**תרגיל: ניתוח פינגווינים (אבל אל תדאגו , אף פינגווין לא יפגע)**

**מטרת התרגיל**

1. להבין את תהליך ניתוח הנתונים (EDA) ולטפל בנתונים.
2. להתנסות בבנייה, כוונון והערכה של מודלים שונים.
3. לנתח את תוצאות המודלים ולהסיק מסקנות.

* **בתרגיל זה יש להשתמש ב random\_state =42 בכל מקום על מנת שיתקבלו תוצאות זהות לכולם**

**שלב 1: טעינת הנתונים ובדיקה ראשונית**

1. טען את נתוני הפינגווינים באמצעות הפונקציה sns.load\_dataset('penguins')
2. הצג את השורות הראשונות והאחרונות של הנתונים.
3. בדוק אם קיימים ערכים חסרים בנתונים.

**שלב 2: זיהוי סוגי עמודות ו EDA ראשוני**

1. בדוק אילו עמודות הן קטגוריאליות ואילו עמודות הן נומריות.
2. זהה אילו עמודות דורשות המרה באמצעות OneHotEncoder ואילו דורשות עיבוד באמצעות StandardScaler. (אל תבצע עדיין המרות )
3. החלף את הערכים החסרים בעמודות ['bill\_length\_mm','bill\_depth\_mm','flipper\_length\_mm','body\_mass\_g'] לחציון
4. החלף ערכים חסרים בעמודת sex לשכיח
5. בדוק האם יש חריגים (outliers ) בנתונים בעזרת חישוב IQR

**שלב 3: ניתוח ראשוני עם העמודות הנומריות בלבד**

1. הגדר את העמודות הנומריות כמאפיינים (X) ואת עמודת target כעמודת המטרה (y).
2. חלק את הנתונים לסט אימון וסט בדיקה (80/20).
3. בנה מילון (dictionary) עם שלושת המודלים הבאים:
   * Logistic Regression עם max\_iter=200
   * Decision Tree עם random\_state=42
   * Random Forest עם random\_state=42

דוגמא לאיך לבנות dictionary:

models = {

'Logistic Regression': LogisticRegression(max\_iter=200),…

1. בנה לולאה שרצה על המילון ומריצה את המודלים( על הנתונים הנומריים בלבד בשלב זה), מפיקה דוח דיוק (accuracy report) )ומציגה את תוצאות המודלים.

* האם התקבלה אזהרה? מה היא אומרת?

**שלב 4: הוספת StandardScaler**

1. בצע Scaling לעמודות הנומריות באמצעות StandardScaler.
   * השתמש ב-fit\_transform על סט האימון וב-transform על סט הבדיקה.
2. הרץ את הלולאה של המודלים שוב והשווה את התוצאות לתוצאות הקודמות.

* מה קרה לאזהרה שהיתה שם קודם?

**שלב 5: One Hot Encoding**

1. בעזרת OneHotEncoder בצע קידוד לעמודות הרלוונטיות ()

* יש למחוק עמודה ראשונה מהקידוד
* שימו לב להשתמש ב index = data.index (בהנחה שסט הנתונים שעובדים איתו נקרא data( בעת הפיכת תוצאת הקידוד ל data frame (בהרצת הפקודה pd.DataFrame) אחרת יכולים לבוא אליכם null ים בסיוטים בלילה...

1. הרץ מחדש את הלולאה, שים לב שלקחת בחשבון גם נתונים מקודדים וגם נתונים בקנה מידה מתאים, והשווה את התוצאות לתוצאות הקודמות.

**שלב 6: כוונון פרמטרים**

1. בחר את שיטת עיבוד הנתונים שהניבה את התוצאות הטובות ביותר.
2. נסה לכוונן פרמטרים עבור כל אחד מהמודלים:
   * **Logistic Regression:** כוונן את הפרמטר max\_iter.
   * **Decision Tree:** כוונן את הפרמטר max\_depth.
   * **Random Forest:** כוונן את הפרמטר n\_estimators.

**שלב 7: הצגת confusion matrix כ-Heatmap**

1. עבור המודל Logistic Regression עם הפרמטרים המכוונונים, הרץ את הנתונים והפק את confusion matrix והצג כ-Heatmap באמצעות ספריית seaborn..

**שלב 8: זיהוי הפיצ'רים החשובים בעץ ההחלטה**

1. שלוף את חשיבות הפיצ'רים מהמודל Decision Tree.
2. הצג את הפיצ'רים החשובים ביותר כגרף אופקי (Horizontal Bar Chart).

**שלב 10: הצגת עץ מספר 7 ב-Random Forest**

1. בחר את עץ מספר 7 מתוך ה-Random Forest.
2. הצג את מבנה העץ באמצעות הפונקציה plot\_tree מספריית sklearn.