יסודות מדע הנתונים – פרויקט קורס 2025

את הפרויקט יש לבצע בקבוצות של 3 סטודנטים. במידה ולא תמצאו קבוצה פנו אלי ואני אשדך ביניכם. יום לפני תאריך ההגנה יש להעלות למודל (בתיבה שנפתחה לכך) את מחברת הפרויקט.

בפרויקט זה נחקור התנהגויות מחירים של טיסות הלוך ושוב בין יעדים באירופה. שימו לב שאת עיקר הדרישות ניתן כמובן לממש באמצעות החומר אותו למדתם. יחד עם זאת, בדרישות הפרויקט שולבו גם מספר משימות המתייחסות לחומר שעליכם ללמוד/להתנסות לבד. מעבר לדרישות עצמן – הרגישו חופשיים לקיים ניסויים עם שיטות נוספות על מנת לשפר את התוצאות – זה רק יכול להוסיף.

הסבר קצר לגבי טרמינולוגיה של עולם הטיסות:

TTT = Time to Travel – כלומר הפרש הימים בין תאריך החיפוש באתר לתאריך יום הטיסה הלוך המבוקשת.

LOS = Length of Stay – כלומר מספר הלילות בין תאריך הטיסה הלוך לתאריך הטיסה חזור.

Snapshot Date = התאריך בו ביצעתם את דגימת האתרים.

הגשת הפרויקט תעשה באמצעות מחברת Ipynb. שימו לב שעליכם לפרט (עם הערות) את כל הניסויים שביצעתם במחברת – גם את מה שהיה "בדרך" אל הפתרון. כך גם במקרה שלא הצלחתם לבצע סעיף מסוים, נדע מה ניסיתם.

בהמשך יתבצעו הגנות בקורס (במועדים שיתפרסמו). בהגנה עליכם להציג את המחברת ולענות לשאלות שתישאלו. השאלות מתייחסות בעיקר למה שביצעתם בפרויקט (בין אם התבקשתם ובין אם בחרתם/יזמתם) אך לא רק – אלא גם לחומר שנלמד במהלך הסמסטר, שעליכם לשלוט בו לקראת ההגנה. כל סטודנט החבר בקבוצת הפרויקט צריך להכיר היטב את הקוד שנכתב על ידי כל אחד מחברי הקבוצה.

שלב א' – Scraping:

השתמשו בספריה המבצעת Web scraping מתוך פייתון (לבחירתכם. לדוגמא: Selenium, Beautifulsoup, Scrapy) על מנת לדגום את תוצאות החיפוש על פי הקרטריונים הבאים:

- 2 מבין האתרים: Google, Skyscanner, Kayak, Expedia
- טיסות (הלוך ושוב) בין פריז, לונדון ורומא (סה"כ 6 אפשרויות של מקור ויעד: פריז ללונדון, פריז לרומא, לונדון לפריז, לונדון לרומא, רומא ללונדון, רומא לפריז)
- 1 מבוגרים, 0 ילדים
- אפשרות גם לטיסות ישירות וגם לא ישירות (עם קונקשן)
- תאריכים: עבור כל אחת מזוגות הערים – יש לבצע חיפוש על כל הקומבינציות של **TTT** בין 1 ל 30 ושל **LOS** בין 1 ל 5 (כלומר סה"כ 150 חיפושים של תאריכי טיסות שונות עבור Snapshot מסוים) עבור 3 תאריכי **Snapshot** שונים (סה"כ לפחות 450 חיפושים). יש לקחת בחשבון שייתכן ותצטרכו לתקן באגים בקוד ה scraping ולכן הערכו בהתאם וסיימו את שלב זה כמה שיותר מוקדם על מנת לאפשר מספיק זמן לסריקות עצמן.
- עבור כל אחת מהקומבינציות של TTT ו LOS עליכם לשמור בקובץ (אפשר CSV) את הפרטים המופיעים על הדף עבור 100 טיסות לפחות (שימו לב שבחלק מהאתרים הסקייפר יצטרך לעבוד לעמודים הבאים ו/או לגלול בדף כלפי מטה כדי להשיג תוצאות נוספות). יודגש שהקוד של ה scraper צריך להיות אוטומאטי לחלוטין ללא התערבות שלכם במהלך הריצה.
- עבור כל טיסה עליכם לאסוף את כל השדות המופיעים על המסך (כולל חברות התעופה המפעילות, שעות הטיסות, האם קונקשן או טיסה ישירה, זמן הקונקשן במידה ומדובר בטיסת

קונקשן, מחיר, שדות התעופה, וכל שדה אחר רלוונטי שעשוי להשפיע על המחיר ומופיע על המסך.

- שימו לב שהנתונים שעליכם לאסוף מופיעים רק בדפי החיפוש – אין צורך להיכנס לפרטים של כל טיסה ספציפית (להיכנס למסך הזמנות) על מנת לאסוף נתונים נוספים.

