משימת תחרות במדעי הנתונים: חיזוי נטישת לקוחות

מטרה:

במשימה זו תתבקשו לחזות נטישה של לקוחות.

מטרת התחרות היא להשיג את ציון F1 הגבוה ביותר בחיזוי נטישת הלקוחות בסט בחינה (holdout).

$$\boldsymbol{F1 = \frac{2 \times TP}{2 \times TP + FP + FN}}$$

תיאור מקרה:

נטישת לקוחות, מבטאת תופעה של לקוחות שמפסיקים לצרוך שירותים מחברה מסוימת (למשל, מבטלים מינוי לחדר כושר או עוברים לסניף בנק אחר). חיזוי הנטישה משמעותית במגוון תעשיות, במיוחד בתחומים כמו טלקומוניקציה, שירותים ופיננסים. חיזוי נטישה מדויקת מאפשרת לחברה לאתר "עריקים" פוטנציאליים מבעוד מועד ולנקוט בפעולות מתאימות כדי לשמר אותם, מה שמוביל בסופו של דבר להגברת נאמנות הלקוחות ולרווחים מוגדלים.

תיאור הדאטה:

סט אימון: מערך נתונים זה מורכב מפרופילי לקוחות, דפוסי ההתנהגות שלהם וסטטוס הנטישה שלהם. בדאטה זה תוכלו להשתמש בכדי לאמן ולבחון את המודלים שלך.

סט בחינה: מערך נתונים זה מכיל פרופילי לקוחות ודפוסי ההתנהגות שלהם, אך לא את מידע הנטישה. מטרתכם היא ליישם את המודלים שאימנתם על בסיס הנתונים הזה כדי לחזות אילו לקוחות צפויים לנטוש את החברה.

תיאור משתנים בדאטה:

- **האם הלקוח נטש** את החברה בחודש האחרון - העמודה נקראת Churn
- **שירותים שכל לקוח נרשם אליהם** - טלפון, קווים מרובים, אינטרנט, אבטחה מקוונת, גיבוי מקוון, הגנת מכשירים, תמיכה טכנית והזרמת טלוויזיה וסרטים
- **פרטי חשבון לקוח** - כמה זמן הוא היה לקוח, סוג החוזה, אמצעי תשלום, חיוב ללא נייר, חיובים חודשיים וסך החיובים
- **מידע דמוגרפי על לקוחות** - מגדר, טווח גילאים, ואם יש להם חברי משפחה אחרים התלויים בחשבון הזה

הערה: העמודה Churn לא תהיה זמינה בסט הבחינה.

פרטי המשימה:

1. **עיבוד מוקדם של נתונים**: בצעו שלבי עיבוד מקדים של נתונים הדרושים, הכוללים טיפול בערכים חסרים, קידוד משתנים קטגוריים, קנה מידה וכו'.

2. **אימון מודלים**: בניית מודלים באמצעות אלגוריתמים שנלמדו בכיתה.

3. **הערכת מודל:** העריכו את המודלים באמצעות מדדים מתאימים.

4. **חיזוי מודל:** החלו את המודלים המאומנים על מערך האימות וחזה את נטישת הלקוחות במסד זה.

פורמט הגשה:

ההגשה הסופית צריכה להיות מורכבת משני קבצים:

1. קובץ CSV עם תחזיות הנטישה שלך בערכת האימות.
   הקובץ צריך להכיל שתי עמודות בלבד: Churn_Prediction , CustomerID
2. קוד r/py מתועד שמייצר את קובץ הפרדיקציה

- ייבדקו רק עבודות בהן הוגשו הקבצים **בפורמט הנכון**

הערות כלליות

-1 הציון יתבסס אך ורק על מדד F1 בין תחזית הנטישה שלכם לבין הנטישה בפועל בסט האימות.
ולא על בסיס איכות הקוד

-2 עם זאת, הקוד חייב לרוץ ולייצר את התוצר שהגשתם

-3 הציון יהיה יחסי לשאר הסטודנטים, והממוצע יהיה 80.
זוהי תחרות פרדיקציה! אין לשתף פעולה עם סטודנטים אחרים.

בהצלחה!!