מדעי הנתונים בתעשייה

תרגיל 1

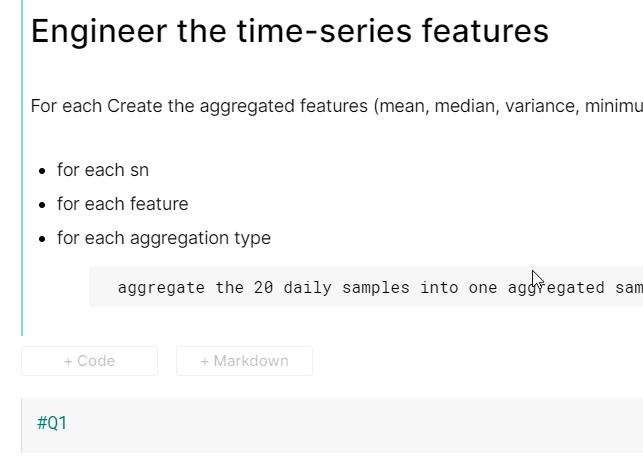
דף הוראות

מסמך זה בא לתאר את הדרישות המדויקת עבור מטלה 1, ולתת הכוונה לפונקציות בהם צריך להשתמש. שימו לב שניתן לפתור את התרגיל במספר דרכים, מובאת פה המלצה אחת בלבד.

השתדלנו לציין את כל הפונקציות החשובות אותם תצטרכו, וגם את הפרמטרים החשובים שתצטרכו להצהיר עליהם. מומלץ מאוד לקרוא את התיעוד של הפונקציות שאתם משתמשים בהם בכל מקרה, הן יחזרו על עצמן לאורך הקורס!

אתם מוזמנים לשאול שאלות נוספות במודל ובשעות הקבלה.

**סעיף 1:**



**הפלט הנרש**: Dataframe (טבלה) שבה כל שורה היא דיסק קשיח, המזוהה לפי מספר סיראלי. העמודות בטבלה הם {ממוצע, חציון, שונות, מינימום, מקסימום} מכפלה קרטיזית עם {כל העמודות המקוריות}. שימו לב שלכל דיסק ידוע אם הוא נכשל או לא (נתון ייחודי שלא משתנה לאורך חיי הדיסק) וגם נתון זה צריך להיות בפלט של שלב זה.

**טיפים להתמודדות**: לבצע group\_by כפי שמופיע במחברת בתאים הקודמים (טיפ: ניתן לבצע group by על יותר משדה אחד, שולחים את השמות בתוך מערך), על הפלט של הgroup by מומלץ להפעיל את הפונקציות mean, median…, כל פונקציה בנפרד. לא מומלץ להשתמש בdescribe.

כדי לחבר את המידע הסטטיסטי שהוא פלט של כמה פונקציות שהרצתם בנפרד, מומלץ להשתמש בpd.merge, שימו לב לפרמטרים left\_index, right\_index וגם לפרמטר suffixes. שימו לב גם שהפונקציה מופעלת כל פעם על 2 טבלאות בלבד ולא יותר, ואתם רוצים למזג מספר גדול של טבלאות, כדי לבצע את המיזוג יש לשרשר את הפלט של הפונקציה עם הטבלה הבאה עד שכל הטבלאות יהיו מחוברות.

להלן דוגמה לרשימה של העמודות שאמורות להיות בפלט:

Output columns for Q1:

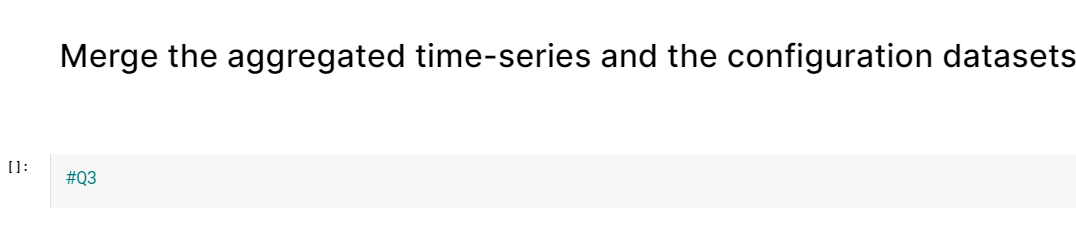
[**'d\_id', 'fail'**, 'bb\_count\_mean', 'r-w\_rate\_mean', 'reconsects\_count\_mean ', 'recovbydrv\_count\_mean', 'xfer\_rate\_mean ', 'bb\_diffs\_mean', 'bb\_count\_median', 'r-w\_rate\_median', 'reconsects\_count\_median', 'recovbydrv\_count\_median', 'xfer\_rate\_median', 'bb\_diffs\_median', 'bb\_count\_var', 'r-w\_rate\_var', 'reconsects\_count\_var', 'recovbydrv\_count\_var', 'xfer\_rate\_var', 'bb\_diffs\_var', 'bb\_count\_min', 'r-w\_rate\_min', 'reconsects\_count\_min', 'recovbydrv\_count\_min', 'xfer\_rate\_min', 'bb\_diffs\_min', 'bb\_count\_max', 'r-w\_rate\_max', 'reconsects\_count\_max', 'recovbydrv\_count\_max', 'xfer\_rate\_max', 'bb\_diffs\_max']

