תרגיל בית 4 (רשות) - **Data science**

מועד הגשה:5.6 בשעה 23:55

הנחיות כלליות להגשת תרגילי הבית :

- הגשת התרגילים תתבצע באמצעות אתר הקורס בלבד.
- יש להגיש את התרגיל לתיבת ההגשה המתאימה, בהתאם לפורמט הבא: **HW#_ID1_ID**2 , כאשר התו # מציין את מספר התרגיל, **ID**1 מציין את תעודת הזהות של הסטודנט הראשון, **ID**2 מציין את תעודת הזהות של הסטודנט השני.

  לדוגמה:**HW1_000000000_111111111.py** זהו קובץ בודד שפותר את תרגיל בית 1.
- בנוסף, יש לכתוב כהערה (בתוך הקוד) בראש כל קובץ את השמות המלאים (שם פרטי ושם משפחה) בלועזית ואת תעודת הזהות של כל אחד מהשותפים, כאשר הפרטים של כל אחד רשומים בשורה נפרדת.
- ההגשה מתבצעת באמצעות קבוצות ההגשה בלבד.
- תרגיל אשר יוגש באיחור לא ייבדק, למעט אישורים מיוחדים כגון: מילואים, אשפוז וכו' המתקבלים לפני מועד ההגשה הרשמי.
- יש לכתוב הסברים ומענה מילולי על השאלות בגוף הקובץ בעזרת הערות ובאנגלית בלבד.
- יש להפריד בין פתרונות של כל שאלה בצורה הבאה. לפני המענה לשאלה מספר 1 כתבו בהערה **Q**1 וכן הלאה. במידה והשאלה מכילה מספר סעיפים (נניח 2 סעיפים) יש לכתוב בהערה **Q1a** לפני פתרון סעיף **a** בשאלה 1, ו- **Q1b** לפני פתרון סעיף **b** בשאלה 1 וכן הלאה.

בהצלחה!

1. Linear Regression , Regression Desicion Trees

הקובץ EnergyEfficiencyHW4.csv שייך לנתוני מחקר של דרישות עומס חימום עבור בניינים (Heating Load), כפונקציה של פרמטרים של הבניין.

שלבי עבודה:

1. חלקו את הדאטה באופן רנדומי ל-80% training set ו-20% test set והגדירו **random_state = 8123**
2. בנו מודל ריגרסיה לינארית לחישוב העומס בחימום (עבור סעיף זה אין צורך לנרמל נתונים).
3. חשבו את שגיאת שורש ממוצע הריבועים ((rmse עבור ה test set והשוו לשגיאה עבור ה train set.
4. הדפיסו את המקדם החופשי של המודל.
5. הדפיסו את שאר המקדמים וענו :

   א. איזה פיצ'ר קיבל את המקדם הקטן ביותר בגודלו ?

   ב. האם ניתן לדעת איזה פיצ'ר הכי משפיע על תוצאת החיזוי ? נמקו.
6. חזרו על סעיפים 2 ו 3 עבור מודל מנורמל .
7. האם ניתן לדעת כעת איזה פיצ'ר הכי משפיע על תוצאת החיזוי ? נמקו.
8. נא להריץ גם עבור עץ החלטה לריגרסיה ולהשוות בין השגיאות .

2. Logistic Regression

מצורף הקובץ memmographic_massesHW4.csv המכיל אינפורמציה על שיטת הממוגרפיה לגילוי סרטן השד, במטרה לזהות מתי הסיכוי יותר גבוה לזיהוי גידול ממאיר לעומת שפיר (ממאיר = 1 ושפיר = 0 בעמודה Severity).

מידע נוסף ניתן לקרוא בקובץ mammographic_massesHW4.txt המתאר את הדאטה בפירוט.

שלבי עבודה:

1. בדקו שאין ערכים חסרים/ בעייתיים. אם יש הסירו את אותן רשומות.
2. חלקו את הדאטה באופן רנדומי ל-80% training set ו-20% test set והגדירו **random_state = 8123**
3. בנו מודל Logistic regression לחיזוי עמודת המטרה Severity . גם כאן הגדירו אותו **random_state = 8123**
4. חשבו את מדדי ה .accuracy, precision, recall

3. KNN Classification

מצורף הקובץ wine.csv , המכיל נתוני אנליזה כימית עבור 3 סוגים שונים של זנים של יין הגדלים באזור ספציפי באיטליה. הפיצ'ר הראשון "Alcohol" מזהה את הזנים.

מידע נוסף ניתן לקרוא בקובץ . wine.txt

שלבי עבודה:

1. כידוע לכם אנו משתמשים עבור KNN בחישוב מרחק אוקלידי ולכן מומלץ לנרמל את הנתונים לפני הרצת המודל. בתרגיל זה נשווה בין מודל על דאטה מנורמל לעומת לא מנורמל .
2. חלקו את הדאטה באופן רנדומי ל-80% training set ו-20% test set והגדירו **random_state = 8123**
3. בחרו k התחלתי קונבנציונאלי ובנו מודל KNN על דאטה לא מנורמל, לחיזוי עמודת המטרה Alcohol . חשבו את מדדי ה accuracy, precision, recall .
4. כעת הריצו את המודל עבור ערכי k שונים . מהו ה k האופטימלי ?
5. כעת בנו מודל KNN על דאטה מנורמל וחשבו עבורו את מדדי ה **.**accuracy, precision, recall איזה מצב עדיף ?

4. Clustering - Kmeans, Hierarchical clustering

מצורף הקבצים wine.csv , wine.txt הנ"ל.

שלבי עבודה:

1. בחרו K המתאים לתיאור הדאטה והכינו את העמודת ה .features
2. גם ב Kmean אנו מחשבים מרחק אוקלידי ולכן מומלץ לנרמל את הנתונים לפני הרצת המודל.
3. בנו מודל Kmeans למציאת הקבוצות (כנל: .(**random_state = 8123**

Clustering - Model Evaluation

4. הדפיסו את הקואורדינטות של מרכזי הקלאסטרים .
5. שמרו את הלייבלים של המודל בעמודה חדשה בשם "cluster" שתתווסף לפיצ'רים הלא מנורמלים.
6. הדפיסו את ממוצעי הפיצ'רים המנורמלים והלא מנורמלים ואפיינו בעזרת שניהם את כל אחד מהקלאסטרים. עבור איזה קלאסטר ערכי ה Magnesium גבוהים יותר ?
7. האם ביצועי המודל טובים ? (מותר להשתמש בידע המוקדם שלנו כדי לבדוק)
8. לבסוף הריצו מודל היררכי עבור הדאטה והדפיסו את דיאגרמת הדנדרוגרם. כמה קלאסטרים התקבלו לדעתכם עפ"י התרשים ?