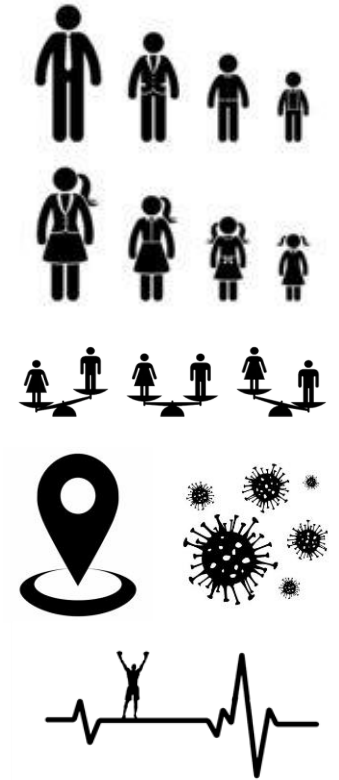
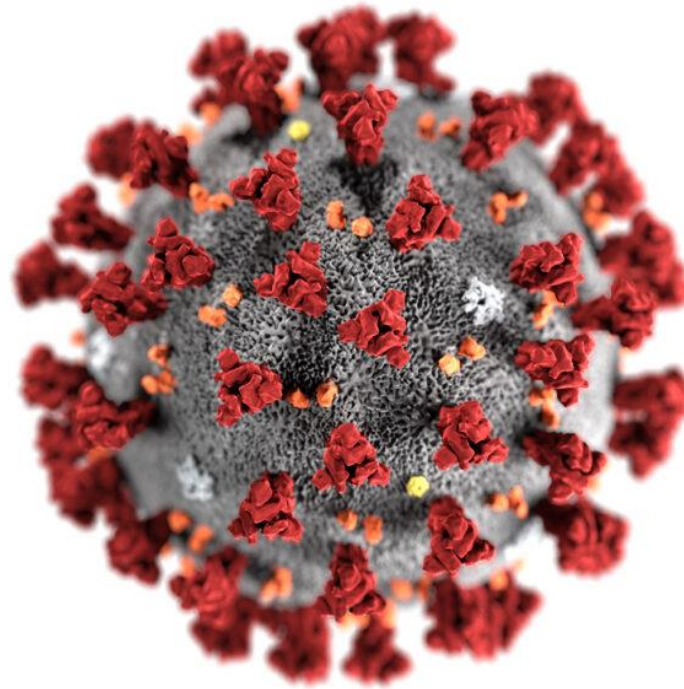


# Coronacast

COVID-19 viral strain  
patient outcome predictor

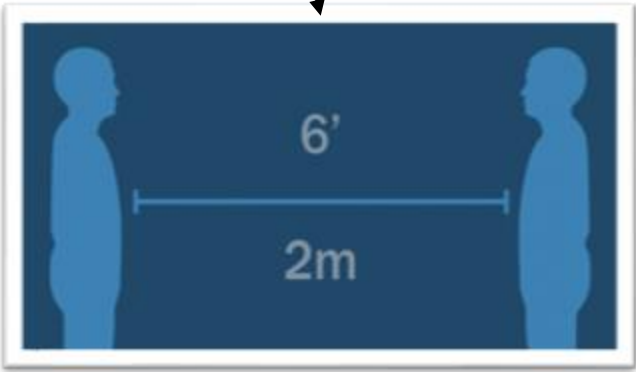
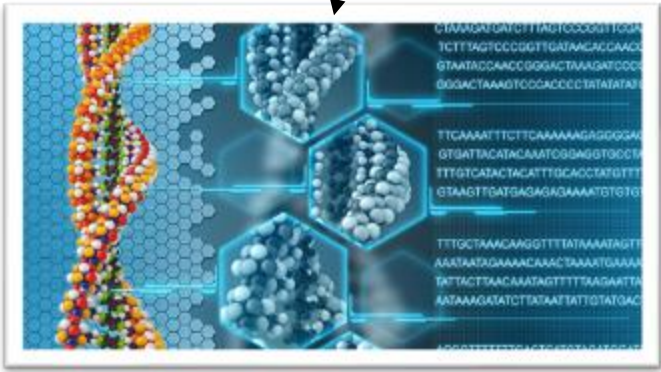


