**CAPSTONE**

**MAPPING DEPRIVED AREAS ON LOW AND MIDDLE INCOME COUNTRIES**

Osemekhian Ehilen

**CONTEXTUAL FEATURES**

**HANDLING MISSING/UNNECESSARY VALUES**

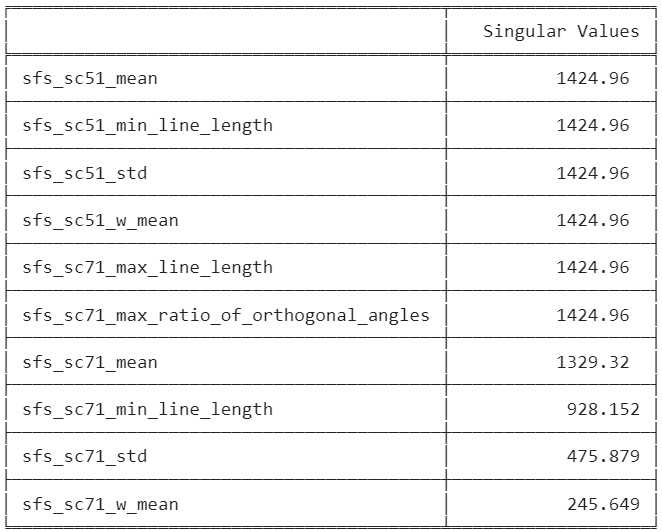
Based on the GIS collaborator, the value -9999 are not necessary in our data frame. Hence, this values are set to zero.

The feature pantex\_sc3\_min was removed because it had over 80% of irregular data (-9999).

**SINGULAR VALUE DECOMPOSITION (SVD)**

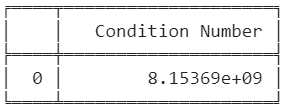
The SVD approach aims at checking for collinearities among features. This will be done on the contextual features on Nairobi’s data.

Below are the last 10 singular values for ten features:



The features with singular values closer to zero have more presence of collinearity.

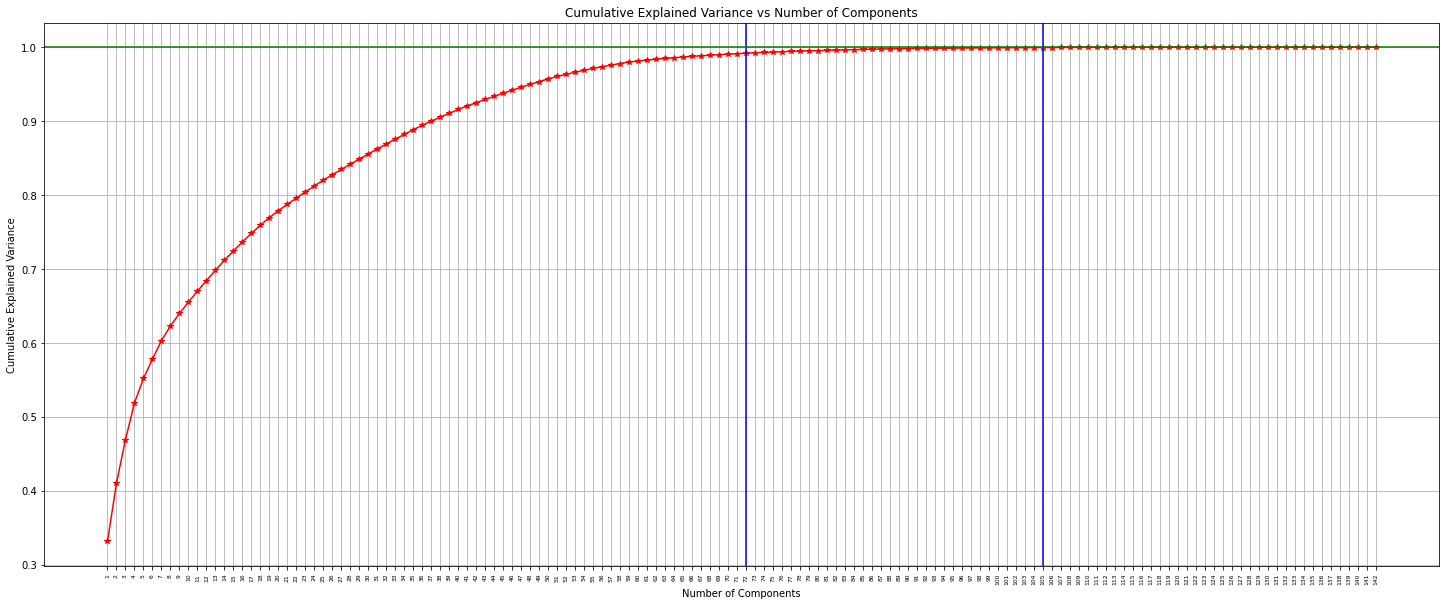
Using the condition number which gives us the degree of collinearity present amongst the features gives:



This indicate a severe degree of collinearity among the features because the condition number is greater than 1000.

**PCA**

The principal component analysis will be used to track most important features that can the original data with less features.



