**Max Kaiser – CA05**

**Evaluate the performance of your model:**

I started my analysis of the dataset by checking data quality and doing basic exploratory analysis. The feature columns were already numerical values and didn’t have to be transformed, which made it easier to build my binary classifier model to predict CVD Risk (Logistic Regression).

First of all, I eliminated features that had p-values over 0.05 (insignificant). This reduced my dataset to 11 instead of 16 feature columns. After defining X and y (response variable) I fitted the model and got an AUC score of 0.73 and accuracy of 0.6. The accuracy is pretty low, but AUC is a better measurement for logit regression performance. The AUC tells us how well the LR Model classified positive and negative outcomes.

Ranging from 0.5 to 1, the ROC curve (see plot in Part 3 - Notebook), our model with AUC of 0.73 is in the acceptable range to diagnose patients’ risk towards CVD, considering that we are not diagnosing a disease directly and further health checks are required (False positives not as dramatic- prevention is the focus).This model can help the doctors to narrow down which patients to focus on and optimize resources/personal.

I displayed the feature importance of all the features sorted in the order of decreasing influence on the CVD risk. **The results** showed that waist and hip measurement in cm had the largest influence (inverse correlation/relationship) on our outcome variable. 🡪 highest absolute coefficients. Average weight in kilograms was also in the top 5 highest coefficients which implied a bigger explanatory power on CVD risk. Finally, the “hlthlm25“feature column which asked about how the person’s health has limited social activities with friends and relatives also had a huge impact.

This data analysis can help to improve questionnaires that are handed out to patients (focus on BMI, weight, health impacting social life) and eliminate questions that are not as important for classifying a patient’s risk.