INTRO TO DATA SCIENCE
מבוא למדע הנתונים

פרויקט גמר

המכון הטכנולוגי חולון,
אביב תשפ"ב

פרויקט הגמר

פרויקט הגמר

מטרה:

רכישת ניסיון מעשי ביישום טכניקות מדע הנתונים

דרישות:

1. הגדרת בעיה וזיהוי שאלות מחקר הולמות
2. זיהוי נתונים בקנה מידה משמעותי (לפחות K50 נתונים לפני ניקוי) והרכשתם (לעיתים ממספר מקורות)
3. ניתוח ראשוני וטיוב נתונים
4. EDA ו-ויזואליזציה
5. ניתוח נתונים מתקדם - בחירת שיטות ואלגוריתמים מתאימים
6. יישום הפתרון ובדיקה של השיטות שיושמו (הערכת ביצועים)
7. הסקת מסקנות ודיווח מסכם

תכולת הפרויקט והגשתו – קוד

- Deliverable: URL of your own **GitHub Pages** site hosting an .ipynb/.html export of your final tutorial
- https://pages.github.com/ – make a GitHub account, too!
- https://github.com/blog/1995-github-jupyter-notebooks-3
- The project itself:
- ~1500+ words of Markdown prose
- ~150+ lines of Python (in other languages it might be MUCH longer.. ☺)
- **Should be viewable as a static webpage – that is, if I (or anyone else) opens the link up, everything should render and I shouldn't have to run any cells to generate output**
- Ppresentations can work as well
- Short clip 5-8 minutes

תכולת הפרויקט והגשתו – הגשה והגנה

- הגשות יתבצעו באופן הבא:
  - כל פרויקט יוגש באמצעות סרטון שישלח למרצה הבודק, עד לדד ליין להגשה (פרטים בהמשך המצגת). במועד הגשת הסרטון, יש לשלוח גם לינק לגיטהב של הפרויקט.
  - בשבועיים שלאחר הדד ליין יתקיימו פגישות של 5–7 דקות לכל פרויקט עם המרצה הבודק. בפגישות אלו תצטרכו להגן על הפרויקט. דהיינו, להסביר מה עשיתם (כמענה לשאלות מטעם הבוחן) ולהצדיק את השיטות שבחרתם.
  - ציון הפרויקט מהווה 34% מהציון הסופי, וייקבע ע"פ הסרטון וההגנה על הפרוייקט.

  - חשוב!
    - לא ניתן לקבל ציון בקורס ללא הגשת פרויקט.
    - לא ניתן לקבל ציון בקורס ובפרויקט על פרויקט שאינו מכיל למידת מכונה.

Grading guidelines

Grading (basic):

- Project planning (proper research question and answer/conclusion) (15)
- Data acquisition (15, using a pre-made dataset does not entitle for points, API entails only for 5 points)
- EDA quality and comprehension (15)
- Machine Learning experiments and insights (15)
- Proper understanding of presented material (15)
- Presentation quality (10)

Bonus (points – one can get more than 100 in theory, in most cases, no one gets a combination of all of them)

- In class presentation (5)
- Special data acquisition (crawling or other very special handling) (10)
- Wow effect (5)
- Single student (5)
- Novelty of project (10)

הנחיות נוספות

- הפרויקטים יתבצעו בקבוצות של 1-2 סטודנטים, אפשרי מקבוצות שונות.
- **דד ליין לאישור נושא:** יש לשלוח את הצעת המחקר (נושא מחקר, נתונים, דרכי ניתוח) **<u>למתרגל</u>** הרלוונטי לאישור עד לתאריך ה- 1.5.22. הגשות מאוחרות של הצעת המחקר יישאו בקנס (ראו למטה), ויתאפשרו רק עד ל- 15.5.22. **מי שלא מגיש הצעה עד למועד זה – לא יוכל לקבל ציון עובר בקורס.**

