פרויקט - חלק א'

זהו החלק הראשון מבין שני חלקים.

הנחיות הגשה: דו"ח התרגיל הראשון וקוד ה-Python שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה-25.12.2024 בשעה 23:59. 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 בהתאם לקבוצות המוגדרות במודל.

דגשים לדו"ח: **אורך הדו"ח לא יעלה על 12 עמודים** (לא כולל 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: Python.

הסבר אודות בסיס הנתונים ב-**Moodel**:

באתר הקורס במודל ישנם שני בסיסי נתונים עם משימות לימוד שונות (Diabetes,Car Insurance). בחרו בסיס נתונים כרצונכם והירשמו בגיליון שבמודל. שימו לב שהביצועים בפרויקט שלכם יימדדו ביחס לשאר הקבוצות שבחרו באותו בסיס הנתונים.

בכל תיקיית נתונים ישנם 3 קבצים:

קובץ וורד עם שם בסיס הנתונים – מכיל הסבר קצר על בסיס הנתונים והעמודות הקיימות בו.

Xy_train.csv – סט האימון שכולל את המשתנים המסבירים ומשתנה המטרה, בו יש להשתמש בחלק זה של העבודה. איתו גם תאמנו את המודלים השונים בחלק ב'.

X_test.csv – משמש לטובת הגשת החיזוי הסופי. מכיל את המשתנים המסבירים 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' – התהליך צריך להיות זהה להשלמה בסט האימון).

הקדמה:

- **ההוראות בתרגיל זה הינן כלליות (כלומר לא נכתבו ספציפית עבור בסיס הנתונים הנתון). מטרתן להנחות אתכם כיצד יש לנתח ולעבוד עם כל בסיס נתונים שהוא. ייתכן כי שאלה מנחה כלשהי המופיעה בהוראות לא רלוונטית או לא 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. על כל הגרפים להכיל כותרות ואת שמות הצירים / המשתנים המוצגים בהם בצורה ברורה ואסתטית.

- 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 קבצי ה Python עליהם עבדתם.
- שאלות בנוגע לתרגיל, יש לפרסם בפורום הייעודי שייפתח במודל.

<u>מבנה העבודה:</u>

<u>הגדרת הבעיה (8 נק')</u>

1. **תיאור כללי של עולם התוכן הנחקר**

- מהי הבעיה המחקרית? מה עשו מחקרים קודמים שעסקו בנושא כדי להתמודד עם הבעיה? (אפשר בקצרה).

2. **הגדרת שאלת המחקר**

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

<u>הבנת הנתונים (50 נק')</u>

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')?
- עבור כל משתנה בסט הנתונים רשמו מה משמעותו בקצרה, ומה סוג המשתנה (רציף, בדיד, קטגוריאלי ניתן לסידור, קטגוריאלי לא ניתן לסידור...). מהו משתנה המטרה ואילו ערכים הוא מקבל?

2. **הסתברויות אפריוריות וקשרים בין מאפיינים**

- הראו מהן ההסתברויות האפריוריות של משתנה המטרה, ושל עוד 3 משתנים רציפים ו3 משתנים קטגוריאליים לפחות. כתבו במשפט קצר מה מלמדות הסתברויות אלה. עבור משתנים רציפים ניתן להשתמש בהיסטוגרמה ועבור קטגוריאליים בגרף עמודות.
- האם סט הנתונים מאוזן? האם לדעתכם הוא מייצג את המציאות?

- חפשו והציגו לפחות 3 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 ניתן להניח שיבחרו ע"י המערכות הלומדות שנבחן בתרגיל השני?
- עבור משתנים אלו, בדקו את ההנחה בעזרת גרפים או סטטיסטיקה תיאורית, והציגו את התוצאות.
- האם יש משתנים שאתם מניחים שניתן להסיר ע"ס ידע אישי? אם כן, בדקו את ההנחה.

3. **איכות הנתונים**

- האם ישנם נתונים חסרים? אם כן, מה ניתן לומר על עליהם? מה אתם מציעים לעשות איתם?
- 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.

<u>הכנת הנתונים (30 נק')</u>

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.
- השלמת מושכלת של ערכים חסרים במידה ואפשרי.

2. **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'. חשוב לנמק מדוע השתמשתם ומה היתרונות/חסרונות של זה.
- גזירת מאפיינים חדשים (פונקציות של משתנים קיימים).

<u>איכות הדו"ח ורמת שימוש בתכנת Python (7 נק')</u> – נק' אלו יינתנו בהתאם להערכה כללית בנוגע לאיכות הדו"ח ובהתאם לרמת השימוש בתכנת Python (האם נעשה שימוש מקיף/חלקי/כלל לא).

<u>איכות התרשימים והטבלאות (5 נק')</u> – כותרות, שמות צירים ושנתות, אסתטיקה.

ניתן לשאול שאלות הקשורות לפרויקט הקורס בפורום השאלות הייעודי שנפתח

בהצלחה!