**Omic Consulting Project**

Azadeh Kamali Tafreshi



Is viral sequences matter



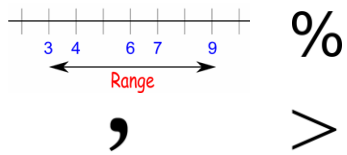
# Data Pre-Processing

## Location



Formatting  
Grouping #84 features  
decreased to #43

## Age & Mutations



Range: Average  
Months: Year  
Digits: Rounded

## Imputation

Age:  KNN  
Imputer

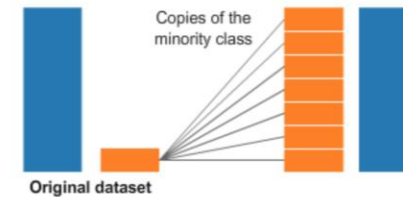
Gender: Third category  
of missing was added

Unknown mutations:  
Dropped


## Balancing the data

Deceased: #41  
Recovered: #969

SMOTE  
Oversampling



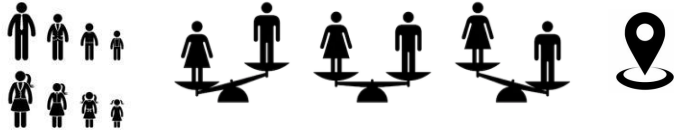
## Scaling

 Standard Scaler

The standard score of a  
sample x is calculated as:

$$z = (x - \mu) / s$$

# Random Forest



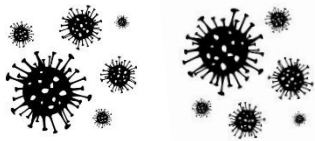
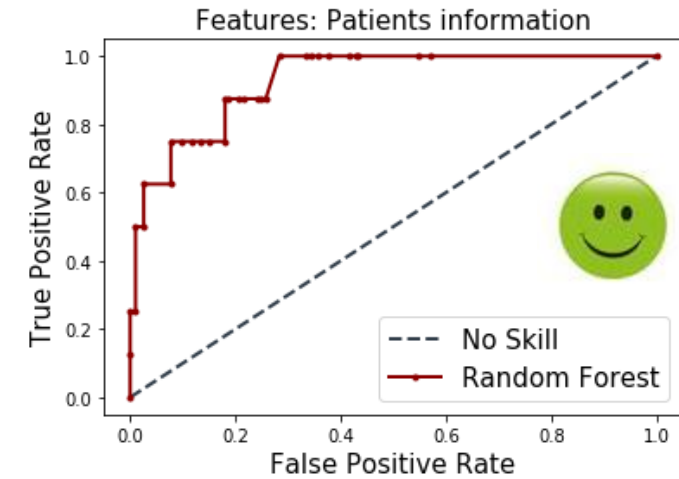
# 197 features vs. # 1010 samples



