Machine Learning Project Poster | Priya Bannur | PB 23

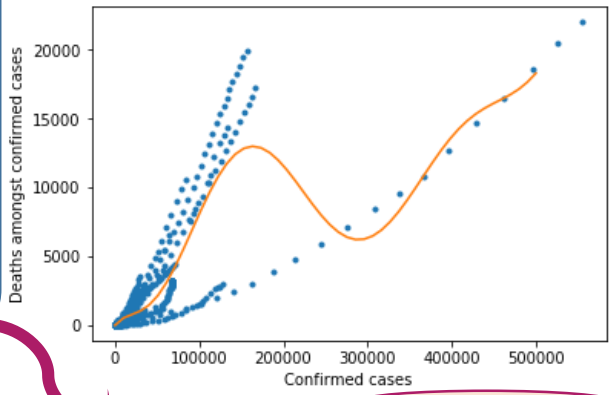
Data Description

- Region wise Confirmed, Recovered and Death cases
- Patient condition information of South Korea
- Patient symptoms data and virus test report

Preprocessing

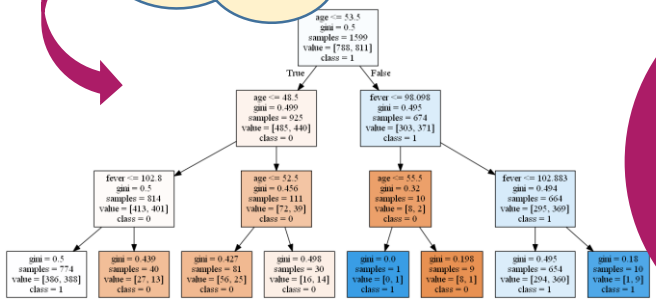
- ✓ SimpleImputer to handle missing data
- ✓ Geocoding location information
- ✓ Label Encoding
- ✓ Principal Component Analysis

Polynomial Regression



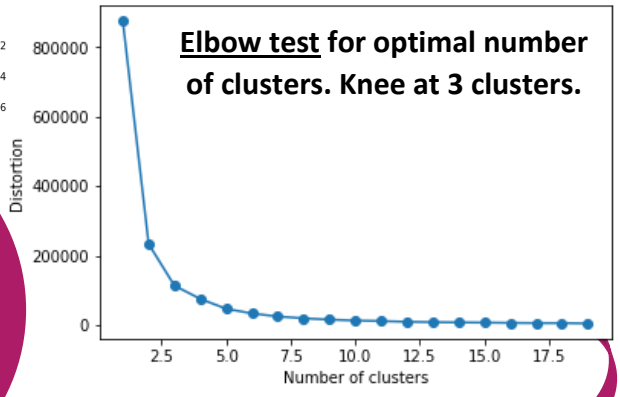
Classification

Decision Tree
Random Forests
Gaussian NB
KNN
Ensemble



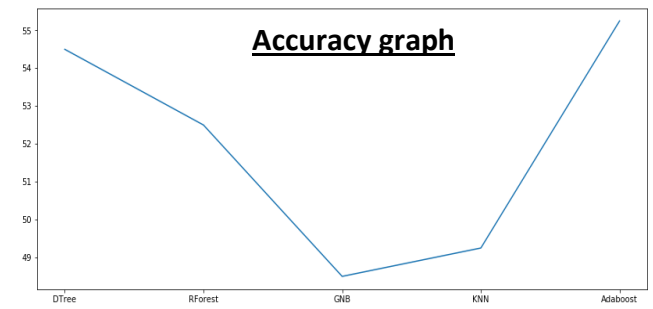
PCA

K-Means Clustering

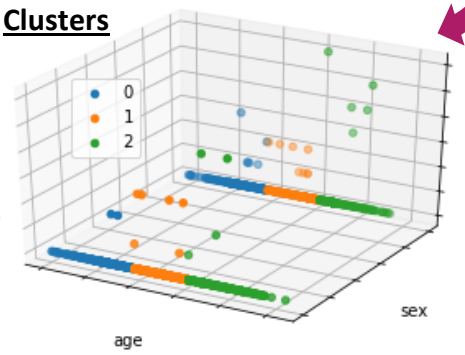


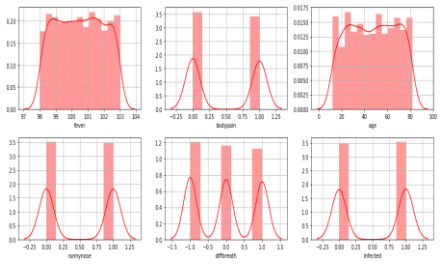
COVID19 Analysis

Accuracy graph



Clusters





OBJECTIVES

1) Pre-process raw/naïve Coronavirus data.
   a) Handle missing values using Imputer.
   b) Label encoding sex column.
   c) Find most contributing attributes by PCA.
2) Regression analysis on death rate.
3) Elbow test to determine optimal no. of clusters & perform K-Means clustering.
4) Compare different classification algorithms and analyse their accuracies.

CONCLUSIONS

1) SimpleImputer class can be used to handle missing text data.
2) Geopy library can be used to encode textual location data into numerical lat & long.
3) Death rate can be predicted using Polynomial regression of degree 8.
4) 3D clusters analysed (sex,age,infection_order).
5) Decision Tree gave maximum accuracy, hence used in Adaboost to improve further.