1. EDA

* Added Precipitation for 5 regions (only East of England doesn’t have this data)

1. Feature Engineering

* Sinusoidal encoding: month\_sine, month\_cos
* Create lagged features for: high-correlated, low correlated features with “Total death”
* Apply 95% PCA on “Weather metrics” to reduce from 35 to 5 features (haven’t done loading for PCA to understand more about each PCA group)
* Plot the correlation again and remove all explanatory variables that has correlation of more than 0.5 (50%) => remove the one that has lower correlation to “Total death”
* Only 20 features remaining.
* Clustering on final dataset (2 clusters – haven’t found centroids and do further explanation)
* Determine threshold:

+ Asked Yakub about mean/std but this method needs to be backed up by scientific papers.

+ He advised to use KBinDiscretized (3 strategies: kmeans, uniform, quantile)

1. Classification models (5 models: Naïve Bayes, SVM, Random Forest, K-neighbors, Gradient Boosting)

* Train\_test\_split (80/20)

+ For 4 threshold types (I still keep mean/std for comparison with the other 3 by KBin)

+ Hyperparameter Tuning on all models.

* K-fold cross validation (10-fold since our dataset is relatively small ~ 156 rows)

+ For 4 threshold types

+ Hyperparameter Tuning on all models.

* Compare both validation methods to choose the best Model:

**BEST MODEL: between train\_test\_split and kfold: SVM**

* 1/ Death Category (kmeans) <- set threshold by kmeans of KBinDiscretized
* 2/ 10-fold Cross validation
* 3/ Because there is no difference after tuning, we'll choose the default model parameters.
* We will only apply further analysis based on this SVM Model

1. Further analysis and optimization

* Visualize and report SVM based on 3 thresholds (by kmeans strategy)
* Visualize the AUC-ROC curve for 3 thresholds
* Feature weight importance for SVM (20 features)
* Apply forward selection based on weight importance to find the top peak of all evaluation metrics => from 20 to 12 top features.
* Run the default model on these 12 features.
* Apply hyperparameter on this one again.
* Apply smote on 12 features.
* Apply hyperparameter on the SMOTE df.
* Visualize important methods based on evaluation metrics.