  שימו לב – במקרה של הגשה של זוג, הצעת המחקר תוגש למתרגל של המגיש בעל שם המשפחה הראשון מבחינת א"ב (למשל דן כהן ודנה לוי, יגישו למתרגל של דן כהן)
- **דד להגשת הפרויקט:** 30.6.22 (לא יינתנו הארכות)
- בונוס מיוחד למי שיסיים לפני סוף הסמסטר ויציג את הפרויקט בכתה
- לאחר הגשת הפרויקט, נדרשת הצגה פרונטלית לפני בוחן הפרויקט (מול אחד ממרצי הקורס)

- על כל יום איחור באישור הצעת המחקר – יורדת חצי נקודה. לא ניתן לאחר יותר משבועיים, ולא יתקבלו הצעות מחקר אחרי ה- 15.5.22

הנחיות נוספות

- **פרויקטים ללא למידת מכונה לא יבדקו. לא תינתן אפשרות להגשה חוזרת**
- הצגת הפרויקט תעשה בידי כלל חברי הצוות – הבוחן של הפרויקט יכול לבקש מכל אחד מחברי הצוות להציג (לפי בקשת הבודק)
- קיימת אפשרות למתן ציון לא אחיד בין חברי הפרויקט
- אין לקחת דאטהסטים מקאגל Kaggle
- אין לקחת פרויקטים מסטודנטים או קורסים אחרים
- במידה ונלקחים נתונים ו/או נתוחים מהאינטרנט – יש לתת קרדיט למקור
- במידה ויתגלה קוד זהה לקוד ממקור אחר, ללא ציון המקור, הפרויקט יפסל
- אפשר להשתמש בכל שפה ובכל חבילה - התייעצויות עם המתרגלים במהלך הסמסטר – בשמחה!

בוחני הפרויקטים והגנה

- כל פרויקט יוגש לאחד מהמרצים, על פי רשימה שתתפרסם לקראת סוף הסמסטר.
- הגנה על הפרויקט תקבע מול המרצה הבוחן (לרוב באמצעות שיתוף של טבלת תאריכים ושעות, לקראת הדד ליין להגשת הפרויקט).
- סטודנטים שיציגו את הפרויקט בכיתה, בשבוע האחרון של הסמסטר, פטורים מהגנה נוספת על הפרוייקט.

מבנה הסרטון

- אורך סרטון: 5-8 דקות
- מבנה:

א. הקדמה – עד דקה לכל היותר על הנושא, מדוע הוא מעניין, ושאלות המחקר סביבן סובב הפרויקט

ב. מקורות הנתונים והרכשה – עד דקה וחצי

ג. ניתוח ראשוני וטיוב – עד דקה

ד. ויזואליזציה ו- EDA – עד דקה וחצי

ה. ניתוח נתונים מתקדם – עד שתי דקות

ו. יישום והערכת ביצועים – עד דקה

ז. סיכום ומסקנות – עד חצי דקה

דוגמאות לפרויקטי עבר

- מה מאפיין להיטים בספוטיפיי, והאם ניתן לחזות אותם?
- איך ניתן לחזות התרחשות רעידות אדמה ואת המיקומים שלהן? *(מצגת לדוגמא במודל)*
- מה הקשר בין דיווחי החדשות של חברות בורסאיות לבין ביצועי המניות והמדדים של בורסת ת"א?
- איך ניתן לחזות אילו שאלות בפורום יהיו פופולריות יותר מאחרות, ואילו שאלות יקבלו מענה קודם?
- מהמו ההבדל ביחס לאיכות החיים בכדוה"א בין מבוגרים לצעירים, ע"ס ניתוח מידע מטויטר?
- איך ניתן לאתר אבני חן, ואילו, בהינתן מאפייני מיקום *(מצגת לדוגמא במודל)*
- נפרסם הקלטות של סטודנטים שסיפרו על הפרויקטים שלהם..