פרויקט סיווג גידולים: שפיר או ממאיר באמצעות למידת מכונה

מסמך זה מציג פרויקט למידת מכונה מעשי, המדגים כיצד ניתן לאמן מודלים לסיווג גידולים כ"שפירים" או "ממאירים". נסקור את השלבים המרכזיים בפיתוח הפרויקט: מטעינת נתונים וחקירה ראשונית, דרך בניית מודלים שונים ואימונם, ועד להערכת ביצועיהם והסקת מסקנות. פרויקט זה מהווה דוגמה קלאסית ל"למידה מונחית", בה אנו מספקים למכונה דוגמאות מתויגות כדי שתלמד לבצע סיווגים מדויקים על נתונים חדשים.

הקדמה ללמידה מונחית וסיווג רפואי

פרויקט זה מתמקד במשימת סיווג קריטית בתחום הרפואה: קביעה האם גידול הוא שפיר או ממאיר. זוהי דוגמה מצוינת ליישום של למידה מונחית (Supervised Learning), ענף מרכזי בלמידת מכונה. בלמידה מונחית, המודל מקבל סט נתונים המכיל גם את ה"תכונות" (features) של כל דוגמה וגם את ה"תווית" (label) הנכונה שלה. במקרה זה, התכונות יכולות להיות מאפיינים שונים של הגידול (כמו רדיוס, מרקם, היקף ועוד), והתווית היא סוג הגידול (שפיר או ממאיר).

מטרת האלגוריתם היא ללמוד דפוסים וקשרים בין התכונות לתוויות, כך שיוכל לבצע תחזיות מדויקות על נתונים חדשים ולא מוכרים. יכולת זו קריטית באבחון רפואי, שכן זיהוי מוקדם ומדויק יכול להשפיע באופן דרמטי על תוצאות הטיפול. בעוד שמודלים אלו אינם מחליפים את שיקול דעתו של רופא, הם משמשים ככלי תומך החלטה עוצמתי.

1

למידה מונחית

אימון המודל על נתונים מתויגים (תכונות + תווית).

2

סיווג בינארי

משימה של חלוקה לשתי קטגוריות בלבד (שפיר/ממאיר).

3

תכונות

מאפיינים נמדדים של הגידול (למשל, גודל, צורה).

4

תווית

התוצאה הרצויה: שפיר (0) או ממאיר (1).

שלב ו: טעינת ספריות ונתונים

הצעד הראשון בכל פרויקט למידת מכונה הוא הכנת סביבת העבודה וטעינת הנתונים. אנו מתחילים בייבוא הספריות החיוניות שישמשו אותנו לאורך הפרויקט. ספריות אלו מספקות כלים רבי עוצמה לטיפול בנתונים, ויזואליזציה ובניית מודלים.

- **Pandas:** שהם הליבה של עבודה עם נתונים במדעי הנתונים, (DataFrames) לטיפול במבני נתונים טבלאיים.
- **Matplotlib ו-Seaborn:** ספריות לוויזואליזציה של נתונים, חיוניות לניתוח ולבחינת התפלגויות וקשרים.
- **Scikit-learn (sklearn):** ספרייה מקיפה לבניית מודלי למידת מכונה, הכוללת אלגוריתמים רבים, כלי עיבוד מקדים ומדדי הערכה.

