Dataset - DKMA-Covid19-Jan-States
This dataset uses statistics on COVID-19 cases in US states in January of 2021.
The original data comes from the following two sources:

- John Hopkins University CSSE COVID-19:
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data
- US 2020 Census

****

The datapoints describe various information about the population and Covid pandemic situation in 50 US States and over each day of the month of January 2021. Each record gives Covid information for that day and demographic data for the whole state as well. Note the demographic data is duplicated for each state since for all days since it does not change that quickly and is based on data from the last census. Also note that the data was combined from multiple sources and not processed for uniform number formats and was not normalized. As far as we are aware there is no missing data, but there are certainly, some states which more extreme values than others, so consideration of outliers is reasonable. All of this will be initial pre-processing you will need to carry out.

File Descriptions:

- covid train.csv - main training data with input features and three binary labels.

Fields Descriptions:

- Day: Date in January 2021 ranging from Jan 2 to Jan 31.
- State ID: Arbitrary ID number for each state, based on alphabetical order. Note there are 51 states since the District of Columbia is also included.
- State: Name of the US State.
- Lat: Latitude for the geographic centre of the state.
- Long : Longitude for the geographic centre of the state.
- Active: Number of active, tracked COVID-19 cases that day in that state.
- Incident Rate: see original data source
- Total Test Results: see original data source
- Case Fatality Ratio: see original data source

- Testing Rate: see original data source
- Resident Population 2021 Census: see original data source
- Population Density 2021 Census: see original data source
- Density Rank 2021 Census: see original data source
- SexRatio: see original data source

Label Description:

Each row of this data dataset represents a particular day in January, 2021 in a particular US State and has three binary labels which are:

- True if is the corresponding count went up from the previous day.
- False if is the corresponding count went down from the previous day.

The three binary labels are:

- Confirmed: Was there a daily increase in the confirmed total cases of COVID-19 in that state on that day?
- Deaths: Was there a daily increase in the number of deaths from COVID-19 in that state on that day?
- Recovered: Was there a daily increase in the number of people recovered from COVID-19 in that state on that day?

Note that the labels are fairly unbalanced, so there are quite a few more True cases than there are False cases. Recovered is the most balanced, so it should be easier to train. Deaths is next and Confirmed is the least balanced.

Main TASK:

1 Data Pre-processing and Preparation

Report a short summary of the pre-processing steps you applied to the data to make it usable for analysis. This should at a minimum include discussion of outliers and normalization (which features, what kind of normalization, if any, is needed and appropriate?). What do you do with "Day", "State" and "State ID"?

2 Representation Learning

Apply PCA and LDA onto the dataset, on an appropriate subset of the features. For LDA use each of the labels individually to train it.

- Use a scree-plot to look at the cumulative variance represented by the PCA eigenvectors, give advice on the best number number of reduced features to use to represent the data.

- Plot the datapoints, or a useful subset of them (select out one day per state? or use color by State?) on the first two PCA and LDA feature vectors.
- Does the LDA method provide better results for one label more than the others?

Now, use the new hybrid dataset which you construct which replaces (some or all) of the numeric value features with a subset of PCA or LDA features.

3 Random Forest Classifier

Classify the data using random forest, Tune the hyper-parameters of the classifier using k-fold cross validation and sklearn functions. Try a range of parameters at least including the following:

- number of trees: f5, 10, 50, 150, 200g
- max depth: f3, 5, 10, None

For this, the plot should be a heat plot. You should have (5 * 4) mean accuracies for different values of number of trees and maximum depth.