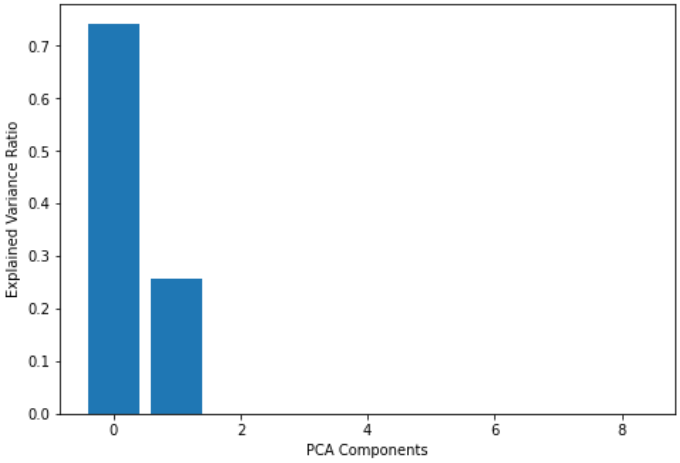
NAIVE BAYES CLASSIFIER

Data Preprocessing :

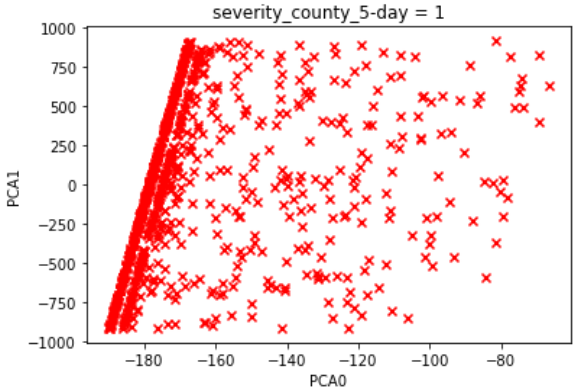
- The attributes ‘statename’, ‘countyname’ encoded using LabelEncoder.
- The attribute ‘countyfips’ dropped as it can’t be considered as a continuous variable and every row has a different value so it doesn’t provide much information as a discrete variable either.
- Missing values : There are no missing values present in the given dataset.

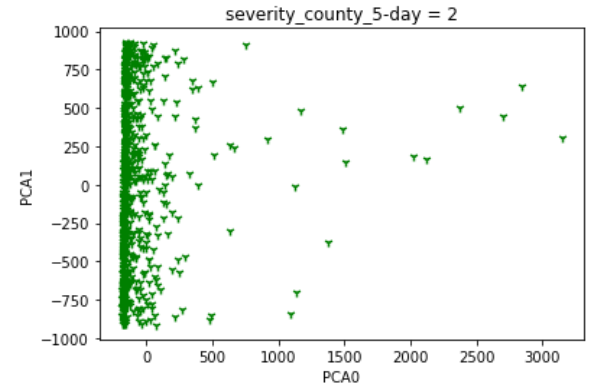
Observations :

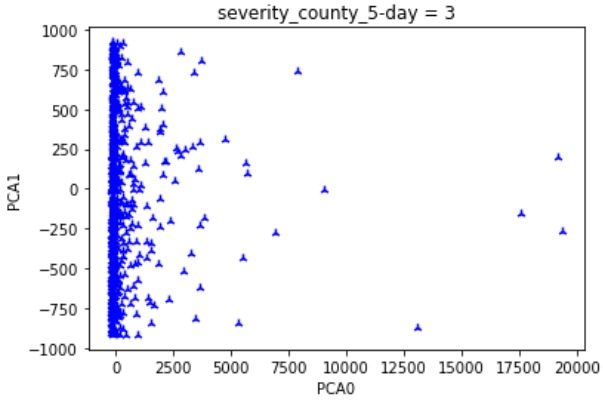
- 5 fold CV on preprocessed data gave the average score ~0.45 each time.
- The value of n = 2 is used to preserve 95% of variance after applying PCA.



- The transformed data, after applying PCA. Each label has a different plot for better visibility.







- 5 fold CV on PCA applied data was quite faster than 5 fold CV on original data with almost similar average score.
- ~30 samples were dropped as outliers.
- After applying Sequential Backward Selection, different features were removed for different splits. But the feature ‘predicted_deaths_by_october_06’ was dropped most often.
- 5 fold CV on the data after applying Sequential Backward Selection, the average score increased slightly (increase in the range 10^{-3}).
- The final test accuracy for all 3 methods varied quite a bit. (in the range 10^{-2}).

Few Conclusions :

- Looking at the not-so-good performance of Naive Bayes on this dataset, it can be concluded that independence of features is a bad assumption on this dataset.
- Similar average score for PCA data and original data indicates that it is better to apply PCA before using Naive Bayes because by using PCA the processing becomes much faster.
- On this dataset, using PCA is better than Sequential Backward Selection, because the latter takes much more time (heaviest computation in the whole assignment) and provides similar results.
- The variation of final test accuracy points to the fact that the training data is not enough.