**Analyzing heart disease through analytics**

**DR. MIGHTY ITAUMA ITAUMA.**

**RICHARD DEVOS Graduate school, NORTHWOOD UNIVERSITY**

**Contributors:**

**Saumya Padhi,**

**Imran Shaik,**

**Srinivas Bhootam**

**Pooja Shinde**

**Repository Link:**

**https://github.com/saumyasam/casestudy\_heart**

**Problem Definition:**

* How can we predict whether a person has heart disease based on their health and demographic information?
* In the dataset provided, the target variable defines whether a person has heart disease Specifically,
* 0=Indicates no heart disease [Negative Case]  
  1=Indicates the presence of heart disease [Positive Case]
* This binary variable Target is the dependent variable in your analysis, and the goal of a logistic regression model is to predict this target based on the other independent variables, such as age, sex, cholesterol, and blood pressure.

**Variables:**

**Dependent Variable:**

* **target:** This variable indicates the presence or absence of heart disease. Our aim here is to predict heart disease, "target" is the dependent variable.

**Independent Variables:**

* **age**
* **sex**
* **cp:** chest pain type (categorical)
* **trestbps:** resting blood pressure
* **chol:** serum cholesterol in mg/dl
* **fbs:** fasting blood sugar > 120 mg/dl (yes/no)
* **restecg:** resting electrocardiographic results (categorical)
* **thalach:** maximum heart rate achieved
* **exang:** exercise induced angina (yes/no)
* **oldpeak:** ST depression induced by exercise relative to rest
* **slope:** the slope of the peak exercise ST segment (categorical)
* **ca:** number of major vessels (0-3) colored by flourosopy
* **thal:** Thallium stress test results (categorical)

A screenshot of a computer

Description automatically generated

A graph of two people

Description automatically generated

**Exploratory Data Analysis:**

* Effect of Chest Pain Type (cp): The data suggest that different chest pain types may have varying associations with heart disease. Individuals with "cp" = 2 appear to have the highest risk, while those with "cp" = 0 have the lowest risk.
* Effect of Thal (Thallium Stress Test): The "Thal" variable also seems to be associated with heart disease risk. Individuals with "Thal" = 2 have the highest proportion of heart disease, followed by those with "Thal" = 3.
* Prevalence of Heart Disease: The "target" variable indicates the presence or absence of heart disease.
  + The proportion of individuals with heart disease (target = 1) varies across the different "cp" categories.
  + For example, individuals with "cp" = 2 have the highest proportion of heart disease (77.11%).

**Binomial Logistic Regression:**

**A screenshot of a document

Description automatically generated**

**Description:**

* **Thal**: OR = 0.39 (95% CI: 0.30, 0.51). A higher thal value reduces the odds by 61% i.e. A decrease in thal is associated with a significant decrease in the odds of the outcome. i.e.
* **sex**: OR = 0.18 (95% CI: 0.12, 0.28), Being female (compared to male) is associated with a significantly lower odds of the outcome.
* **cp**: OR = 2.20 (95% CI: 1.84, 2.62). Chest pain type is significantly associated with the outcome. Individuals with a certain type of chest pain have over twice the odds of the outcome compared to those with the baseline chest pain type.
* **restecg**: OR = 1.53 (95% CI: 1.11, 2.09) Resting electrocardiographic results are associated with an increased odd of the outcome.
* **thalach**: OR = 1.02 (95% CI: 1.01, 1.03). Maximum heart rate achieved has a weak positive association with the outcome.
* **slope**: OR = 2.49 (95% CI: 1.86, 3.34) The slope of the peak exercise ST segment is associated with an increased odd of the outcome.

**Summary:**

* Significant predictors (p < 0.05) include: **age, thal, sex, cp, restecg, thalach, exang**, and **slope**.
* Predictors such as **age** and **exang** decrease the odds of the outcome, while **cp** and **slope** increase it.
* No multicollinearity issues were detected (VIF < 5).