כריית נתונים - מעבדה 1 - היכרות עם הנתונים ועיבוד מקדים

מעבדה זו יש לממש בעזרת מחברת ולהגיש **תוצר PDF או HTML בלבד של המחברת - מחברות בפורמט IPYNB לא ייבדקו.**

(בעת מימוש המעבדה - ניתן להעזר במקורות המצורפים, מצופה גם לבצע חיפוש ובדיקה עם מקורות מידע נוספים מהרשת)

1. **היכרות עם הנתונים**

בעזרת קובץ הנתונים - dmc2010_train.txt (כרפרנס ניתן גם תיאור של הנתונים בקובץ dmc2010_features.pdf). שימו לב כל עמודה מופרדת בעזרת ;

יש לבצע/לבדוק ולפרט עבור כל עמודת נתונים בעזרת ספריית פנדס, וספריות נוספות לבחירתכם:

שימו לב ישנם כלים שיתנו את כל התשובות בפקודה או שתיים, שווה לחקור ...

a. פירוט העמודות וסוגי הנתונים על פי פנדס.
b. סוג העמודה במילים שלכם על פי הנלמד בכיתה (נומינלי,סודר,רציף וכו')
c. שכיח או ממוצע בהתאם לסוג העמודה
d. ערכי מינ/מקס או פירוט ערכים ייחודי (בהתאם לסוד העמודה)
e. כמות ערכים חסרים בכל עמודה (אם קיים בכלל)
f. תלות (קורולציה או מבחן אחר) בין כל עמודה לכל עמודה (בהתאם לסוג הנתונים)
g. בעבור העמודות הרציפות הציגו נתונים סטטיסטיים (ממוצע, סטיית תקן, רביעים וכו'
h. עבור ערכים בדידים יש להציג כמה יש מכל ערך ייחודי
i. עמודת ה target90 הינה עמודה, בינארית שבעזרתה ננסה ללמוד, מיהם האנשים שעבורם יהיה לנו אינטרס לשלוח קופון (אלו שעשו קניה נוספת ב90 יום שאחרי) - יש להתייחס לעמודה זו באופן פרטני:
   i. מה ניתן לעשות במידה ויש לנו ערכים חסרים בעמודה.
   ii. בכדי "ללמוד" מעמודה זו - מה לדעתכם צריכה להיות התפלגות הערכים, ומה ניתן לעשות במידה והתפלגות זו לא עומדת בדרישות (קראו על imbalanced dataset)
   iii. כיצד לדעתכם יהיה ניתן להעריך (לתת ציון לטיב המודל) למודל שנבנה מנתונים אלו - הציעו מספר שלבים לטיפול בנתונים בכדי ליצור מסגרת הערכה שכזו.

https://www.analyticsvidhya.com/blog/2021/04/top-python-libraries-to-automate-exploratory-data-analysis-in-2021/
https://www.kdnuggets.com/2021/02/pandas-profiling-one-line-magical-code-eda.html
https://pandas.pydata.org/docs/getting_started/intro_tutorials/06_calculate_statistics.html
https://chartio.com/learn/data-analytics/what-is-exploratory-data-analysis/
https://towardsdatascience.com/how-to-build-an-eda-app-in-python-af7ec4b51528

2. **נירמולים**

- יש לסכם בקצרה מהו נרמול נתונים, מהי סטנדרטיזציה של נתונים וההבדל ביניהם
- יש לסכם בקצרה על כל סוג נרמול/סטנדרטיזציה (סעיפים a-c)
- יש למשש פונקציה המקבלת סדרת נתונים ופרמטרים רלוונטיים ומחזירה אותה לאחר נרמול בהתאם לפונקציה המתבקשת (עבור כל סעיף פונק נפרדת) **בעזרת פייתון וספריות סטנדרטיות בלבד**
- יש לחקור ולמצוא ספריות פייתון שיודעות לבצע נרמול/סטנדרטיזציה ובצע בעזרתן את אותן פעולות

a. Min-max
b. Z-score
c. Decimal Scaling

מקורות לקריאה :

https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html
https://medium.com/@dataakkadian/standardization-vs-normalization-da7a3a308c64
https://www.geeksforgeeks.org/data-normalization-in-data-mining/
https://cs.sounak.in/custom-min-max-z-score-mad-z-score-and-decimal-scaling-normalization-in-python/
https://www.geeksforgeeks.org/python-decimal-normalize-method/

3. **דיסקרטיזציה לא מונחית של ערכים רציפים - Discretization using binning**
   - הסבירו מהי דיסקרטיזציה
   - מהי דיסקרטיזציה של עומק שווה Equal-frequency discretization
     ממשו פונקציה המקבלת סדרת מספרים ופרמטרים רלוונטיים ומחזירה סדרה של ערכים בדידים כנדרש **בעזרת פייתון וספריות סטנדרטיות בלבד**
   - מהי דיסקרטיזציה של רוחב שווה Equal-width discretization
     ממשו פונקציה המקבלת סדרת מספרים ופרמטרים רלוונטיים ומחזירה סדרה של ערכים בדידים כנדרש **בעזרת פייתון וספריות סטנדרטיות בלבד**
   - יש לחקור ולמצוא ספריות פייתון שיודעות לבצע דיסקרטיזציה בעזרת דליים (bins) ולבצע בעזרתן את אותן פעולות

https://www.saedsayad.com/unsupervised_binning.htm

4. החלקה
   a. הסבירו מהי החלקת נתונים (smoothing)
   b. הסבירו מהו ממוצע נע ואיך הוא עוזר בהחלקת נתונים
   c. פרטו על השיטות וממשו פונקציות המבצעות את ההחלקה **בעזרת פייתון וספריות סטנדרטיות בלבד**
      i. Simple Moving Average
      ii. Weighted Moving Average
      iii. Exponential Moving Average
   d. הסבירו על Binning Methods for Data Smoothing
      פרטו וממשו את השיטות השונות המפורטות **בעזרת פייתון וספריות סטנדרטיות בלבד**

      i. Smoothing by bin means:
      ii. Smoothing by bin boundaries
   e. יש לחקור ולמצוא ספריות פייתון שיודעות לבצע החלקה ולבצע בעזרתן את אותן פעולות

https://www3.nd.edu/~rjohns15/cse40647.sp14/www/content/lectures/06%20-%20Data%20Quality.pdf
https://t4tutorials.com/binning-methods-for-data-smoothing-in-data-mining/#Binning_Methods_for_Data_Smoothing
https://machinelearningmastery.com/moving-average-smoothing-for-time-series-forecasting-python/
https://www.sciencedirect.com/topics/economics-econometrics-and-finance/smoothing-technique

https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc42.htm