מכון טכנולוגי חולון
Holon Institute of Technology

50058 מדע נתונים- תאוריה ומעשה | סמסטר קיץ תשפ״א

עבודה מספר 1- ניתוח נתוני רעילות של פטריות

בעבודה זו נתמקד בניתוח נתוני רעילות של פטריות. לצורך כך נעזר בסט הנתונים אשר פורסם באתר Kaggle-
https://www.kaggle.com/uciml/mushroom-classification

המטרה של הפרויקט היא לבחון אילו מאפיינים קשורים לרעילות של הפטריות השונות והאם ניתן לדעת, על סמך המאפיינים, האם מדובר בפטריה רעילה, אם לאו.

הנחיות:

סקירת נתונים (data exploration)-

א. הורד את סט הנתונים מהאתר וקרא אותם לתוך data frame.
ב. סקור את המאפיינים השונים שבבסיס הנתונים (אילו מאפיינים הם נומרים, אלו אלפא-נומריים, תאריכים וכו׳).
ג. בצע קידוד (encoding) למאפיינים האלפא-נומריים בשתי שיטות: label encoding ו-one-hot encoding.
ד. שרטט גרף של התפלגות הערכים של המאפיינים השונים (ניתן, לדוגמא, להיעזר ב״תרשים קופסה״ (box plot)).
ה. בדוק אם ישנם ערכים חסרים. אם כן, בדוק מה האחוז שלהם בכל אחד מהמאפיינים, והשלם אותם בהתאם לצורך (ככל שניתן).
ו. חשב והצג את הקורלציות ההדדיות בין המאפיינים (ראה דוגמא: https://datatofish.com/correlation-matrix-pandas/). עבור המאפיינים הקטגוריאליים, יש לבצע את הקורלציות על הקידודים הנומרים.
ניתן (אך לא חובה) להיעזר גם בהסברים בקישור הבא: https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9
ז. זהה את המאפיינים שנחזים להיות הכי קשורים לרעילות.

פיתוח המודלים (model development)-

ח. חלק את סט הנתונים לשלושה חלקים- training set (70%), validation set (15%), test set (15%).
ט. השתמש בשניים מבין המסווגים שלמדנו, בכדי להעריך את רמת הרעילות של הפטריות. בחן והשווה את ביצועי המסווגים.
י. בחן את השפעת שיטת הקידוד (label encoding vs. one-hot encoding) על ביצועי המסווגים.

מכון טכנולוגי חולון
Holon Institute of Technology

הערות:

- ההגשה יכולה להתבצע באחת מבין שתי אפשרויות:

א. הגשת דו"ח הכולל פירוט (תרשים + הסבר תמציתי) של הנעשה, תוצאות, ניתוח תוצאות ומסקנות. יש להגיש את הדו"ח ב**קובץ PDF אחד בלבד** בהתאם להנחיות.

ב. הגשה במתכונת של מחברת Jupyter, ובלבד שהמחברת כוללים הסברים של מה שנעשה, ניתוח התוצאות והמסקנות. במקרה כזה, יש לעדכן בקישור שיפורסם לצורך כך, את הקישור למחברת שלכם ב- Google colab או GitHub ולהעלות את המחברת עצמה למודול.

- עיקר העבודה הוא ההסברים שלכם- הסבר תמציתי על הנעשה, ניתוח תוצאות ומסקנות.
- התרגיל יוגש ביחידים או בזוגות, באמצעות המודל בלבד. בהגשה בזוגות, מצופה מכל אחד מבני הזוג לשלוט בכל נדבכי העבודה והדו"ח. רק אחד מבני הזוג יגיש את העבודה במודל. יש לרשום שמות + מס' תעודות זהות בראש העבודה.
- עבודות דומות תיפסלנה ויינקטו צעדיים משמעתיים.
- מומלץ מאוד לבצע את העבודה באמצעות שפת פייתון.
- לוח הזמנים להגשה- בהתאם למוגדר במודל.

בהצלחה!

צוות הקורס