**סעיף 2:**



**מה צריך לעשות?** ראשית להשתמש בפונקציה change\_capcity, היא מוכנה. פשוט להשתמש בה ע"פ התיעוד.

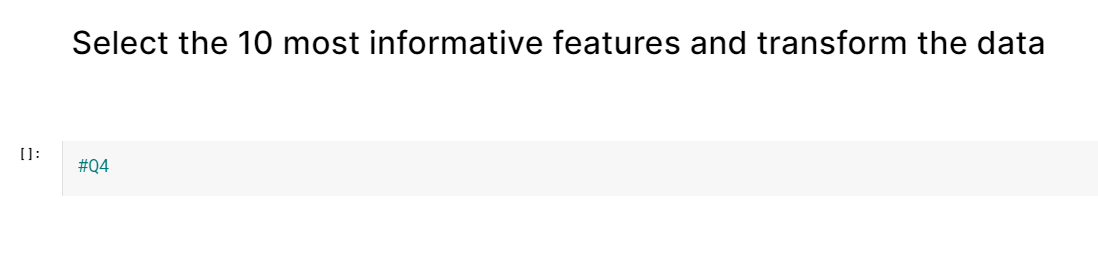
בנוסף עליכם להמיר את העמודה model ליצוג של [one-hot-vector](https://en.wikipedia.org/wiki/One-hot). מומלץ להשתמש בפונקציה pd.get\_dummies.

**סעיף 3:** 

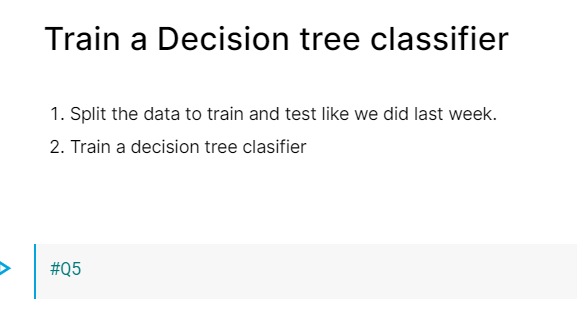
עליכם למזג (JOIN) את הdataframe אותו יצרתם עד עכשיו, ביחד עם הנתונים שנמצאים בconf\_data

מומלץ להשתמש בpd.merge שימו לב לפרמטרים on, how.

שימו לב שלאחר מכן אין צורך בשדה d\_id', יתרה מכך הוא יפריע לתהליך! כמו כן צריך לשמור בdataframe נפרד את השדה fail מאחר והוא מכיל את התשובות לחיזוי.

**סעיף 4:** 

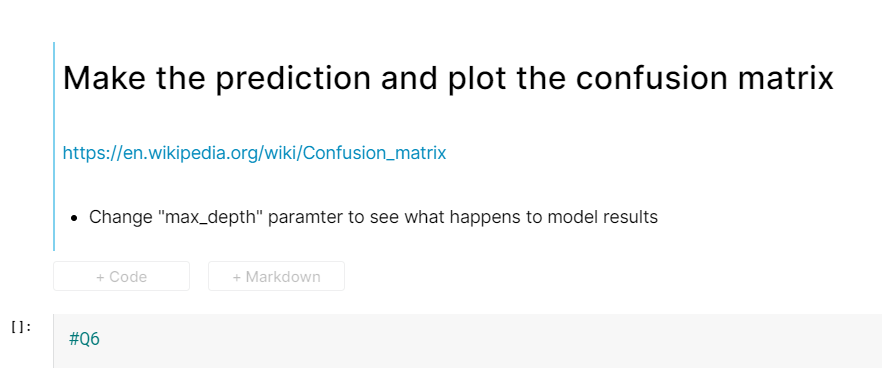
עליכם לבחור את 10 הפיצ'רים הכי משמעותיים בנתונים. הפלט של סעיף זה יכיל את הדאטא מהסעיף הקודם, אך ישארו בו רק 10 העמודות הכי חשובות. קיימות אין ספור דרכים לבצע את המשימה הזו הן תיאורטית (איזה אלגוריתם לבחור) והן מעשית (מימוש).

**סעיף 5:** 

עליכם לאמן מודל מסוג עץ החלטה. הפלט של הסעיף זה בעצם העץ.

מומלץ להשתמש בDecisionTreeClassifier אשר עשינו לו import בתא הראשון. שימו לב למתודה fit.

שימו לב שצריך להשתמש בחלוקה לtrain וtest כפי שעשינו בשיעור הקודם.

**סעיף 6: **

באמצעות המודל מהסעיף הקודם, עליכם לבצע חיזוי לנתונים. שימו לב לפונקציה predict של עץ ההחלטה. בנוסף עליכם ליצרconfusion matrix. עשינו Import לפונקציה המתאימה בתא הראשון, אך רצוי למצוא באמצעות חיפוש באינטרנט דרך להדפיס את המטריצה עם השמות על העמודות והשורות של המטריצה.

נסו ערכים שונים של הפרמטר max\_depth וכתבו (במחברת – הערה או בMarkdown) הסבר עם התוצאות של הניסויים שלכם.

**סעיף 7:**

גיא טוען שאם עקבתם אחר ההוראות עד סעיף 6 כולל אז יש טעות קריטית במחברת ושלא ניתן להסתמך על הדיוק שיצא לנו במחברת כשנקבל מידע על דיסק חדש.

1. הסבירו מדוע במילים
2. הסבירו מה צריך לעשות בשביל לתקן את הטעות
3. תקנו את הטעות (מצופה כי תגישו מחברת לאחר התיקון)

לא כל הסעיפים יבדקו.

**סעיף 8:**

הציעו שיפור למודל החיזוי שלנו.