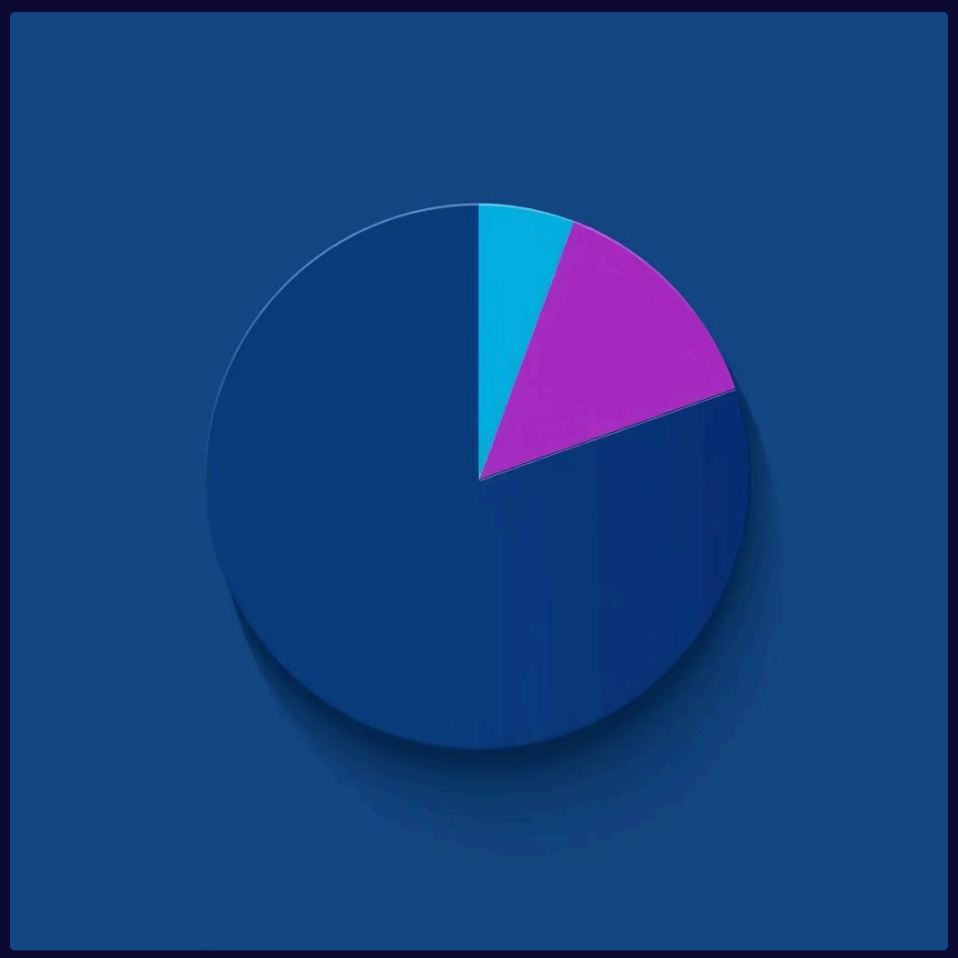
לאחר מכן, אנו טוענים את קובצי הנתונים. נשתמש בשני קבצים עיקריים:

קובץ אימון (Train Data): סט הנתונים שעליו נלמד את המודלים. הוא מכיל דוגמאות של גידולים יחד עם התווית הנכונה שלהם (שפיר/ממאיר).

קובץ בדיקה (Test Data): סט נתונים נפרד ששמור מראש ואינו נחשף למודל בשלב האימון. הוא משמש להערכת ביצועי המודל על נתונים חדשים, שלא ראה מעולם, כדי להבין עד כמה המודל מוכלל היטב ולא סתם "שינן" את נתוני האימון.

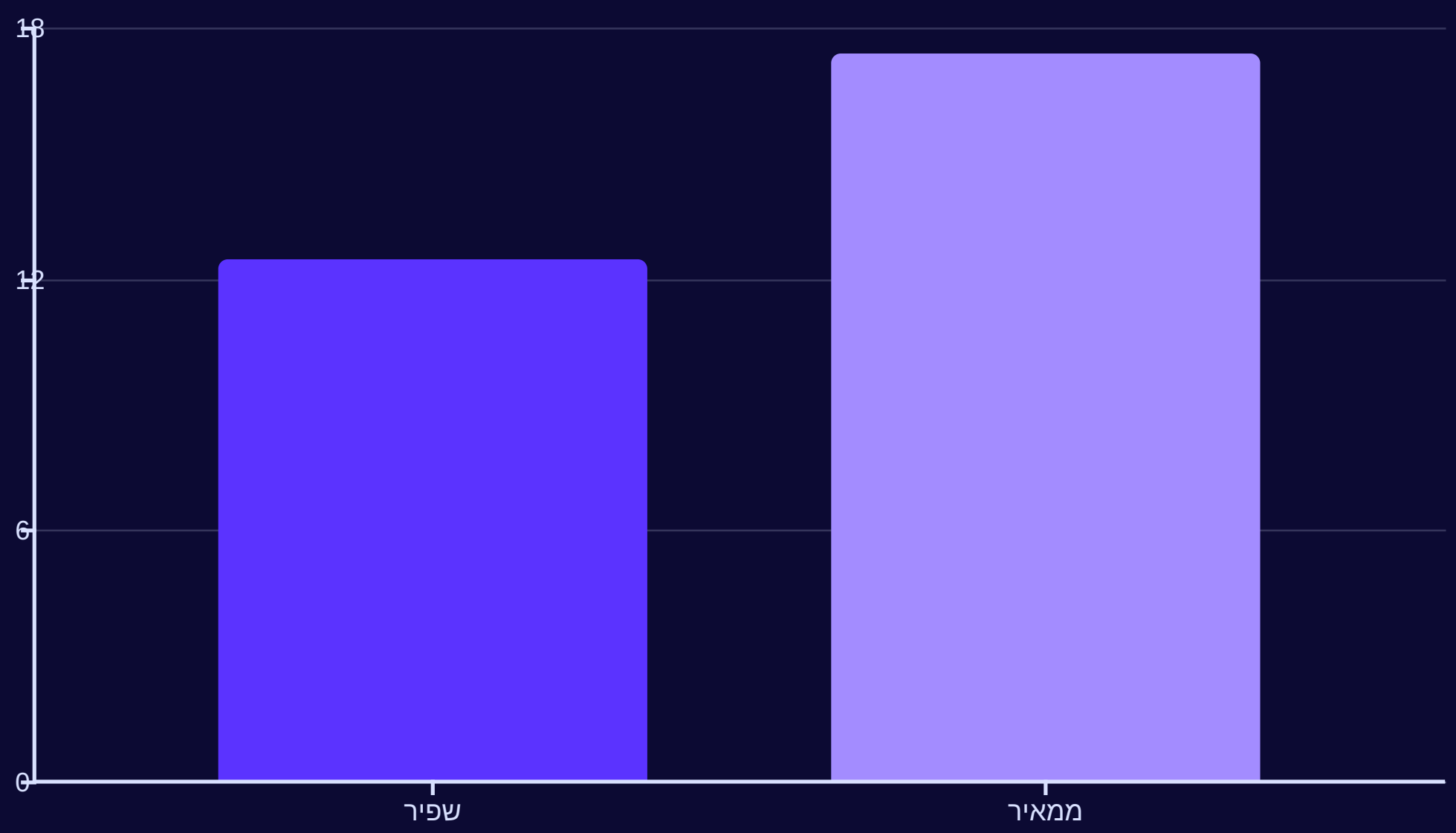
שלב 2: בדיקת נתונים וניתוח ראשוני (EDA)