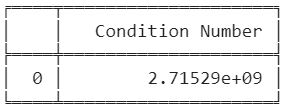
The graph above displays the degree of explained variance by number of features.

The 143 principal components generated gives each correlation value of which I took the highest absolute value. Then I got 105 unique important features from the PCA.

With half (72) the feature size, you get 99.2% variance explained and with 105 feature size, you get 99.97% variance explained and with 108 (68 unique features from PCA and 40 features from collaborator’s Random Forest feature importance) feature size, you get 99.98% variance explained.

Also, with 120 features you get 100% variance explained.

The unique features from the first 105 features are 68 features which I performed SVD on. This resulted in condition number coming down significantly:



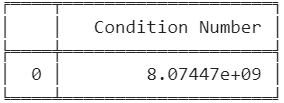
**Variance Inflation Factor (VIF)**

The VIF also helps in checking multi-collinearity. And with VIF value 1 means features are not correlated, between 1 and 5 means features moderately correlated then greater than 5 means feature are highly correlated.

The formula is given below:

Where R is the un-adjusted R-Squared.

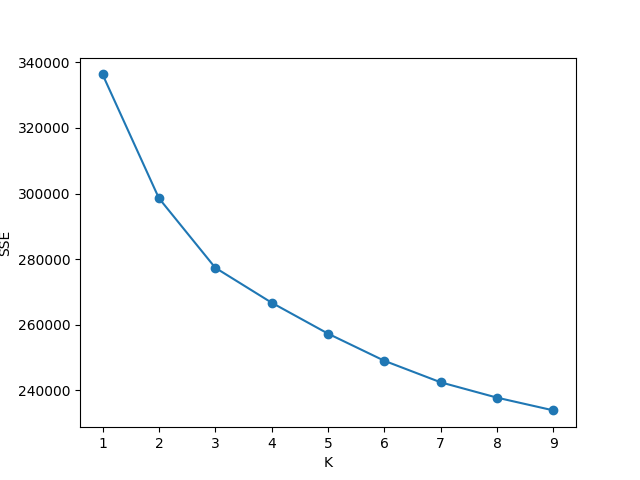
The condition number only reduced slightly and thus PCA performed better than VIF in this scenario.

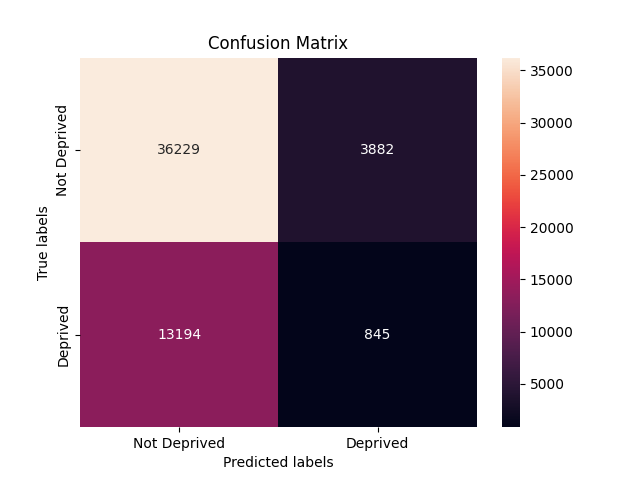


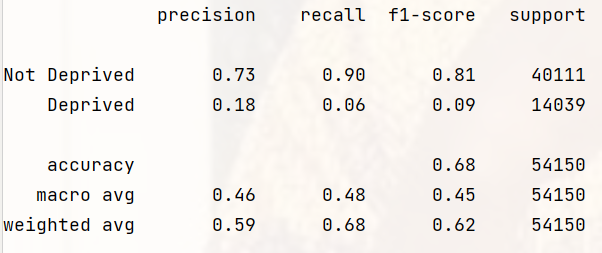
**K-MEANS**

K-Means clustering is an unsupervised learning which gives labels based on features given.

The number of clusters to be used by the algorithm was set to 2 based on the elbow method which at 2 clusters gave a significant drop in sum of square error (SSE):



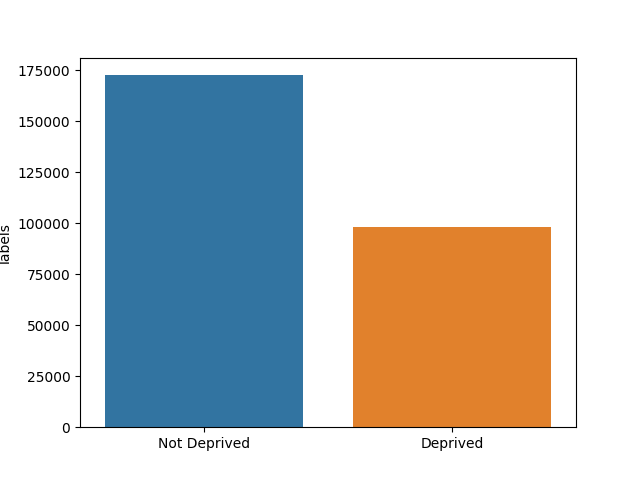
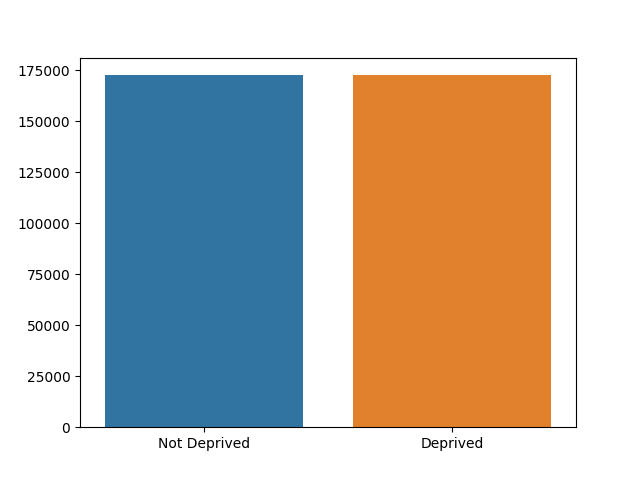