Metric : Area under the ROC curve

AUROC: 93 %

Recall: 96 %

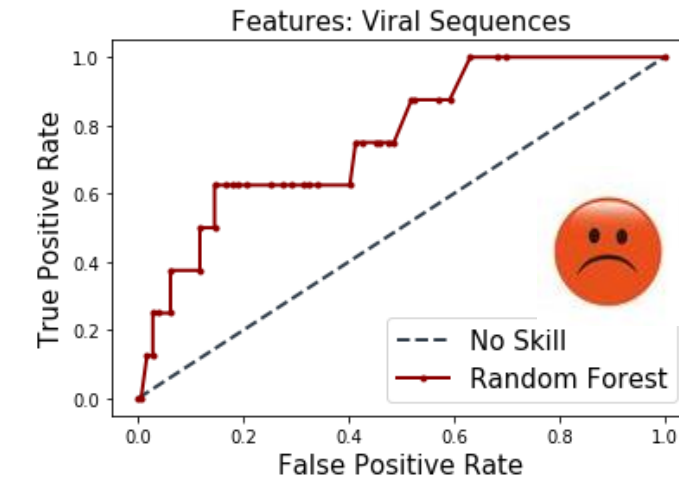


# 410 feature vs. # 933 samples



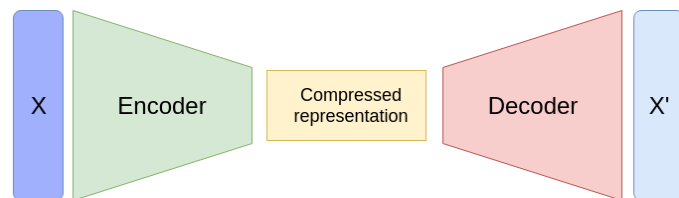
AUROC: 76 %

Recall: 89 %

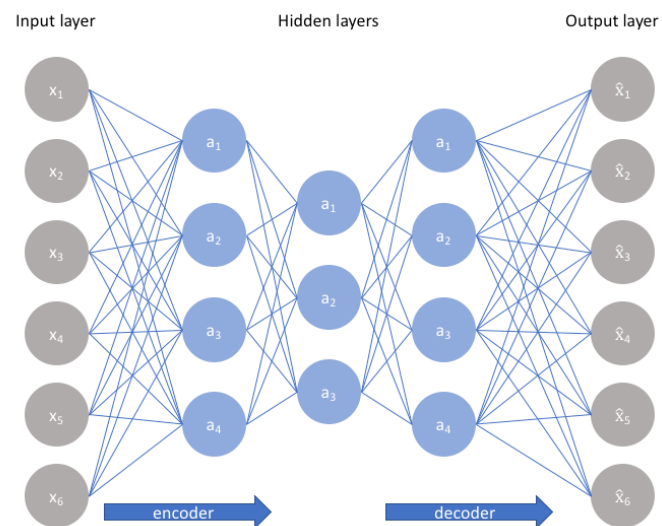


# Random Forest

## Autoencoder for Viral Sequences Keras

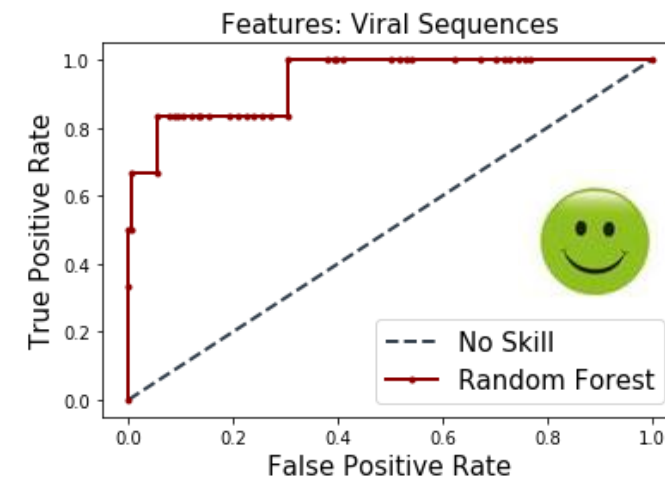


