Some information about the data

Name	Email	Country	College/Company	Specialisation
Kelvin Mpofu	mpofukelvintafadzwa@gmail.com	South Africa	n/a	Data science
Purity Nyagweth	purityeverniter@gmail.com	Kenya	n/a	Data Science
Reshma Jayapalan	reshma.jayapalan@gmail.com	UAE	n/a	Data Science
Hanouf Hazza	hanouf.haz@gmail.com	Saudi Arabia	n/a	Data Science

The data-set we have has 2 discrete value features/columns, 1 continuous value feature and 66 categorical value columns/features including the target which is the consistency flag.

The data set has many missing values specifically in the features 'Race', 'Ethnicity','Region','Ntm_Speciality','Risk_Segment_During_Rx', 'Tscore_Bucket_During_Rx','Change_T_Score','Change_Risk_Segment'.

There are outliers in the feature sets Dexa_Freq_During_Rx. The count of risks column is positively skewed. We neglected to deal with outliers because we plan on using random forest algorithm and decision trees which are robust against outliers.

To deal with the missing values we could drop all the missing value columns but this would result in a lot of data getting lost. One strategy would be to using the missing value as a label on its own, this may not be the most efficient approach. Another approach might be to fill the missing data with the most popular label. Another approach might be to use a machine learning algorithm to fill the missing data.

The approach we used was, we looked at features whose missing data was less than 5 % of the feature set and we dropped those. We then used one hot encoding to encode the remaining data and encoded the missing values as a feature. This may not be the best approach.