The K-Means with two clusters gave an f1-macro of 45% on test set which confirms the claim of previous collaborator indicating k-means not good for this project.

**LOGISTIC REGRESSION & RANDOM FOREST CLASSIFIER**

**Handling Class Imbalance**

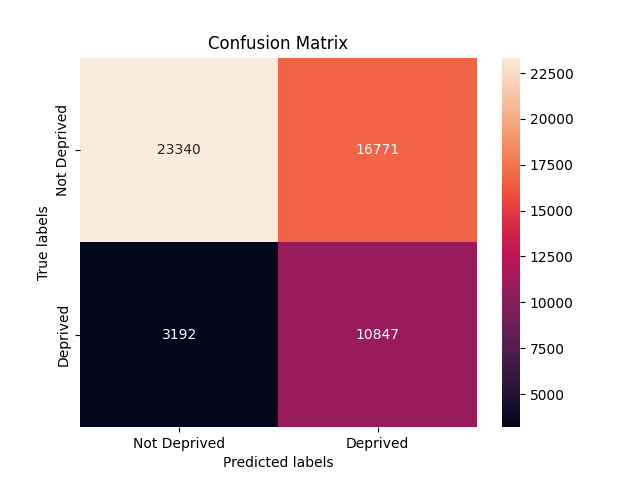
**** ****

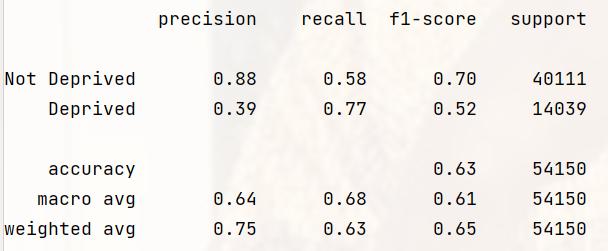
**For 68 Features**

The best estimator was logistics regression with parameters:

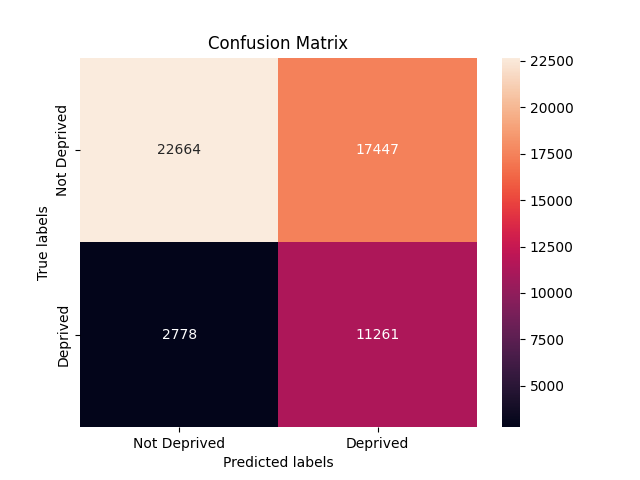
'model\_\_objective': 'binary:logistic', 'model\_\_n\_estimators': 30, 'model\_\_max\_depth': 20, 'model\_\_eval\_metric': 'logloss'.

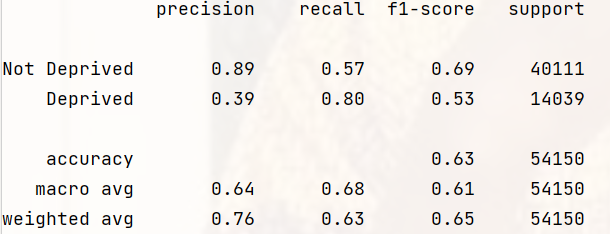
For null values handled with zeros:





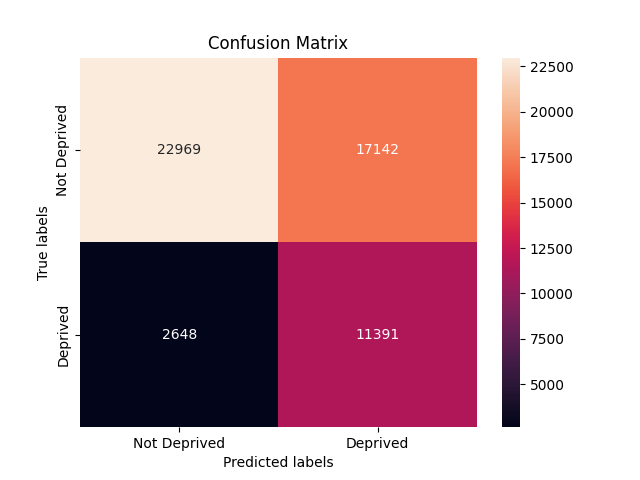
For null values handled with column’s mean value:

