**At the beginning**

We encountered issues with our dataset sourced from Kaggle, namely related to counterintuitive findings that raised concerns about the accuracy of certain attributes. For instance, the dataset suggests that experiencing chest pain during exercise correlates with a lower likelihood of heart disease, while showing no chest pain increases the risk of having heart disease, an interpretation that contradicts science (Hamada, 2025). This anomaly suggests that the target variable in our Kaggle dataset may have been reversed by the uploader - where a value of 0 actually represents a diseased heart and 1 indicates a healthy heart, rather than the original convention where 0 = healthy and 1 = diseased. The dataset also seems to contain multiple duplicate rows of information, potentially raising questions regarding the reliability of the source (Kalsi, 2024). This can be problematic as there are 723 rows of duplicated entries, which could introduce bias and subsequently negatively impact model performance (Rehman, 2023). Moreover, there are incorrect entries, perhaps reflective of ‘fat finger’ errors when transcribing data from the original data source (Mahan, 2022). For example, the predictor thal should only have entries 0-2, as can be seen in the screenshot of attribute information as provided by kaggle, but there are 410 entries containing 3. Here, what 3 stands for can be questioned.

A screenshot of a computer

AI-generated content may be incorrect.

Thus, while the coefficients discussed in our report remain valid in terms of magnitude and relative importance, their directional interpretation should be considered in reverse.

**In the body of our essay**

Please note the aforementioned issues with our dataset, namely in regards to the coding of our outcome variable, target – the directional interpretation should be considered in reverse.