מערכות תבוניות – פרויקט

בהתייחס לנתונים בקובץ data.csv , יש לממש את הנדרש בתכנית פייתון. יש להגיש את התכנית עצמה (קובץ project.py) וקובץ word, המתאר את התוצאות (גרפית) וכולל הסברים, עפ"י הנדרש בסעיפים הבאים.

א. טען את הנתונים לתוך dataFrame (יש להשתמש בספריה Pandas).

ב. טפל בנתונים החסרים באופן הבא:

- בדוק האם הנתונים תקינים. נתון שאינו נמצא בטווח הנכון או אינו מהסוג המתאים לשדה – יש למחוק אותו

- השלם נתונים שניתנים להשלמה מתוך נתונים אחרים.

- באם בעמודה מסוימת למעלה מ 25% מהנתונים חסרים – הורד את העמודה

- באם בשורה מסוימת חסרים יותר מ 2 נתונים – מחק את השורה.

ג. בחר 2 זוגות של עמודות מתוך הנתונים והצג כל זוג כגרף scatter. (בחר את ציר ה X וציר ה Y). (השתמש בספריה matplotlib)

ד. בהתייחס לנתונים שבחרת בסעיף הקודם – נרמל אותם כך שיהיו בטווח 0-1. שרטט scatter לאחר הנירמול. האם יש שינוי מהותי? (השתמש בפונקציה minmax)

ה. חשב רגרסיה לינארית עבור כל אחד מהזוגות בסעיף ג'. הצג את התוצאות. (השתמש בספריה sklearn)

ו. חשב רגרסיה לינארית עבור כל אחד מהזוגות בסעיף ד'. הצג את התוצאות.

ז. דון בהבדל בין תוצאות הסעיפים ה' ו-ו'.

ח. בהתייחס לסעיף ד' - האם יש נקודות המהוות ouliers? אם כן – ציין מהן והורד אותן מסט הנתונים. חזור על סעיף ד' ללא הנקודות הללו. השווה את התוצאות שקיבלת עכשיו לאלו שהתקבלו בסעיף ד'.

ט. השתמש בספריה metrics והצג את המדדים הבאים עבור כל אחת מהאנליזות שביצעת בסעיפים ה' ו' ח' ו-ט': mean absolute error, mean squared error ו root mean squared error. השווה ודון בתוצאות.

י. בחר פרט אחד שהסרת אותו בסעיף ב' - מאחר והיו חסרים נתונים עבורו, ונסה להעריך את אחד הנתונים החסרים, בהתאם לתוצאות שקיבלת בסעיפים הקודמים.