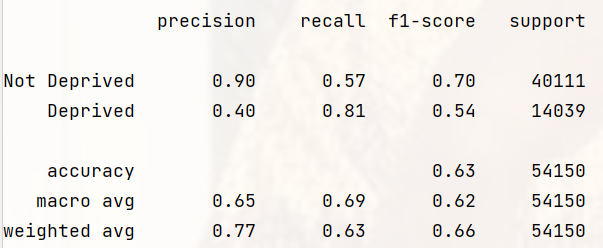




**For 81 Features**

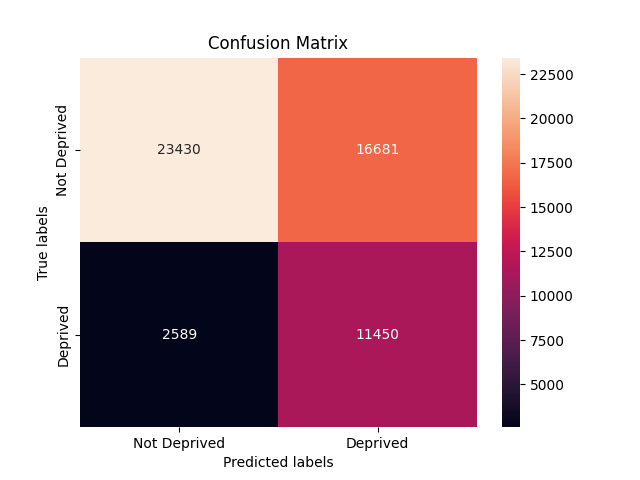
For null values handled with column’s mean value:

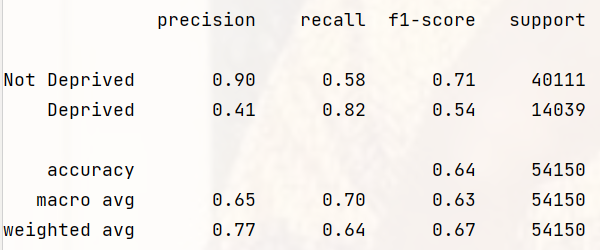
****

****

**For 97 Features**

For null values handled with column’s mean value:

****

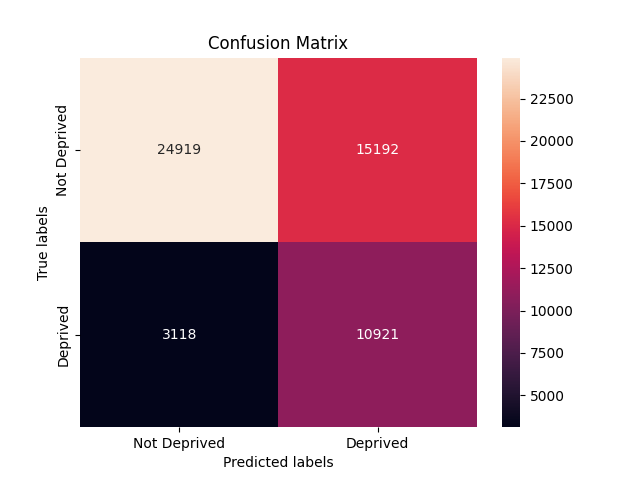
****

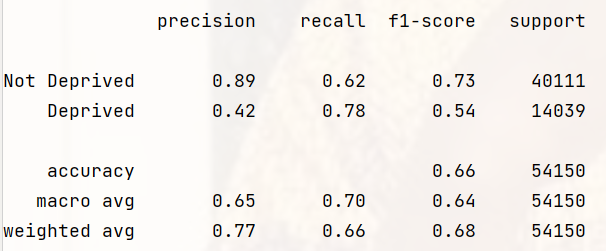
**For 108 Features**

The best estimator was logistics regression with parameters:

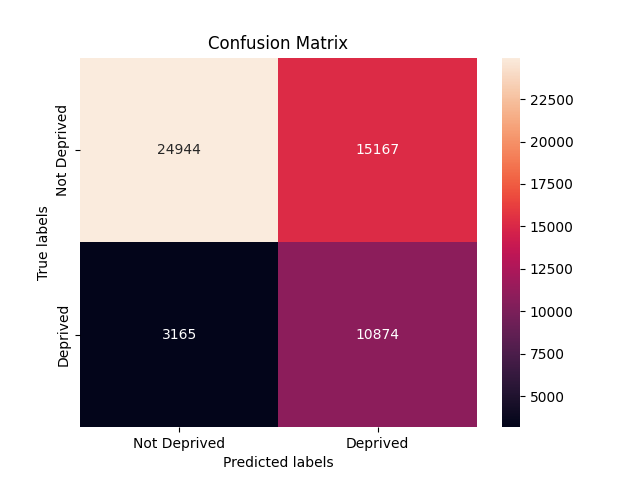
'model\_\_objective': 'binary:logistic', 'model\_\_n\_estimators': 30, 'model\_\_max\_depth': 20, 'model\_\_eval\_metric': 'logloss'.

