לימוד מכונה 364-1-1811

**פרויקט - חלק א'**

זהו החלק הראשון מבין שני חלקים.

**הנחיות הגשה:** **דו"ח התרגיל הראשון וקוד ה-R/Python שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה-10.5.2019 בשעה 23:55. מספיקה הגשה של אחד מבני הזוג.**

מטרת התרגיל: בתרגיל זה נתרגל טעינת, עיבוד, הצגת וניתוח נתונים. פעולות אלה גם ישמשו אותנו בהבנת והכנת הנתונים לקראת התרגיל השני בו נשתמש בנתונים לאימון ובחינה של מערכות לומדות והשוואה ביניהן.

צוותי הגשה: הגשת התרגיל הינה בזוגות, בהתאם לקבוצות המוגדרות במודל.

דגשים לדו"ח: **אורך הדו"ח לא יעלה על 8 עמודים** (לא כלל עמודים נלווים כמו שער ותוכן עניינים), בגודל כתב 12, פונט Arial ורווח של שורה וחצי. **חריגה מהגדרות אלו תגרור הורדת נק'**. יש לשמור על תמציתיות ולהתמקד בתובנות המרכזיות שלכם בכל סעיף.

שפת תכנות: R או Python.

הסבר אודות בסיס הנתונים לפרויקט:

בסיס הנתונים מיועד לסיווג מחלת לב אצל מטופלים (האם מדובר במחלת לב או לא).

הסבר אודות בסיס הנתונים ב-**Moodel**:

**זהו בסיס הנתונים שישמש אתכם לפרויקט.**

סט הבחינה

X\_test.csv – משמש לטובת הגשת החיזוי הסופי. מכיל את המשתנים מסבירים בלבד של סט הבחינה. בחלק זה אינכם צריכים להגיש עדיין חיזויים, אבל הקובץ פורסם כדי שתוכלו לראות איך נראות הרשומות עבורן יש לבצע חיזוי (זה חשוב בעיקר לטובת ההחלטה איך להתמודד עם נתונים חסרים וכד').

הסבר על המשתנים בבסיס הנתונים:

id – patient identification number

age (in years)

gender (0 = female, 1 = male)

cp (chest pain type) – 0 – typical angina, 1 – atypical angina, 2 – non-anginal pain, 3 -asymptomatic)

trestbps (resting blood pressure (mm/Hg)

chol - serum cholestoral (mg/dl)

fbs - fasting blood sugar (0 – below or equal to 120 mg/dl, 1 – else)

restecg – resting electrocardiographic results (0 – normal, 1 – having ST-T wave abnormality, 2 - showing probable or definite left ventricular hypertrophy by Este's criteria.

thalach – maximum heart rate achieved.

exang – exercise induced angina (0 - no, 1 – yes)

oldpeak – ST depression induced by exercise relative to rest.

slope – the slope of the peak exercise ST segment (0 – upsloping, 1 – flat, 2 – downsloping).

ca – number of major vessels (0-3, 4 – unknown)

thal – (0 – missing value, 1 – fixed defect, 2 – normal, 3- reversable defect)

y (1 – heart attack, 0 - else)

הקדמה:

* **ההוראות בתרגיל זה הינן כלליות (כלומר לא נכתבו ספציפית עבור בסיס הנתונים הנתון) ומטרתן להנחות אתכם כיצד יש לנתח ולעבוד עם כל בסיס נתונים שהוא. ייתכן כי שאלה מנחה כלשהי המופיעה בהוראות לא רלוונטית או ניתנת למענה עבור בסיס הנתונים הנתון. במקרה כזה, הסבירו מדוע לא ניתן/רלוונטי לענות עליה.**
* בתרגיל יושם דגש על שימוש בתוכנה לצורך מענה על שאלות הקשורות בנתונים ובניתוחם. לא פחות חשובות הן התובנות משימוש זה לגבי עולם התוכן של הבעיה הנחקרת, כשהשאלה המרכזית הינה: מה בעצם למדנו מתרגיל זה? יש לשלב טבלאות, גרפים וכו', לנתחם ולהשליך מהם על עולם התוכן הנחקר.
* עבור כל פעולה שבוצעה בתוכנה יש גם לתעד: מדוע ביצענו אותה? מה התקבל בפלט? מה למדנו מהפלט? מהן ההשלכות על השלבים הבאים?
  + דוגמה: ע"ס ידע אישי, חשדנו שישנו קשר חזק בין שני משתנים🡨בחרנו להציג מתאם זה כדי לבחון את הקשר ביניהם🡨הגרף מראה מתאם גבוה🡨ניתן ללמוד כי שני המשתנים מתארים את אותה התופעה🡨נמחק את אחד מהמשתנים, כך שבשלב הבא נוכל להשיג מודל פשוט יותר.
* יש להגיש את כל קבציR ו/או Python עליהם עבדתם.
* שאלותבנוגע לתרגיל, יש לפרסם בפורום הייעודי שייפתח במודל.

