The metadata file is attached in this pull. The data consists of 280 columns and 452 observations.

Column 14 which is a parameter called ‘J’ has 83% missing values and hence it is removed from the dataset.

10% of the dataset is set aside as test set. Parameter hypertuning is done using GridSearch/cross validation on the training dataset.

I trained the model using KNN, Logistic, SVM(Linear and Kernel models) decision trees and random forest. On top of this, I used gradient boosting, adaboost and bagging techniques to improve accuracy and treat bias/variance trade off.

Since there are 280 indicators and only 452 observations, I applied PCA to reduce dimensionality and hence reduce model complexity.

**Conclusion**

* Boosting and Bagging seem to rectify the problem of overfit seen in almost all models except one where AdaBoost did not prove much effective on decision tree. The train and test accuracies have become more comparable.
* Principal Component Analysis (PCA) has been used for dimensionality reduction. All the models that were previously run on the actual data are now run on the PCA components. There are no notable improvements in terms of the accuaries for almost all models and the problem of overfitting seems to still persist except in the case of decision tree.
* We have used the scoring parameter as **recall** as it is a good measure of accuracy, especially in medical applications where the **null accuracy** is high