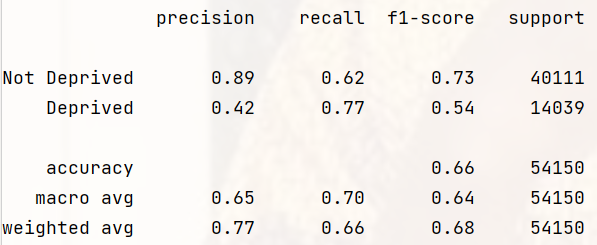
For null values handled with zeros:





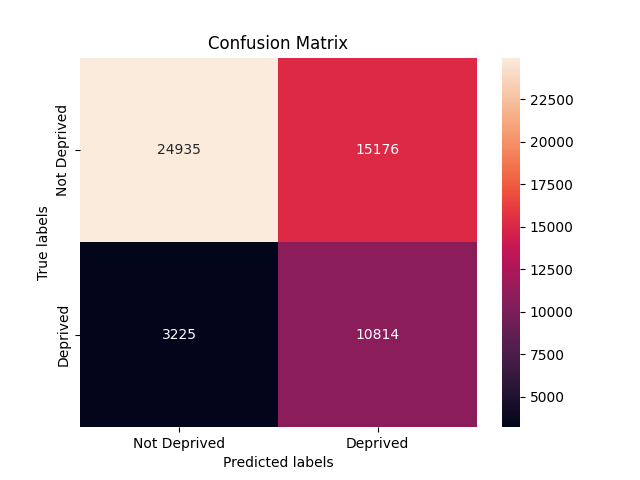
For null values handled with column’s mean value:

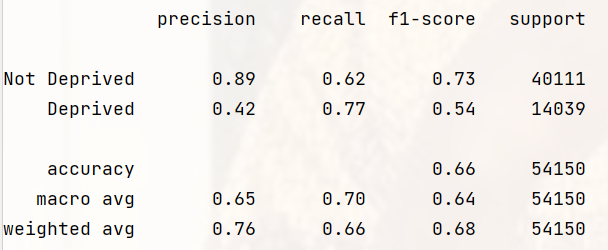




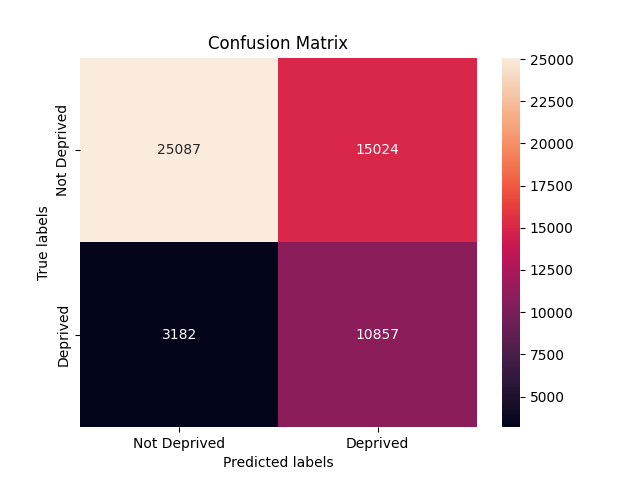
**For 143 Features**

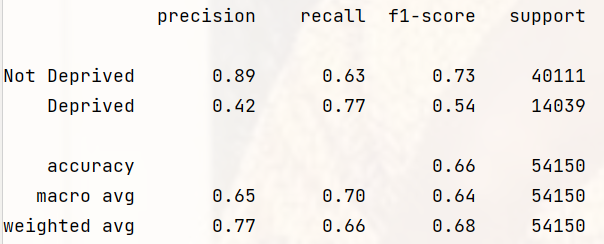
For null values handled with zeros:





For null values handled with column’s mean value:





**COVARIATE FEATURES**

The covariate feature which includes 53 features are going to be examined below.

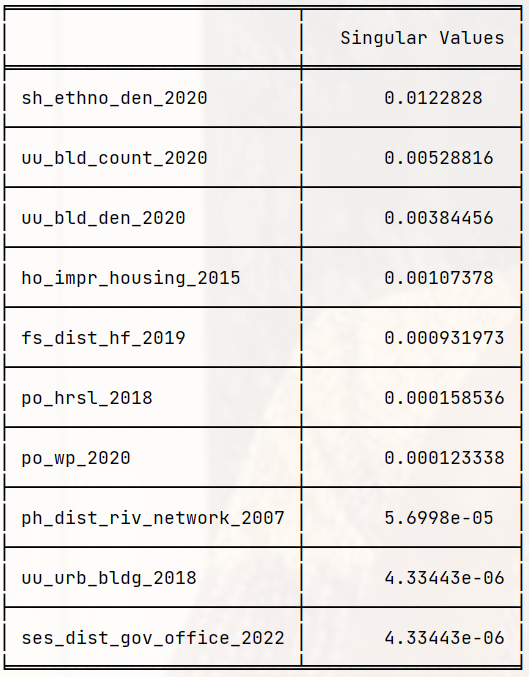
**HANDLING MISSING/UNNECESSARY VALUES**

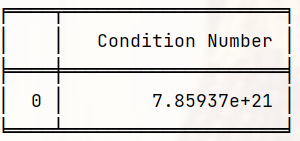
The data frame is free from null values.

