The Data File seems structured but missing a lot of values in between and containing NaN values. I made some changes to the file *'echocardiogram.data'* as the column names were not there (I am attaching these files back within the solution package).

The *'echocardiogram.names'* file explicitly said that for a person to be alive we should check the second column and in response should also take a look at the column 'alive-at-1'. Taking a closer look at those, I was able to infer that the given data in those columns is irregular and uncertain while not making it clear that if they are alive or dead at the end of the said period.

Looking at the problem statement I can say that it can be framed as a classification problem. To make it look more coherent for training and prediction part, I add a column <u>'2-years-surv'</u> which checks all rows in <u>'survival'</u> column if the person survived for more than 24 months or not and assign a Boolean truth value (True or False). This column becomes the dependent variable, and all the others are independent variable in the model definition. I chose Logistic regression for my basic classification task as its one of the most powerful classifiers present. It has been proven to give better unbiased results with lower variances. In addition, it does not make any assumptions about distribution of classes and is less inclined towards overfitting.

I applied the classifier and method on the data (The code is attached in *'1-Logistic Regression.ipynb'*). The accuracy score came out to be close to 97% and the confusion matrix showed only one prediction outside the actual one. This is a result of small dataset, and prediction is easier in this kind of situation.

As there are lots of models and approach to classification. I chose Random forest approach as the next method to test on this dataset. They are made through multiple decision trees which provides an ensemble learning method with average prediction of the decision tree.

I have applied this model on the dataset with same data cleaning techniques from the Logistic regression problem (The code is attached in *'2-RF Classifier.ipynb'*). The only difference is the amount of decision trees or estimators supposedly given. After carefully going through the results you can see that the accuracy is 100%, which is expected as the dataset is so small and training set was even smaller.

The Random Forest was able to create better predictions than Logistic Regression just because this classification problem fitted better with tree approach rather than variance approach. Also having a large dataset might change the scores on Logistic regression's favor.

In case of unlabelled records, the accuracy will depend majorly on the size of the dataset but given if the variables behave linearly. The accuracy will go close to 90% for sure for both models.

Moving forward to modeling and finding the missing values in the *'echocardiogram.test'*. As the problem statement best fitted a classification problem, the prediction models I chose above can only be applied to categorical classes or Boolean labels. While the missing values in the test dataset has continuous values, I applied a Linear regression model to the problem of imputing missing values in the dataset (the code is attached in the file *'3-LRegressor.ipynb'*). The basic difference is the training and testing sets. I made an iterator for the 4 missing columns and trained the model with columns from the *'echocardiogram.data'* file. I took the dependent variables from the data file and then deleted the missing columns one by one as independent variable and iterating through it for training. In terms of predicting those values, the test set is the *'echocardiogram.test'* file with the columns already present and cleaned for the purpose of testing so that they do not have to be reshaped. Ultimately, you get the results for all the 19 missing rows for the 4 columns with their mean squared error.

The dataset can be made more uniformed and sounder through clear demarcation of 'still-alive' column. Also, adding an age they lived until column can add more biasness towards prediction of survival rate. The addition of a feature of their diagnosis through different methods can add some interesting pointers and views.