ניתוח נתונים אקספלורטורי (Exploratory Data Analysis - EDA) הוא שלב קריטי להבנת הנתונים שלנו לפני בניית מודל. בשלב זה, אנו "מציצים" לנתונים, מחפשים דפוסים, אנומליות וקשרים, ומקבלים אינטואיציה לגבי אופיים.



- **הצצה ראשונית:** הצגת 5 השורות הראשונות של ה-DataFrame מספקת תצוגה מהירה של מבנה הנתונים והתכונות הקיימות.
- **התפלגות תווית המטרה:** יצירת גרף התפלגות עבור תווית המטרה (שפיר/ממאיר) היא חיונית. אם הנתונים אינם מאוזנים (Imbalanced), כלומר, יש הבדל משמעותי בכמות הדוגמאות מכל סוג, המודל עלול ללמוד לנחש באופן מוטה. לדוגמה, אם יש 95% גידולים שפירים, המודל יכול פשוט לנחש תמיד "שפיר" ולהשיג דיוק גבוה, אך הוא יהיה חסר תועלת בזיהוי מקרים ממאירים.

ניתוח תכונות בודדות: אנו בוחנים את התפלגות המאפיינים (כמו רדיוס ממוצע) ביחס לתווית המטרה. במקרים רבים, נראה שמאפיינים מסוימים נוטים להיות שונים באופן מובהק בין גידולים שפירים לממאירים. לדוגמה, גידולים ממאירים עשויים להיות בעלי רדיוס ממוצע גדול יותר. הבנה זו מסייעת לנו להעריך אילו תכונות יהיו חשובות למודל בסיווג.



הגרף מעלה מדגים את ההבדל המובהק ברדיוס הממוצע בין גידולים שפירים לממאירים, ומאשש את האינטואיציה כי זוהי תכונה חשובה לסיווג.

שלב 3: בניית מודלי למידת מכונה

לאחר שהכרנו את הנתונים, אנו עוברים לשלב המרכזי של בניית המודלים. עולם למידת המכונה מציע מגוון רחב של אלגוריתמים, וכל אחד מהם ניגש לבעיה בצורה שונה. נבחר כמה מודלים קלאסיים כדי להשוות ביניהם:

K-Nearest Neighbors (KNN)

אלגוריתם המבוסס על 'הרוב קובע' – מסווג נקודה חדשה על פי הרוב מבין k השכנים הקרובים ביותר שלה במרחב התכונות.

Decision Tree

מודל המקבל החלטות על ידי סדרה של שאלות כן/לא, בדומה לעץ זרימה. קל להבנה ופרשנות.