**SINGULAR VALUE DECOMPOSITION (SVD)**

The covariate features seem to have very high degree of collinearity with about 19 features less than 10 in singular value.

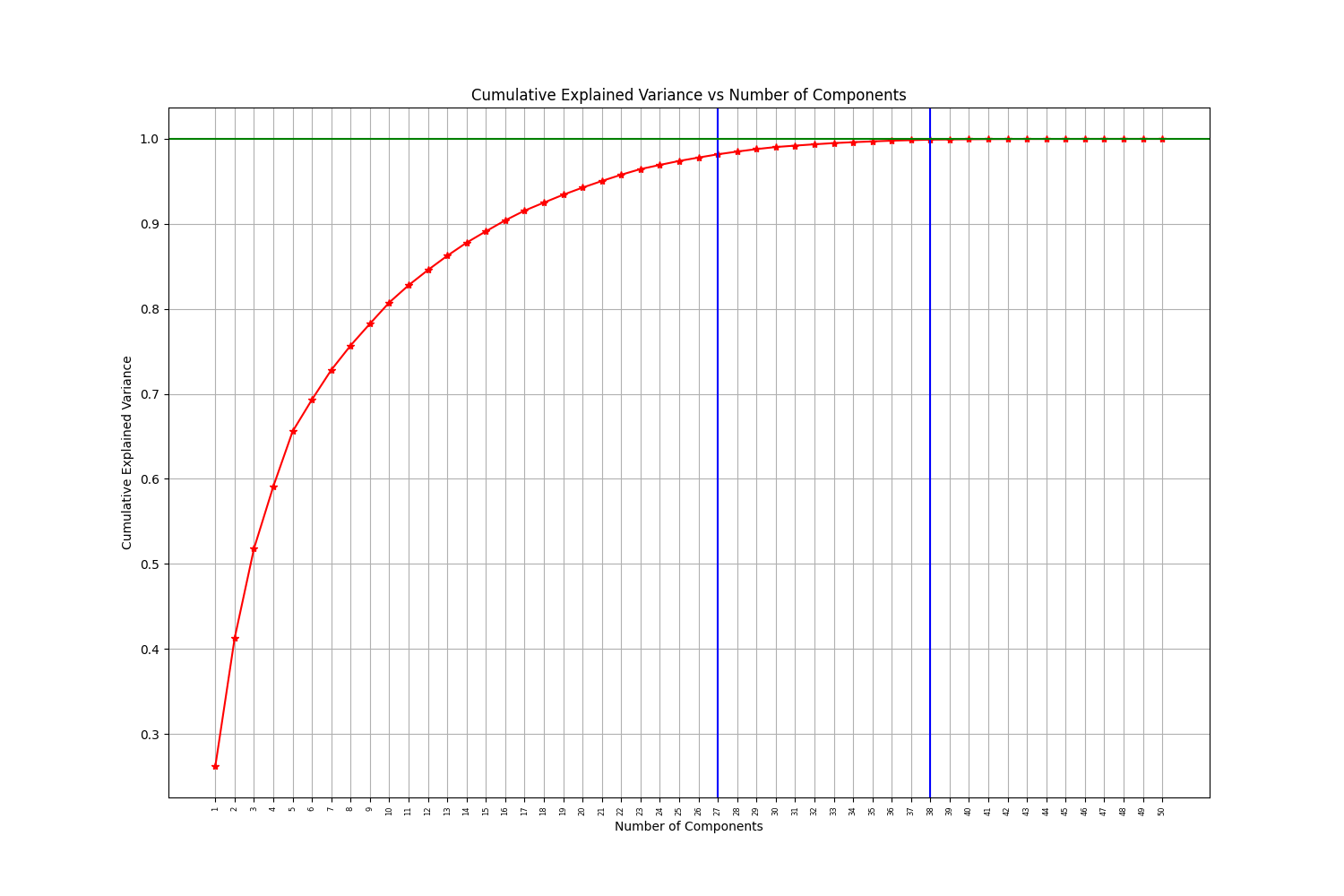
The condition number also depicts very high degree of collinearity (>1000).





