מבוא ל – Science Data – עם Python למנהלים
2024-2025
תרגיל 3

<u>הבהרה כללית</u>
ההנחה היא כי כל המטלות המוגשות לרבות מטלה זו, הן יצירות של הסטודנטים <u>החברים בקבוצה בלבד</u>. פלגיאט (גניבה ספרותית) מתייחס להצגת המילים, הקוד, הרעיונות ו/או העבודה של אחרים (בני אדם או מחשבים) כאילו זו העבודה שלנו. לא מקובל ולא ראוי להגיש עבודה שלא נכתבה על ידכם. העתקה והדבקה של מאמרים שפורסמו, דוחות, מסטודנטים אחרים, אתרי אינטרנט או מקורות כתובים אחרים, נחשבים פלגיאט אלא אם כן מוזכר במפורש המקור סופקה התייחסות מלאה אליו כולל מספר עמוד וביבליוגרפיה מלאה, כמו גם הסבר לנחיצותו של כל מקטע קוד. פלגיאט חשוד יטופל בקפדנות ובהתאם לכללי האוניברסיטה.

<u>מטרת התרגיל</u>
התנסות מעשית בעבודה עם דאטה, ביצוע ניתוחים והפקת תובנות. בתרגיל זה תבחרו דאטה, תציעו שאלות מעניינות שניתן לענות עליהן בעזרת הדאטה, תבצעו אנאליזה ראשונית ותתנו אינטרפטציה לתוצאות, תגדירו ותבנו מודל קלסיפיקציה מבוסס רגרסיה לוגיסטית, תבצע אנליזה סמנטית ותציגו את הממצאים בצורה ויזואלית ומילולית בעזרת Python. הדגש הוא על היכרות מעמיקה עם הנתונים (לרבות גרפים של קשרים ושכיחות בחיתוך קטגוריות), בניית שאלה טובה, הכנת הדאטה לניתוח, ביצוע קלסיפיקציה, בדיקת הביצועים של מודל הקלסיפיקציה, ביצוע אנליזה סמנטית ותקשור טוב של תהליך העבודה שלכם, התוצאות והתובנות שעלו מכל שלב ושלב.

<u>הגשה:</u> בזוגות.
כל זוג סטודנטים יגיש את התרגיל המסכם כלינק קולאב במודל (מספיק להגיש פעם אחת). יש לכתוב בראשית הקולאב את שמות הסטודנטים המגישים + מספרי הזהות.
הקולאב יכלול שילוב של קוד, פלט של כל פקודה והסברים אודות תהליך העבודה, התוצאות והמסקנות.
קטעי הטקסט יכולים להופיע תחת תיבת טקסט, בתוך תא קוד (שימוש ב #), או כתמונה מצורפת מתוך מסמך word- העיקר שיהיה קל וברור לעקוב!

<u>מועד הגשה</u>:
עד ה-16.2.2025 בחצות. **לא תתקבלנה הגשות באיחור** (פרט למילואימניקים מוכרים וכו).

<u>הנחיות כלליות</u>:
- העבודה על כל השלבים תעשה בפייתון בעזרת שיטות שנלמדו במהלך הקורס.
- קוד- הדרישה היא לקוד מלא שכולל את כל שלבי העבודה (מדאטה גולמי ועד לתוצאות הסופיות) ולכלל את כלל הפלט (גם עבור שלבי ביניים).
- שימוש בקוד שלא נלמד בשיעור/תרגול דורש הסבר על תוכן הקוד ונחיצותו/תרומתו.
- יש לוודא שהקוד רץ במלואו מתחילתו ועד סופו וההגשה כוללת את הפלט של כל פקודה ופקודה. עבודה שבה הקוד לא ירוץ במלואו תיפסל.

<u>הנחיות מפורטות:</u>
1. **בחירת דאטה-** יש לבחור דאטה מתוך תקיית ההגשה במודל. כדי לבחור, עברו על כל האופציות וחישבו מה יהיה לכם הכי מעניין לעבוד איתו.
2. **סקירה כללית של הדאטה (10 נקודות)**

ערכו סקירה קצרה וכיתבו- מה מבנה הדאטה (כמה טורים, תצפיות, איזה סוגי משתנים, מה המשמעות של כל טור, האם יש חוסרים וכו׳). מה המשמעות של תצפית, איזה טורים רלוונטיים כנראה לשאלות המחקר ומה המשמעות הספציפית של כל טור (חיוני לפרט במידה ויש קידוד של תשובות למשל, 1=כן, 2 = לא וכדומה). אל תשכחו להסתכל בדוקומנטציה. שימו לב שהקוד והפלט הינם חלק אינטגרלי מהסקירה.