Naive Bayes

מודל הסתברותי המבוסס על משפט בייס, מניח עצמאות בין התכונות. פשוט ויעיל למשימות סיווג רבות.

Pipeline-I StandardScaler:

אלגוריתמים מסוימים, כמו KNN, רגישים לקנה מידה (Scale) שונה של התכונות. לדוגמה, אם תכונה אחת נמדדת במילימטרים (מספרים קטנים) ואחרת במיקרוגרמים (מספרים גדולים), התכונה עם המספרים הגדולים תשלוט במרחק ולכן גם בסיווג. כדי להתמודד עם בעיה זו, אנו משתמשים ב-StandardScaler, אשר מבצע נרמול לנתונים (ממוצע 0 וסטיית תקן 1).

כדי לארגן את שלבי עיבוד הנתונים והמודל באופן יעיל ומוגן מ"זליגת נתונים" (Data Leakage), אנו משתמשים ב-Pipeline של Scikit-learn. ה-Pipeline מאפשר לשרשר מספר שלבים (לדוגמה, קודם Scaler ואז מודל) כך שכל השלבים מיושמים באופן עקבי על נתוני האימון ונתוני הבדיקה, ללא חשיפת מידע מנתוני הבדיקה בשלב האימון.

שלב 4: אימון ובדיקת המודלים באמצעות Cross Validation

לאחר שהגדרנו את המודלים, הגיע שלב ה"למידה" בפועל. אנו מאמנים את כל אחד מהמודלים על קבוצת האימון (Train Data). בשלב זה, המודל מכוונן את פרמטריו הפנימיים כדי למזער את שגיאות הסיווג על הנתונים שראה.

אך כיצד נוכל להיות בטוחים שהמודל לא פשוט "שינן" את נתוני האימון ויתפקד היטב גם על נתונים חדשים? כאן נכנס לתמונה Cross Validation (אימות צולב). זוהי טכניקה חיונית להערכה אמינה של ביצועי המודל:

- **חלוקת נתונים:** קבוצת האימון מחולקת למספר תתי-קבוצות (folds).
- **אימון והערכה איטרטיביים:** בכל איטרציה, המודל מאומן על חלק מה-folds ונבדק על ה-fold הנותר (שלא שימש לאימון באותה איטרציה).
- **תוצאה ממוצעת:** התוצאות מכל האיטרציות ממוצעות, ומספקות אומדן חזק יותר לביצועי המודל מאשר הערכה יחידה על קבוצת בדיקה סטטית. הדבר מפחית את ההשפעה של חלוקת נתונים מקרית.

מדד ה-F1 Score:

בסיווג רפואי, טעויות אינן שוות ערך. זיהוי שגוי של גידול ממאיר כשפיר (False Negative) עלול להיות הרסני, בעוד שזיהוי שגוי של גידול שפיר כממאיר (False Positive) יוביל לבדיקות נוספות אך לרוב פחות מסכן חיים. לכן, אנו משתמשים במדד ה-F1 Score, המהווה ממוצע הרמוני בין דיוק (Precision) לכיסוי (Recall):

- **Precision:** מודד כמה מתוך התחזיות החיוביות של המודל היו באמת חיוביות.
- **Recall:** מודד כמה מתוך המקרים החיוביים האמיתיים המודל זיהה נכון.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

מדד זה חשוב במיוחד כאשר יש חוסר איזון בנתונים או כאשר עלות של סוג טעות מסוים גבוהה יותר.

שלב 5: ניתוח תוצאות והסבר

לאחר אימון המודלים והערכתם באמצעות Cross Validation, אנו משווים את ביצועיהם כדי לקבוע איזה מודל הציג את התוצאות הטובות ביותר, בדגש על ה-F1 Score. המודל המוצלח ביותר הוא זה שמשיג איזון אופטימלי בין דיוק לכיסוי, במיוחד בהתחשב ברגישות של היישום הרפואי.

כדי להבין טוב יותר את ביצועי המודל ואת סוגי הטעויות שהוא מבצע, אנו יוצרים Confusion Matrix (מטריצת בלבול). זוהי טבלה המסכמת את תוצאות הסיווג של המודל ומראה:

	תחזית המודל: חיובי (ממאיר)	**תחזית המודל: שלילי (שפיר)**
אמת: חיובי (ממאיר)	True Positive (TP) - נכון חיובי	False Negative (FN) - שגוי שלילי (החמצה מסוכנת)
אמת: שלילי (שפיר)	False Positive (FP) - שגוי חיובי (אזעקת שווא)	True Negative (TN) - נכון שלילי

מטריצת הבלבול מאפשרת לנו להבין את ה"עלות" של כל טעות. למשל, במקרה של סיווג גידולים, אנו שואפים למזער את מספר ה-False Negatives (גידול ממאיר שאובחן כשפיר), גם אם זה בא על חשבון מספר קטן יותר של False Positives (גידול שפיר שאובחן כממאיר ויוביל לבדיקות נוספות). ניתוח זה מכוון אותנו לשיפורים אפשריים במודל או לבחירה של מודל שמתאים יותר למגבלות היישום.