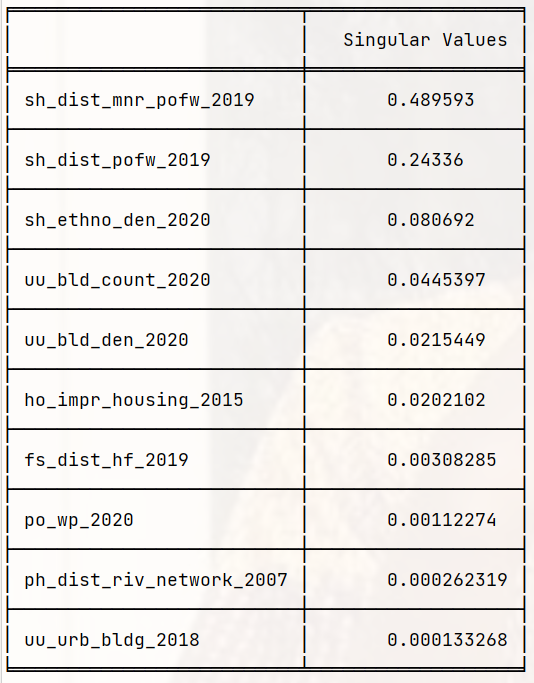
**PCA**

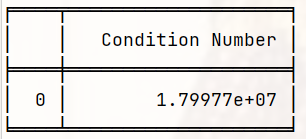
The PCA graph of cumulative explained variances below gives us an idea of the variation of original data to be explained by specific number of feature(s).



From the above, with 27 features one can get 98.18% variance of the original data while with 38 features, you get 99.91%.

SVD analysis performed on 38 important features gives:



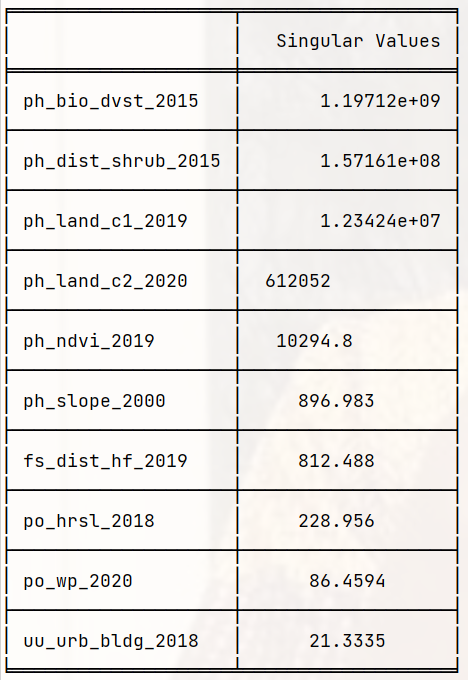


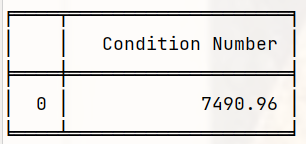
The condition number is significantly noticed to have reduced which is very good with singular values increased slightly.

**VIF**

The VIF analysis was carried out on the covariate features with output features with bearable collinearity between 1 and 5.

The analysis gave 10 features with bearable collinearity and their singular values are given below:

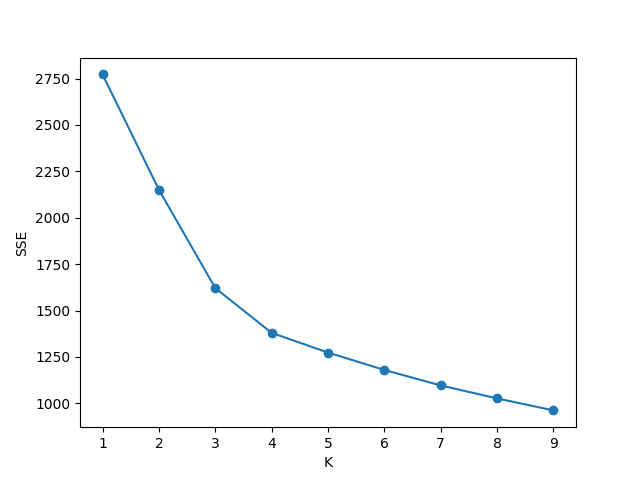


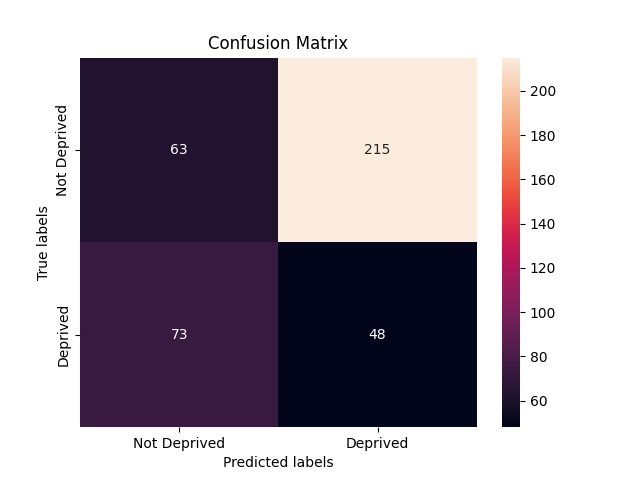


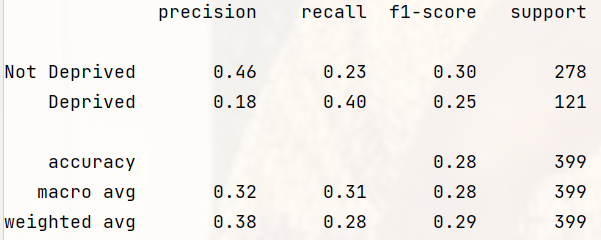
The singular values are higher with VIF result compared to PCA’s result.

