מטלה 2 : מבוא למדעי הנתונים

כללי:

- מטלה זאת תכלול 3 שאלות בנושאים שונים אשר נלמדו בכיתה במהלך הקורס.
- הפרויקט מחייב תיעוד מפורט של הפעולות אשר נעשו בכל שלב לטובת בדיקה איכותית וקבלת ביקורת בונה ומלמדת.
- שימו לב לעקוב אחר ההוראות בכל שאלה.

קווים מנחים:

- יש להגיש מחברת jupyter מסודרת הכוללת כותרות, הסברים מילוליים ותיעוד של הקוד.
- שאלות על הפרויקט ייענו בפורום בלבד
- בבדיקה יינתן דגש על יעילות נראות הקוד והחשיבה האנליטית שלכם במהלך התרגיל, שימו לב!
- זמן הריצה של המחברת כולה לא יעלה על שעה.
- השתמשו בתרשימים וגרפים בשלבים השונים של התרגיל.
- אנא וודאו שמחברת ההגשה רצה עם כל בסיסי הנתונים כאשר הם באותה תיקיה ללא ניתובים שונים

הגשה:

- מועד הגשה 20.2.20 – במידה ויש בעיות חריגות אנא עדכנו מראש
- הפרויקט יוגש בזוגות , יש להירשם לקבוצות במודל.
- מייל עוזר ההוראה : mr.marudi@gmail.com

בהצלחה!!!

שאלה 1 – Explainability AI

1. מידע על בסיס הנתונים:

i) בסיס הנתונים לתרגיל נקלח מאתר UCI :
https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

ii) לבסיס הנתונים המקורי הורדנו את עמודת "communityname string" הכוללת תווים

iii) לבסיס הנתונים המקורי הוספנו כותרות לעמודות לנוחיותכם.

iv) הקובץ "Communities and Crime Data Set.names " מכיל מידע על בסיס הנתונים והפיצ'רים השונים.

2. Preprocessing:

i) קובץ הנתונים מכיל ערכי null , עליכם לטפל בהם בכל צורה שתבחרו
(מכוון שהתעסקנו בהשלמת נתונים במטלה הראשונה אתם רשאים לבצע כל השלמה שתרצו גם הבסיסית ביותר ולהסביר בקצרה).

ii) במידה ותחליטו לבצע פעולות נוספות על בסיס הנתונים אנא ציינו אותם במפורש והסבירו מדוע בחרתם לבצע אותן.

3. SHAP:

i) בחרו מודל רגרסיה ואפטמו אותו כך שיחזה בצורה המיטבית את תוצאת המודל. הראו את המדד עליו התבססתם.
טיפ : מומלץ להשתמש ב GridSearch
(יש אפשרות להשתמש גם במודל קלסיפיקציה אך עליכם לטפל בעמודת הlabel בהתאם על ידי חלוקה לbin)

ii) הסבירו בקצרה את המושג shap value ומה המספר מייצג

iii) Global interpretability : הציגו summary_plot והסבירו את פיזור הנקודות עבור שלושת הפיצ'רים החשובים ביותר עם דגש על כיוון הצבעים (אדום וכחול) וכיצד הם משפעים על ערך המטרה. האם ההסבר עולה בקנה אחד עם המשמעות מאחורי הפיצ'ר כפי שמופיע בקובץ המידע ?

iv) Local interpretability: הגרילו 3 שורות בצורה רנדומלית והסבירו את תוצאת המודל עבורן. הציגו את הפלטים המתאימים מתוך החבילה

4. LIME:

i) הריצו את אלגוריתם lime על שלושת השורות שהגרלתם בחלק הקודם והסבירו את תוצאת המודל

ii) האם שני האלגוריתם חזו תוצאה דומה? נסו להסביר את הדומה והשונה בין התוצאות.

שאלה 2 – imbalanced dataset

1. מידע על בסיס הנתונים:

i) HTRU2 הוא בסיס הנתונים המכיל עבור כל שורה אינדקס האם מדובר בכוכב פולסאר או לא.
קישור ומידע נוסף על בסיס הנתונים : קישור

2. מטלה:

i) עליכם לחלק את הדאטה בצורה רנדומלית כך ש10% ישמר לוולידציה ו90% לאימון ובדיקת המודל

ii) השתמשו ב -4 שיטות הדגימה שנלמדו בכיתה כדי להתמודד עם imbalanced dataset והריצו מודל חיזוי מתאים:
(1) Under sampling
(2) Over sampling
(3) SMOTE
(4) ADASYN
(5) Combine approach - בונוס של עד 5 נקודות לשימוש בשיטה זאת

iii) עבור כל שיטות הדגימה הסבירו את עיקרי השיטה והציגו כיצד היא שינתה את בסיס הנתונים (כמות הדגימות לכל class, כמות הדגימות בכללי וכו...)

iv) הציגו השוואה בין תוצאות המודלים השונים על גבי סט הוולידציה, בחרו מדדים מתאימים לבעיה והסבירו מדוע בחרתם אותם.
דגש: לבעיות אלו קיים סט מדדים מתאים, הקפידו להשתמש בהם
בונוס 3 נקודות : בחרו מדד לא מתאים והסבירו מדוע הוא מוטעה בבעיות אלו.

v) חוו דעה: איזה מדד מתאים יותר לבעיות מסוג זה F1 score או F2 score

שאלה 3 – Clustering

3. מידע על בסיס הנתונים (כמו בשאלה 2):

i) HTRU2 הוא בסיס הנתונים המכיל עבור כל שורה אינדקס האם מדובר בכוכב פולסאר או לא. קישור ומידע נוסף על בסיס הנתונים :

ii) במהלך המטלה הקפידו להוריד את שורת ה-target- והשתמשו בה רק לווליידציה של המודל

4. מטלה:

i) עליכם להריץ שני מודלים לclustering לבחירתכם ולבצע את הפעולות הבאות:

(1) בצעו שימוש בנרמול לבחירתכם וטרנספורמציה של הדאטה לשני ממדים על ידי PCA.

(2) מצאו את מספר הקבוצות האופטימלי, הקפידו להסביר את הפעולות שביצעתם ולהציג גרפים מתאימים

(3) השתמשו בסט המדדים כפי שנלמד בכיתה לבעיות מסוג זה על מנת לבחון את המודלים שלכם ולבצע השוואה של התוצאות.

(4) הציגו את תוצאות המודל על גרף אחד כך שכל class יוצג בצבע אחר

.5

בהצלחה !