנים מסבירים בלבד של סט הבחינה. קובץ זה ישמש אותכם בחלק ב׳.

XY\_train -קובץ המשתנים המסבירים והמשתנה המוסבר עבור האימון (עליו עובדים בחלק זה)

# הסבר על המשתנים בבסיס הנתונים

* enrollee\_id : Unique ID for candidate
* city: City code
* city\_ development \_index : Developement index of the city (scaled)
* gender: Gender of candidate
* relevent\_experience: Relevant experience of candidate
* enrolled\_university: Type of University course enrolled if any
* education\_level: Education level of candidate
* major\_discipline :Education major discipline of candidate
* experience: Candidate total experience in years
* company\_size: No of employees in current employer's company
* company\_type : Type of current employer
* lastnewjob: Difference in years between previous job and current job
* training\_hours: training hours completed
* target: 0 – Not looking for job change, 1 – Looking for a job change

# הקדמה:

* אחד האתגרים בכניסה לעולם של machine learning הוא הכרה והתמצאות בכל המונחים ושיטות העבודה. אנחנו מאמינים שכדי לצלוח זאת עליכם לתרגל ולהשתמש במינוחים מקצועיים כמה שיותר כדי לחבר למונחים את ההבנה של הרעיונות שאותן הן מייצגים. כל המונחים בעולם זה הם באנגלית ולכן גם המצגות כתובות באנגלית, בנוסף בדוח הפרויקט נרצה לראות אותכם משתמשים במונחים מקצועיים ומכניסים אותם בהקשרים הנכונים בעבודה.
* בתרגיל יושם דגש על שימוש בתוכנה לצורך מענה על שאלות הקשורות בנתונים ובניתוחם. לא פחות חשובות הן התובנות משימוש זה לגבי עולם התוכן של הבעיה הנחקרת, כשהשאלה המרכזית הינה: מה בעצם למדנו מתרגיל זה? יש לשלב טבלאות, גרפים וכו', לנתחם ולהשליך מהם על עולם התוכן הנחקר.
* יש להגיש את כל קבצי Python עליהם עבדתם.

# מבנה העבודה

# Data collection and Sensing (10 נק')

1. **הנתונים אותם קיבלתם כבר עברו את 2 השלבים הראשונים.**

* ענו בקצרה: מהו הData collection ? איזה סוג Sensing בוצע על הדאטה (סטטי \ דינמי), הסבירו ?
* הציעו סוג Sensing שלא בוצע על הדאטה, והסבירו איך הוא יכול לעזור למשימת הלימוד (אין צורך להוסיף אותו)
* מהי קטגוריית וסוג משימת הלמידה (מצגת 3), הסבירו את תשובתם ? האם ניתן להשתמש בנתונים כדי לבצע עוד סוג של משימת למידה ?

# Dataset Creation (75 נק')

**[1. Exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis)** – **20 נק'**

* הציגו את התפלגות הנתונים של כל המשתנים בסט הנתונים כולל משתנה המטרה, מה מסקנותיכם בנוגע למשימת הלימוד (השתמשו בגרפים ומדדים סטטיסטיים כדי להמחיש את מסקנותיכם ושמרו על תמציתיות בהסברים)?

**אין לחזור במילים על מה שניתן לראות בגרף – הסבירו את מסקנותיכם עבור כל משתנה.**

**שאלות מכוונות:**

* מהו טווח הערכים שכל משתנה מקבל?
* מה משמעות המשתנה בהקשר של משימת הלימוד ?
* האם הוא קשור להבנתם של עוד משתנים בDataset?
* על מה מלמדות הסתברויות אלה?
* האם סט הנתונים מאוזן? במה זה תלוי?
* קשרים מעניינים בין משתנים מסבירים

**2. כעת עבדו עלפי שלבי העבודה שלמדנו בהרצאה 2 – שימו לב, ייתכן ולא כל השלבים רלוונטיים! במידה ושלב מסוים אינו רלוונטי, הסבירו מדוע הוא אינו רלוונטי (על ההסבר להיות משכנע)**

* כאשר אתם בוחרים לבצע פעולה על הנתונים, הסבירו בתמציתיות מהו התהליך שאתם רוצים לבצע, ואיך בחרתם לבצע אותו ?
* עבדו באופן סיסטמטי על פי השלבים, ציינו כל שלב בכותרת מתאימה, ושאבו רעיונות מהמצגת. כאן נצפה לראות שימוש במונחים והשיטות אותם למדנו בהרצאה המראה על הבנת החומר הנלמד.
* בשלב זה אין תשובה אחת נכונה, כל קבוצה יכולה לבחור איך לבצע שלבים מסוימים על פי שיקול דעתכם ובמידה וניתן לכך הסבר מספק.
* שימו לב לתוכן של כל שלב על פי המצגת. **הכנסת פעולה לא בשלב הנכון תוריד ניקוד!**
* הניקוד על כל שלב יהיה:

1. dimensionality reduction – 5 נק'

2. feature extraction; pre-processing - 10 נק' עבור כל סעיף

3.feature selection, feature representation – 15 נק' עבור כל סעיף

* חלק מציון החלק יהיה על שלבי עבודה ברורים והצגתם באופן כזה שיהיה ניתן להבין מה עשיתם. כל פעולה שאתם מבצעים ציינו עם הפרדה כשיהיה קל להבחין (לא לרשום פסקה אחת ארוכה). **דוח מבולגן יוריד ציון !**
* **בסוף כל שלב, במידה וביצעתם שינויים בנתונים, הציגו סיכום קצר בו אתם מציינים איזה שינוי ביצעתם בנתונים. הציגו פלט של כמה רשומות מתוך הdataset שיצרתם על מנת שיהיה ניתן לראות ויזואלית מה עשיתם!**

**ניתן להוסיף כל פלט או הסבר שתומך ומסכם כל שלב.**

# Model Training (5 נק')

* על סמך החומר שלמדנו בכתה בהרצאה 3, בחרו שיטת ולידציה לנתונים. הסבירו מדוע בחרתם בה ? מה היתרונות שלה על פני שאר השיטות שלמדנו ?
* הסבירו את תהליך הולידציה שתבצעו בעזרת השיטה שבחרתם. איזה מטריקה תבחרו להשתמש

איכות הדו"ח ורמת שימוש בתכנת Python (10 נק')

* איכות הדוח – ויזואלית, סידור, קריא
* איכות השימוש בתוכנה – מורכבות הגרפים, הפלטים, ושיטות

בהצלחה !