הפקולטה להנדסה

המחלקה להנדסת תעשיה וניהול

מבוא לניתוח נתונים 4212410

סמסטר קיץ תשפ"ב

מועד ב' 9.11.2022

ד"ר ליהי דראי

משך הבחינה: 2 שעות

המבחן בכיתת מחשב ללא אינטרנט

חומר עזר מותר: כל חומר מודפס. חומר דיגיטלי על דיסק-און-קי

חומר עזר אסור: מכשירים דיגיטליים (לפטופ, טאבלט, שעון, טלפון, מחשב)

אין להשתמש או להגיש מחברות

הנחיות לסטודנטים:

- במודל תמצאו: קובץ פייתון (test_b.ipynb) וקובץ דאטה (data_file_b.csv)
- העתיקו את שני הקבצים למחשב, ושימו אותם באותה התיקייה.
- כתבו את התשובות בתוך קובץ הפייתון (test_b.ipynb).
- יש להגיש את הקובץ test_b.ipynb בלבד. טופס הבחינה ומחברת הבחינה לא יבדקו.
- הגשת המבחן – דרך מתזמן הבחינות בלבד.
- אין להשתמש בלולאות for או בספריות שלא נלמדו בקורס.
- ניתן להשתמש בכל הפקודות מהספריות: pandas, numpy, matplotlib, seaborn, sklearn
- עבור כל השאלות, יש להראות בקוד ביצד הגעתם לתשובה. פתרון ללא קוד לא יבדק.

בהצלחה!!!

הקובץ data_file_b.csv מכיל נתונים על לקוחות של חברת סלולר. להלן חלק מהשדות:

gender	מגדר (של הלקוח)
seniorCitizen	האם הלקוח אזרח ותיק (0=לא, 1=כן)
partner	האם ללקוח יש בן זוג
dependents	האם ללקוח יש ילדים
tenure	וותק בחברת הסלולר (בשנים)
techSupport	האם יש מנוי לתמיכה טכנית
paymentMethod	שיטת חיוב
monthlyCharges	חיוב חודשי
totalCharges	סך כל החיובים

<u>שאלה 1 (25%)</u>

א. צרו טבלה המראה עבור כל שיטת חיוב (paymentMethod) את הממוצע והחציון של החיוב החודשי (monthlyCharges) ואת מספר החיובים סה"כ (totalCharges).

ב. מה היא שיטת החיוב (paymentMethod) שאזרחים וותיקים (seniorCitizen) משתמשים בה הכי הרבה? יש להראות בקוד כיצד הגעתם לתשובה.

<u>שאלה 2 (25%)</u>

הציגו שני איורים אחד ליד השני:

א. האיור הראשון צריך לאפשר לענות על השאלה: מהו התפלגות החיוב החודשי (monthlyCharges), באחוזים, עבור אזרחים ותיקים ולא ותיקים (seniorCitizen)? יש לדאוג שציר הערכים בציר הX יהיו ב log.

ב. האיור השני צריך לאפשר לענות על השאלה: האם אזרחים ותיקים (seniorCitizen) משתמשים בתמיכה הטכנית (techSupport) יותר ממי שאינם אזרחים ותיקים? אין להציג באיור מקרים בהם לא ידוע (unknown) האם יש תמיכה טכנית.

<u>שאלה 3 (25%)</u>

א. הציגו בטבלה את המאפיינים של שלושת קבוצות הלקוחות שנוטשים הכי הרבה (churn) את חברת הסלולר. השתמשו בארבעת המאפיינים הבאים בלבד: gender, seniorCitizen, partner, dependents.

***דוגמה אפשרית למה שניתן לראות בטבלה**: הלקוחות שנוטשים הכי הרבה הם גברים שהינם אזרחים ותיקים, יש להם בת זוג ואין להם ילדים. הלקוחות השניים שנוטשים הכי הרבה הם.. הלקוחות במקום השלישי שנוטשים הכי הרבה הם... אין צורך לכתוב במלל את התשובה. מספיק להציג את הנתונים.*

ב. לאילו מבין המאפיינים בסעיף א' (gender, seniorCitizen, partner, dependents) יש את הקורלציה הגבוהה ביותר עם נטישה (churn)? הציגו במפת חום.

<u>שאלה 4 (25%)</u>

א. צרו מודל המנבא האם הלקוח ינטוש (churn) על סמך המאפיינים הבאים: gender, seniorCitizen, partner, dependents.

ב. מהו הדיוק (באחוזים) של המודל שיצרתם?

ג. הציגו באיור את אחוז הלקוחות שנוטשים מול אלו שלא נוטשים (churn). אם לוקחים בחשבון את אחוז הלקוחות הנושטים (churn), מה ניתן להגיד על הדיוק שחישבתם בסעיף ב'?