שלב ב' – Exploration + Data preprocessing:

- יש ליצור גרפים של התפלגויות של:
  - מחירי הטיסות בכללי
  - מחירי הטיסות בהינתן חברות מפעילות
  - מחירי הטיסות על פי יעדים שונים
  - התפלגות הטיסות של החברות השונות על הטיסות פר יעד
  - התפלגות זמני הקונקשן וזמני הטיסות בין היעדים השונים
  - השפעת שדות התעופה בכל עיר על המחירים של הטיסות בין היעדים
- עליכם להסיר מהנתונים טיסות שעל פי המחיר מהווים outlier על פי שיטת Tukey (1.5IQR) פר כל זוג מקור ויעד
- הציגו PairGrid (של ספריית seaborn) עבור המשתנים וכתבו בפיסקה מסקנות/תובנות מהגרף. כמו כן התייחסו לקשר בין המשתנים השונים X לבין משתנה המטרה (מחיר)

שלב ג' – תחזית מחירי הטיסות

בסעיף זה עליכם לממש במחברת פתרון לחזות את מחירי הטיסות (בהינתן קומבינציית הפרמטרים שהופיעו על המסך שאספתם קודם לכן) על פי הנתונים שאספתם לגבי כל טיסה בשלבים הקודמים.

- שימו לב שעליכם להתנסות במספר אלגוריתמי רגרסיה (לפחות את אילו שנלמדו: LinearRegression, DecisionTreeRegressor, GaussianProcessRegressor) ובנוסף עליכם לבחור עוד 3 אלגוריתמי רגרסיה שלא נלמדו מתוך ספריית sklearn.
- עליכם לוודא שחילקתם את הדאטה ל Train, Test בחלוקה של 30-70 לטובת ה Train בבחירה אקראית.
- עבור כל אלגוריתם עליכם לבצע התנסויות בפרמטרים שונים (למשל עבור עץ רגרסיה לנסות הגדרות שונות לעומק העץ, לבחינת קריטריון החלוקה וכד'. עבור תהליכים גאוסיאנים להתנסות עם קרנלים שונים.
- הציעו פיצ'רים נוספים (המחושבים על סמך הקיימים) ובידקו כיצד הם משפיעים על תוצאות התחזית שלכם. למשל אולי כדאי לקחת לא רק את ערכי ה TTT אלא גם את היום בשבוע של ה טיסת ההלוך ו/או טיסה החזור, את הקירבה לסוף החודש, קירבה לחג בתקופה וכד'
- עבור לפחות 2 אלגוריתמים - דרגו את המשתנים השונים על פי מידת השפעתם על תוצאות הרגרסיה. הציעו לפחות 2 דרכים שונות לבצע זאת (נקרא גם feature importance).
  - דרך אחת תתבסס על הדרך בה האלגוריתם הספציפי עובד
  - דרך אחרת תתעלם מאיך האלגוריתם עובד ותתייחס אליו כקופסה שחורה
- עבור כל הרצה של אלגוריתם, חשבו את השגיאה על ה Test ועל ה Train והדפיסו Residual Plot כפי שביצענו בכיתה. כתבו פסקה מסכמת לגבי התובנות מהגרף.
- מדדי שגיאות נדרשים עבור כל אלגוריתם: RMSE, MSE, MAE ,R2.
- נתחו את התוצאות שקיבלתם באלגוריתמים השונים. התייחסו לחולשות/יתרונות של אלגוריתמים שונים שהשתמשתם בהם ונסו להסביר מה הסיבה לכך (במיוחד המתקדו במקרים בהם אלגוריתם מסוים קיבל תוצאות טובות/גרועות משמעותית יחסית לאלגוריתמים אחרים).
- התנסו בשיטות נרמול שונות על הנתונים (לא חייבים על כל השדות) טרם הרצת האלגוריתמים ודונו בהשפעם השיטות על ביצועי האלגוריתמים השונים
- הפרידו את הניתוחים בין הנתונים שהורדתם מהאתר הראשון לבין אילו של האתר השני שסרקתם. מצאו הבדלים בין המודלים על האתרים השונים, לא רק בביצועים אלא גם במשתנים המשפיעים על הפרדיקציות, חוזקות/חולשות אלגוריתמים מסוימים וההבדלים הניכרים לעין

- הציגו גרף המראה את התפלגות השגיאות R2 של המודל הטוב ביותר שלכם על פני הטיסות השונות
- בצעו את הניסוי שוב על האלגוריתם הטוב ביותר שקיבלתם – הפעם על חלוקה שונה של Train ו Test :
  - ה Train יכיל את הנתונים עבור TTT<=25
  - ה Test יכיל את הנתונים עבור TTT>25 (כך שלעשה אנו בודקים כאן את היכולת לחזות את המחירים ב"עתיד"

שלב ד' – למידת פערי מחירים באתרים מתחרים

- בנו מודל החוזה את הפרש המחירים בין האתר הראשון שסרקתם לבין האתר השני שסרקתם בהינתן הנתונים המופיעים על המסך שאספתם בסקרייפינג. בדומה לדרישות לעיל בסעיף הקודם – גם כאן התייחסו לדרישות הטכניות (חלוקת Train/Test, הרצת מספר אלגוריתמים, בחינת השגיאות וכד')

שלב ה' – ביצוע קליסטור לחברות התעופה

- בסעיף זה עליכם לקלסטר את החברות המפעילות את הטיסות על פי הדמיון ביניהן במדיניות התמחור. בסעיף זה תרצו לסנן החוצה מהדאטה את הטיסות שאינן מופעלות על ידי אותה חברה בהלוך ובחזור. הציעו דרכים שונות להציג את הקלאסטרים של החברות

ב ה צ ל ח ה !