סיכום ותובנות עיקריות

במסגרת פרויקט זה, עברנו דרך כל השלבים המרכזיים בפיתוח מודל למידת מכונה לסיווג גידולים כשפירים או ממאירים. ראינו כיצד:

- **הכנת נתונים**

  התחלנו בטעינת הספריות והנתונים (אימון ובדיקה).

- **ניתוח אקספלורטורי**

  ביצענו EDA להבנת הנתונים, התפלגותם וזיהוי תכונות רלוונטיות.

- **בניית מודלים**

  בנינו מספר מודלים שונים (Naive Bayes, Decision Tree, KNN) ויישמנו StandardScaler ו-Pipeline.

- **אימון והערכה**

  אימנו את המודלים והערכנו אותם באמצעות Cross Validation ומדד F1 Score.

- **ניתוח ביצועים**

  ניתחנו את התוצאות והבחנו את סוגי הטעויות באמצעות Confusion Matrix.

פרויקט זה מציג לא רק את היכולות הטכניות של למידת מכונה, אלא גם את הפוטנציאל העצום שלה בתחומים כמו רפואה.

הפוטנציאל של למידת מכונה ברפואה

יישומים של למידת מכונה אינם מיועדים להחליף אנשי מקצוע רפואיים, אלא לשמש ככלי עוצמתי לתמיכה בהחלטות קליניות. היכולת לחזות מחלות, לסייע באבחון מוקדם, ולזהות דפוסים בנתונים רפואיים כבירים, משפרת באופן משמעותי את יעילות הטיפול והאבחון.

אבחון מוקדם

זיהוי סימני מחלה בשלבים התחלתיים, לעיתים לפני הופעת תסמינים ברורים.

רפואה מותאמת אישית

התאמת טיפולים למאפייניו הגנטיים והקליניים הייחודיים של כל מטופל.

גילוי תרופות

ייעול תהליכי מחקר ופיתוח תרופות חדשות.

תמיכה בהחלטות

מתן המלצות מבוססות נתונים לרופאים, המשפר את איכות ההחלטות.

בעתיד, אנו צפויים לראות שילוב הולך וגובר של טכנולוגיות למידת מכונה בכל שלבי הטיפול הרפואי, מה שיוביל לשיפור משמעותי באיכות החיים ובתוחלת החיים.

המלצות והמשך פיתוח

פרויקט זה הציג בסיס איתן לסיווג גידולים, אך תמיד קיימים כיווני התפתחות ושיפור. להלן מספר המלצות להמשך:

- **חקר מודלים נוספים:** ניסוי עם אלגוריתמים מתקדמים יותר כמו Support Vector Machines (SVM), Random Forest, או רשתות עצביות עשוי לשפר את הביצועים.
- **הנדסת תכונות (Feature Engineering):** יצירת תכונות חדשות מתוך התכונות הקיימות, שעשויות לחשוף דפוסים חבויים ולשפר את כוח הניבוי של המודל.
- **טיפול בנתונים חסרים או חריגים:** יישום טכניקות מתקדמות יותר להתמודדות עם נתונים חסרים (Missing Values) או חריגים (Outliers) שעשויים להשפיע על דיוק המודל.
- **כוונון היפר-פרמטרים (Hyperparameter Tuning):** אופטימיזציה של הפרמטרים הפנימיים של המודלים (למשל, K ב-KNN, עומק עץ ב-Decision Tree) באמצעות טכניקות כמו Grid Search או Randomized Search.
- **פרשנות מודלים (Model Interpretability):** במיוחד בתחום הרפואה, חשוב להבין מדוע המודל קיבל החלטה מסוימת. שימוש בכלי פרשנות כמו SHAP או LIME יכול לסייע בהבנה זו ולבניית אמון במערכת.

על ידי יישום המלצות אלו, ניתן להפוך את המודל לכלי אבחוני חזק ואמין עוד יותר, שיכול לתרום תרומה משמעותית למאבק במחלות ולהצלת חיים.