3. **הגדרת שאלת מחקר מרכזית (5 נקודות)**- קלסיפיקציה בעזרת רגרסיה לוגיסטית
הסבירו בבקשה מה אתם מנסים לחזות ואיזה פיצ׳רים (טורים) אתם משערים יכולים לשמש אתכם כדי לערוך את הקלסיפיקציה ומדוע.

4. **אנאליזה ראשונית והכנת הדאטה לניתוח (30 נקודות)**
בצעו אנליזה ראשונית (exploratory analysis) כדי להבין את הדאטה טוב יותר ולבחון את מידת הרלוונטיות של הפיצ׳רים השונים לקלסיפיקציה.
- השתמשו בכלים כגון ניתוחי שכיחויות, קורלציה, ויזואליזציה וכל כלי אחר שלמדנו שיכול לעזור לכם להבין טוב יותר את הדאטה. למשל, בחנו את הקשרים בין המשתנים השונים, נסו לזהות טרנדים וכו בעזרת סטטיסטיקות תאוריות, סטטיסטיקות לפי קבוצות, ויזואליזציה, וגרפים בחיתוך לפי קבוצות (שימו לב לבחור תצוגה גרפית מתאימה וכו׳).
- יש לכלול הסברים מפורטים על ההיגיון מאחורי כל שלב, להסביר את התוצאות ולהתייחס לתובנות שעלו מן הניתוח הראשוני (וכיצד הן משפיעות על שלב הניקוי ועל הקלסיפיקציה).
- שימו לב למקרים בהם חלוקה לקבוצות או הגדרה מחדש של משתנים עוזרת לייצר תובנות רלוונטיות.
- נקו והכינו את הדאטה לניתוח. שלב זה יכלול למשל גם איחוד של קבוצות במשתנים קטגוריאליים או רציפים, יצירת טורים חדשים, הסרת טורים לא רלוונטיים, קידוד מחדש של קטגוריות, התייחסות לחוסרים, נרמול של הטורים וכו׳. זכרו להסביר מדוע הסרתם טורים מסוימים/תצפיות מסוימות או ערכתם שינויים בדאטה.

5. **אנליזה סמנטית (13 נקודות)**
בצעו אנליזה סמנטית למשתנים המתאימים מתוך הנתונים כפי שנלמד בקורס. לאחר ביצוע האנליזה הסמנטית, חשבו על ניתוח מעמיק נוסף העושה שימוש בתוצאות אנליזה זו. בצעו את הניתוח, הציגו את התוצאות בתיבת טקסט וספקו את הקוד עבור הניתוח. נתחו את התוצאות בטקסט ובעזרת גרפים, וכיתבו את התובנות הרלוונטיות. האם התוצאות תאומו את ציפיותיכם?

6. **חידוד או עדכון של שאלות המחקר (2 נקודות)**
במידת הצורך, בעקבות האנאליזות שכבר ערכתם, חדדו את שאלת המחקר, וראו אם נדרשת הוספה או הורדה של טורים בעקבות היכרות עם הדאטה

7. **ביצוע קלסיפיקציה + הערכת ביצועים (15 נקודות)**
בצעו רגרסיה לוגיסטית בעזרת הכלים שנלמדו בקורס והעריכו את הביצועים של המודל עם קרוס ולידציה. חשבו האם בעקבות התוצאות שקיבלתם ניתן לשפר ולטייב את המודל ופעלו בהתאם.

8. **דיון (15 נקודות)** הצגת דיון אודות התובנות שעלו מן הניתוח.
בדיון התייחסו בבקשה גם ל: 1) האם הביצועים של המודל טובים מספיק לדעתכם לשימוש המיועד? פרטו ונמקו תוך התייחסות לסוג הטעות שנראה לכם חשוב למזער. הערה: אין הכרח שביצועי המודל יהיה טובים, נא הפעילו שיקול דעת וביקורת. 3) הסבירו את תוצאות האנליזה הסמנטית שביצעתם. 4) האם עולות סוגיות אתיות כלשהן מהשימוש בדאטה או במודל הזה? 5) מהן מגבלות הניתוח והממצאים 6) הציעו כיוונים ורעיונות למחקר המשך פרקטי (למשל, בהתבסס על התוצאות שלכם איזה מחקרי המשך מעניין לבצע ואיזה דאטה היה יכול לעזור לכם לקחת את הניתוח לרמה הבאה).

9. **הערכה כללית (10 נקודות)**
בהירות, איכות כתיבה, בהירות וכדומה. שימו לב שלא להשאיר סתם ניתוחים או גרפים שלא תורמים להבנת הנתונים, הקלסיפיקציה או תהליך העבודה שלכם (לפעמים כל המוסיף גורע...)