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 ב14 מהם בלבד, כאשר חלקם מספריים וחלקם קטגוריים:

Id – מזהה התצפית עבור כל מטופל

Age גיל המטופל בשנים (משתנה מספרי בדיד) –

Gender – מין המטופל (משתנה בינארי [0-נקבה, 1-זכר]

Cp –

….

למלא את זה על כל המשתנים (?) יש עוד משהו להוסיף חוץ מהסבר, אילו ערכים מקבל וסוג הערך?

עיבודים שונים מקאגל:

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

עיבוד נוסף לדאטא-סט הזה:

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