# Technical Report

Models

| **Model Name** | **Model Type** | **Target Values** | **Features** | **Input Shape** | **Output Value** |
| --- | --- | --- | --- | --- | --- |
| model6\_LogisticRegression | Logistic Regression | Healthy (0), Esophageal conditions (1) | Age at Collection, BMI (kg/m2), Gender\_F, Gender\_M | (1, 4) | Healthy (0) or At risk for esophageal conditions (1) |

Scaling the data   
  
Using the same scaler that was used in training is important because it ensures that the data pre-processing applied to the training data is also applied to any new data in a consistent manner.

When you fit a scaler on the training data, it computes the mean and standard deviation of each feature in the dataset. These values are used to scale the features so that they have zero mean and unit variance. This process helps to improve the performance of many machine learning algorithms, as they are sensitive to the scale of the input features.

When you apply the scaler to new data, it uses the same mean and standard deviation values that were computed during fitting. This ensures that the new data is scaled in the same way as the training data. If you were to use a different scaler or fit a new scaler on the new data, it would likely produce different scaling parameters, which could lead to inconsistent or incorrect predictions.

In summary, using the same scaler for both training and new data ensures that the preprocessing is consistent, which is essential for obtaining accurate and reliable predictions from your machine learning model.

## Future Work

**Personalized Risk Factors:** Allow users to input additional risk factors such as smoking history, alcohol consumption, family history of esophageal cancer, and diet. These factors can be used to further refine the risk prediction.

**Confidence Level of Predictions:** Display the confidence level (probability) of the model's predictions, which can help users understand the uncertainty associated with the prediction.

In the Streamlit app, use the predict\_proba() function of the chosen model to calculate the probability of the prediction. Display the probability along with the risk of cancer.