הגדרת הבעיה:

1. תיאור עולם התוכן הנחקר:

הבעיה עליה אנו עובדים היא מציאת דרך לחזות מקרה בו אדם חווה התקף לב, על בסיס פיצ'רים שונים הניתנים למדידה.

קיימים מחקרים רבים בנושא וישנם אלגוריתמים שפותחו, אשר מצליחים לחזות מצב שכזה ברמת דיוק גבוהה (כ-90 אחוז).

1. הגדרת שאלת המחקר:

אנו מנסים לאמן מודל אשר יצליח לחזור ברמת דיוק טובה מצב של אדם החווה התקף לב. על ידי שימוש בכלים ושיטות של מערכות לומדות אנו יכולים לוודא את איכות ומהימנות הנתונים, לעבדם ולאן בעזרתם את המודל כך שייתן את החיזוי הטוב ביותר על בסיסם.

הבנת הנתונים:

1. מקור הנתונים:

סט הנתונים נקרא “The Cleveland heart disease data”, אשר סופק על ידי אוניברסיטת קליפורניה ומקורו מאתר UCI Machine Learning Repository.  
בסיס הנתונים מכיל 303 תצפיות אשר 212 מתוכם סווגו לאימון המודל והשאר לבחינתו. כמו כן, הוא מכיל במקור 76 פיצ'רים אך אנו משתמשים ב13 מהם בלבד, כאשר חלקם מספריים וחלקם קטגוריים:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| שם הפיצ'ר | הסבר | סוג דאטא | טווח ערכים | משמעות | מדידה |
| age | גיל המטופל בשנים | מספרי | 34-80 1 |  | ידני |
| gender | מין המטופל | קטגורי | 0-1 | 0-נקבה  1-זכר | ידני |
| cp | סוג כאב בחזה | קטגורי | 0-3 | 0-תעוקת לב טיפוסית  1-תעוקת לב לא טיפוסית  2-לא נובע מתעוקת לב  3-אסימפטומתי | מומחה |
| trestbps | לחץ דם במנוחה | מספרי | 94 - 200 |  | מדידה |
| chol | רמת כולסטרול | מספרי | 126-409 |  | מדידה |
| fbs | רמת סוכר בדם | קטגורי | 0-1 | 0-עד 120 מ"ג/דצ"ל 2  1-אחרת | מדידה |
| restecg | תוצאות אק"ג | קטגורי | 0-2 | 0-תקין  1-עיוות בגל T במקטע ST  2-סיכוי להיפרטרופיה בחדר שמאלי | מומחה |
| thalach | דופק לב מקסימלי | מספרי | 88-195 |  | מדידה |
| exang | תעוקת חזה עקב מאמץ | קטגורי | 0-1 | 0-לא  1-כן | מומחה |
| oldpeak | צניחה בסגמנט ST באק"ג, במאמץ לעומת מנוחה | מספרי | 0-5.6 |  | מומחה |
| slope | שיפוע בסגמנט ST בשיא מאמץ | קטגורי | 0-2 | 0-חיובי  1-שטוח  2-שלילי | מומחה |
| ca | מספר כלי דם גדולים | אורדינלי / קטגורי | 0-4 | 0-3-כמות  4-לא ידוע | מומחה |
| thal | מחלת תלסמיה | קטגורי | 0-4 | 0-ערך חסר  1-פגם קבוע  2-נורמלי  3-פגם ניתן לריפוי | מומחה |
| y | אירוע התקף לב | קטגורי | 0-1 | 1-התקף לב  0-אחרת | מומחה |

1 לאחר מחיקת תצפיות לא הגיוניות מבחינת מספרים

2 נחשב נמוך או תקין.

תעוקת לב – מחלה של שריר הלב הגורמת לכאבים ולחץ בחזה משום שלא מסופק מספיק דם עם חמצן.

לחץ דם נורמאלי נחשב – 120 (פחות מ-90 תת לחץ דם, מעל 140 יתר לחץ דם, מעל180 – משבר) **\*\*אפשר לחלץ לקטגוריות אם נרצה**

כולסטרול – עד 200 נורמלאי, מעל 240 זה גבוה ( במקרים נדירים יכול להיות יותר – הערכים הגיוניים)

סוכר בדם – הממוצע התקין הוא סביב 90-130

**\*\*\* יש צורך לרשום ממש במספרים הסתברויות לכל לייבל בכל פיצ'ר?**

**או שההיסטוגרמה מספיקה?**

עיבודים שונים מקאגל:

<https://www.kaggle.com/ronitf/heart-disease-uci>

עיבוד נוסף לדאטא-סט הזה:

<https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>