CA05 README

Conduct logistic regression analysis on cardiovascular disease data

Cardiovascular disease is labelled 0 when there is no risk and 1 when there is risk present.

In the first part, I will be cleaning the data before I split it into training and testing.

After that, I will run logistic regression on the dataset I split and check the accuracy of these models based on different c values. Making a model with logistic regression is better because we are dealing with a classification problem and the output it generates is a discrete variable (0/1) instead of continuous (linear regression). When c value is the highest, “1”, it generated the highest accuracy.

I decided to keep all of the variables because I am more concerned about the accuracy and other performance metrics of my model compared to the significance each variable have on cardiovascular disease. After performing the test, I found out that waist has the highest positive impact while hip size has the highest negative impact in terms of risks.

After running these tests, the AUC graph along with other performance metrics indicate that this linear regression model is fairly accurate, scored around the 80% range.