Also, the condition number is far lower than PCA’s result though collinearity is still high.

**KMEANS (38 Features)**





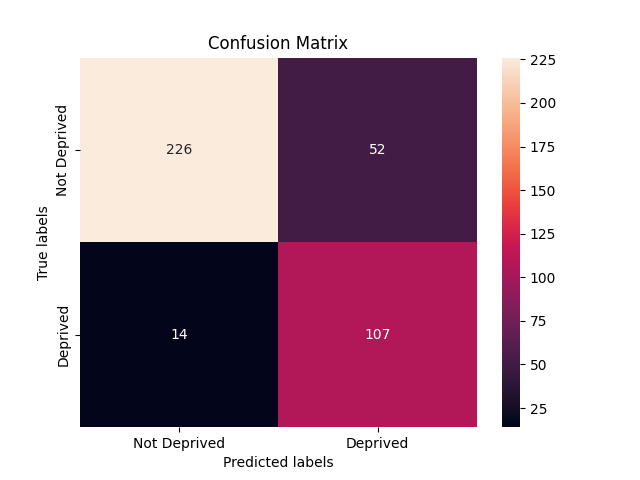


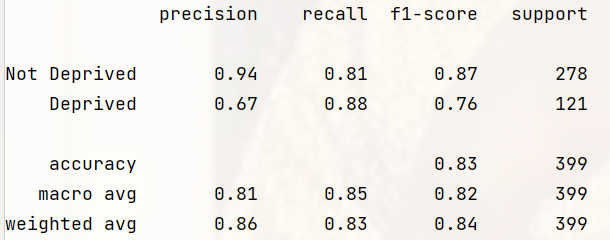
**RANDOM FOREST CLASSIFIER**

**38 Features**

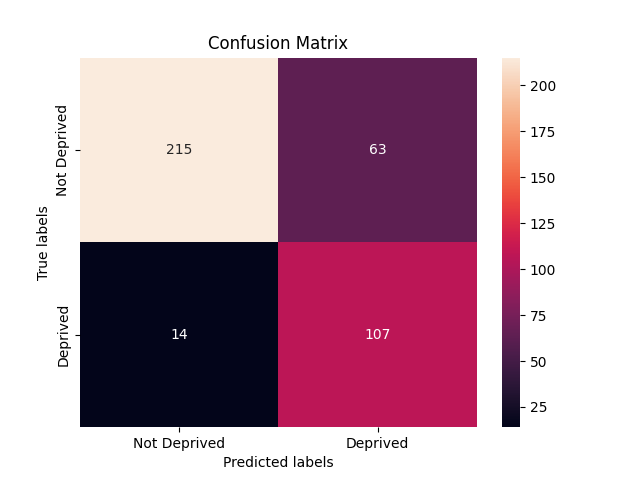
The best parameters for Random Forest Classifier are:

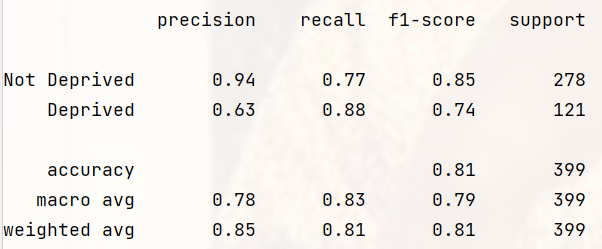
*class\_weight='balanced', max\_depth=30, min\_samples\_leaf=20, min\_samples\_split=100, random\_state=123*



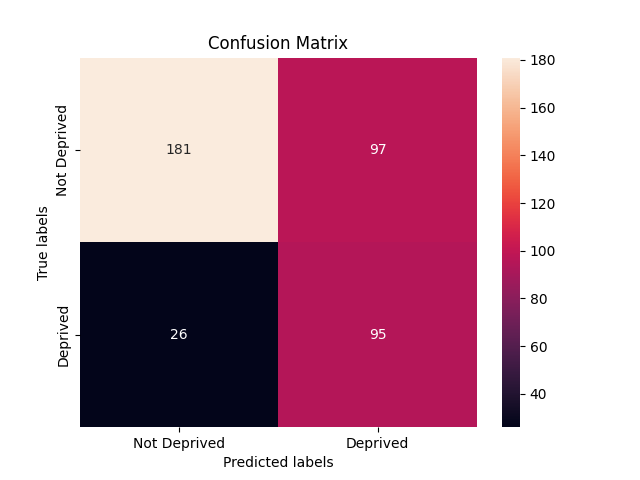


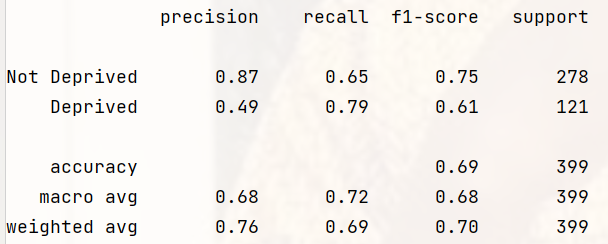
**53 Features**

****

****

**For 10 Features**

****

****

**REFERENCES**

5 Stages of Preprocessing for K-means Clustering- [Medium](https://medium.com/@evgen.ryzhkov).