**מבנה העבודה:**

**הגדרת הבעיה (8 נק')**

1. **תיאור כללי של עולם התוכן הנחקר**

* מהי הבעיה המחקרית? מה עשו מחקרים קודמים שעסקו בנושא כדי להתמודד עם הבעיה?

1. **הגדרת שאלת המחקר**

* מה אנו מצפים לפתור בעזרת הכלים והשיטות של מערכות לומדות?

**הבנת הנתונים (53 נק')**

1. **תיעוד מקורות הנתונים ומשמעותם** (אם יש צורך, היעזרו במידע מהאינטרנט)

* מהו מקור הנתונים במאגר אתו אתם עובדים וכיצד הנתונים נוצרו (למשל: מדידות חיישנים, נתונים סטטיסטיים, ידנית, נתוני מומחה, וכו')?
* מהם המאפיינים המופיעים בסט הנתונים? מה משמעותם? מהו משתנה המטרה ואיזה ערכים הוא מקבל?

1. **הסתברויות אפריוריות וקשרים בין מאפיינים**

* מהן ההסתברויות האפריוריות של משתנה המטרה ושל שאר המאפיינים? על מה מלמדות הסתברויות אלה? עבור משתנים רציפים ניתן להשתמש בהיסטוגרמה.
* האם סט הנתונים מאוזן? האם לדעתכם הוא מייצג את המציאות?
* חפשו והציגו קשרים "מעניינים" בין מאפיינים - צפויים ולא-צפויים. הסבירו המשמעות של קשרים אלה.
* מהם המאפיינים בהם ניתן "לחשוד" כבעלי השפעה על משתנה המטרה (ע"ס ידע אישי/מוקדם, מחקרים, ניסיון בתחום וכו'), ולפיכך כנראה יבחרו ע"י המערכות הלומדות שנבחן בתרגיל השני?

1. **איכות הנתונים**

* האם ישנם נתונים חסרים? אם כן, מה ניתן לומר על עליהם?
* האם ישנם נתונים שאינם הגיוניים (למשל: גיל>250)? מה אתם מציעים לעשות עם נתונים אלה?

אם יש לכם תובנות נוספות לגבי הנתונים, זה המקום להציגן.

**הכנת הנתונים (32 נק')**

1. **על פי הצורך, בצעו ונמקו בחירת מאפיינים שביצעתם**

* השמטת מאפיינים "רועשים" או חסרי חשיבות.
* השמטת מאפיינים בעלי איכות נמוכה מדי (שגיאות, ערכים חסרים וכו').
* השמטת תצפיות בעלות חוסר רב.
* התמקדות בפלחי אוכלוסייה נבחרים.

1. **על פי הצורך, תנו טיפול פרטני במאפיינים**

* דיסקרטיזציה של משתנים רציפים, למשל בצורה שמייצגת את ההתפלגות או בצורה שמייצגת דרישות של עולם התוכן.
  + - דוגמה: משתנה רציף "גיל" ניתן להפוך למשתנה בדיד ע"י הגדרת סיפים כגון 6, 18, 21, 65 וכו'.
* גזירת מאפיינים חדשים (פונקציות של משתנים קיימים).

**איכות הדו"ח ורמת שימוש בתכנת R (7 נק')** – נק' אלו ינתנו בהתאם להערכה כללית בנוגע לאיכות הדו"ח ובהתאם לרמת השימוש בתכנת ה-R (האם נעשה שימוש מקיף/חלקי/כלל לא).

בהצלחה !