Addendum:

Here are some assumptions in this assessment that I did not address much at the time I was doing the coding, but I want to mention them here.

1. Dataset size: My assumption is that we have a large dataset. Roughly there are around 3500-4000 rows of data. The statement also aligns with this assumption (It says 'fair volume of data').

2. Feature types: I also overlooked what type of features we collected other than the mentioned ones. Some of the features are mentioned in the statement. Here is what the feature data could look like:

Attributes: [Type of procedure (Object), How long it lasted (int), Severity (Object), Gender (Object), Age(int), Smoking (Boolean), Number of days since last follow up (int), Blood Pressure (float)].

3. Relationship between features: My assumption would be some of the features has complex relationship with target variables and between themselves (non-linear).

4. Latency of new prediction: My assumption is, we don't need a very fast model, in other words, speed is not a crucial factor in our model.

The mentioned assumptions are necessary for Random Forest model that I used. Random Forests need large datasets, the feature relationship need to be complex, and they require more time

to train than linear models.

If we had small datasets, fairly simple relationships between variables and if speed of training and making new prediction were crucial, I would probably choose a linear model like

a logistic regression or a Support Vector Machine.

One issue that I need guidance about would be how to use the granular features about the outcome. My assumption in this problem was not to include them while training the model because.

we won't have access to them while making final predictions.