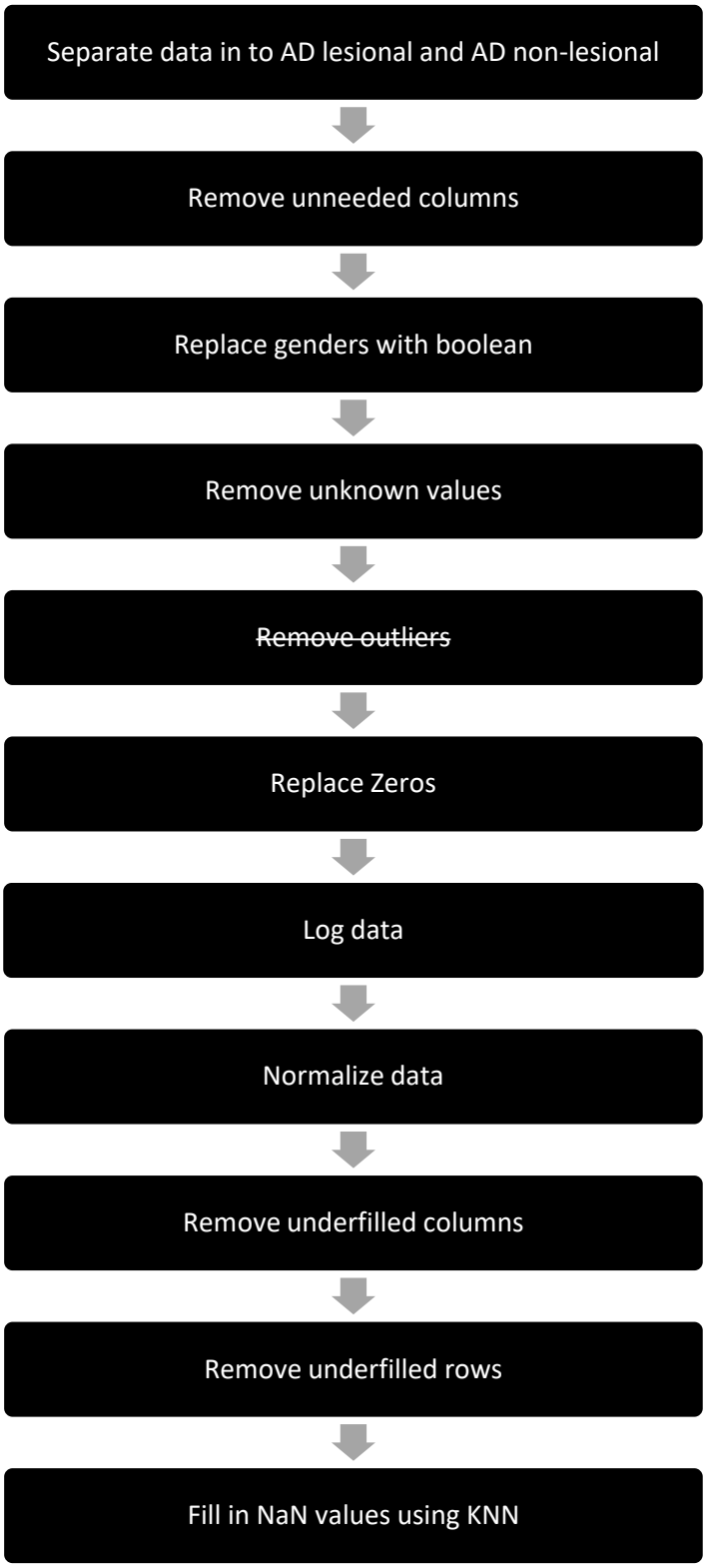
Pre-processing the BIOSCAD data

Flowchart:



Separate data in to AD lesional and AD non-lesional:

The dataset includes data from those who do not suffer from atopic dermatitis (AD), lesional skin from AD sufferers, and non-lesional skin from AD sufferers. As the aim is to correlate SCORAD with the compound levels in the skin, the data from those who do not suffer from AD was discarded. The data for AD sufferers was split in to lesional and non-lesional. This was done as there were significant gaps in the data from lesional skin for a number of patients. As such, when we remove patients with too much missing data we would lose a significant amount of data from the non-lesional skin. The two skin types have been processed separately and can be joined later if necessary.

Remove unneeded columns:

Certain columns of data contained duplicate information and were removed. For example, the date of birth and date of visit columns are not needed as we have the age at visit which is more useful. There are also columns stating whether the data is for lesional or non-lesional skin. This is no longer useful as the data has already been split in to lesional and non-lesional as outlined above.

Replace genders with a Boolean:

The genders were replaced with a Boolean in order to make processing the data easier. Female was represented as 0, male was represented as 1.

Remove unknown values:

Unknown values were shown in the data using words (strings). These were replaced with NaN (not a number) so that the data could be stored as a float.

~~Remove outliers:~~

~~In order to find the outliers in the data, the mean and standard deviation were found for a given attribute. These calculations were done excluding the NaN values. Any item more than 1.96 standard deviations (95% certainty) from the mean was deemed to be an outlier and replaced with NaN.~~

Replace zeros in continuous data:

Any zeros in the continuous data (too small to measure) were replaced with a value 10 times less than the minimum non-zero value. This was done because $\log(0) = \infty$ which will be replaced by NaN. Thus, data would be lost.

Log data:
All continuous data was replaced with its log value.

Normalize data:
All continuous data was linearly scaled so that the minimum value was at 0 and the maximum value was at 1. This was done so that no single attribute would have a larger effect than the other when using K-nearest neighbours.

Remove underfilled columns:
Columns (attributes) with less than 85% of the values filled (not NaN) are removed from the dataset. This is done first because there are a large number of attributes but, relatively, not many data points.

Remove underfilled rows:
Rows (data points) with less than 85% of the values filled (not NaN) are also removed from the dataset. As the underfilled columns have already been removed it reduces the chance that rows will be removed.

Fill in NaN values using KNN:
The values that are NaN, either because they were originally unknown or they were outliers, are filled in by using the corresponding value from the data point's nearest neighbour. When finding the nearest neighbours, only attributes that are not NaN in the data point are used. If a value for such an attribute is not present in a neighbour then a penalty distance of 1 (the maximum it could possibly be) is applied. Once the neighbours have been ranked by distance, the closest neighbour with a non-NaN value in the target attribute is used.

General points:
Wherever possible, processes which might be done in other data analysis tasks are abstracted away in to their own functions. Many of these functions use parallel computing to speed up processing time. It is worth noting that, when used on a small data set such as this one, parallel computing may actually increase the processing time due to the initial overhead of splitting up the task. Nonetheless, this increase is slight and the advantage gained when processing large data sets is great so they have been written in parallel.