- Unsupervised NN learning
- Non-linear dimension reduction



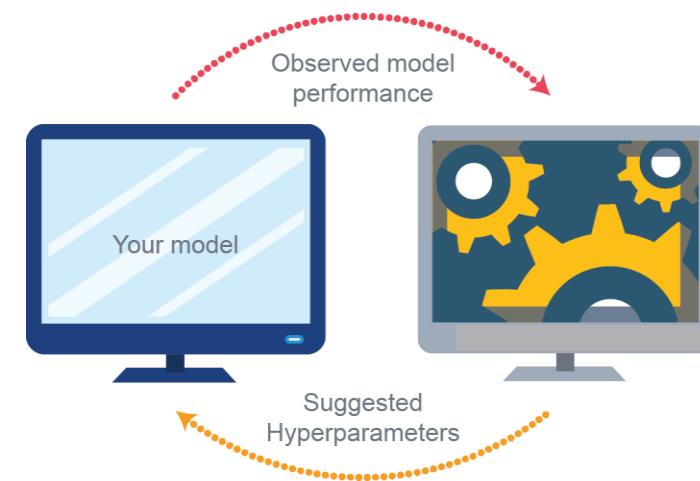
AUROC: % 94

Recall: % 92



~~AUROC: % 76~~

~~Recall: % 89~~



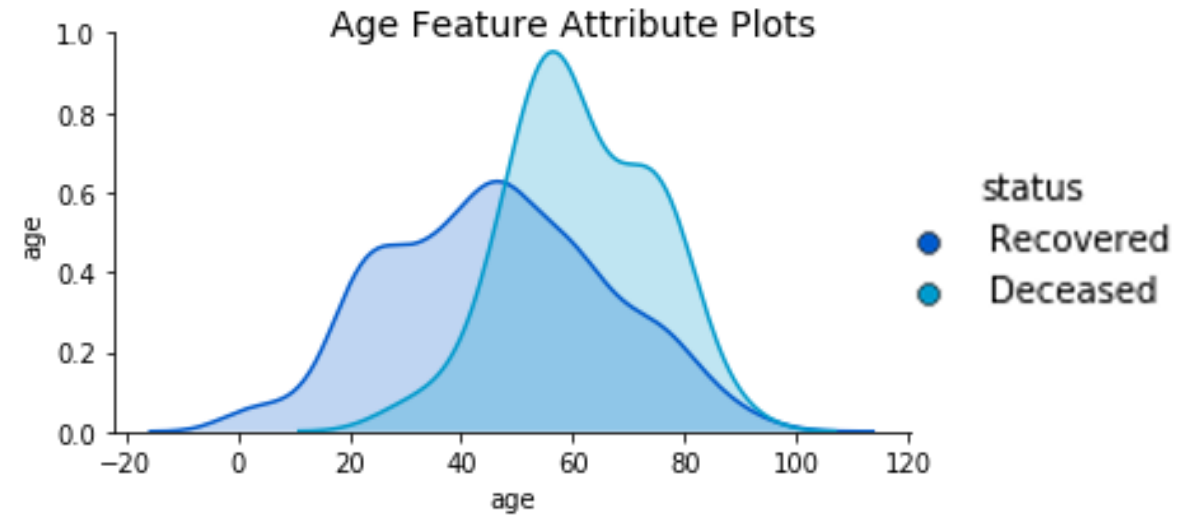
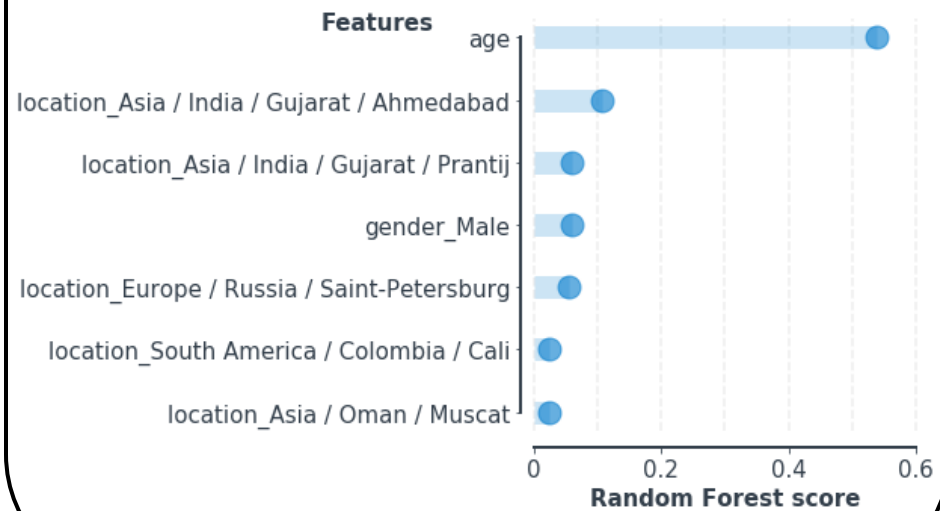
# Feature Importance



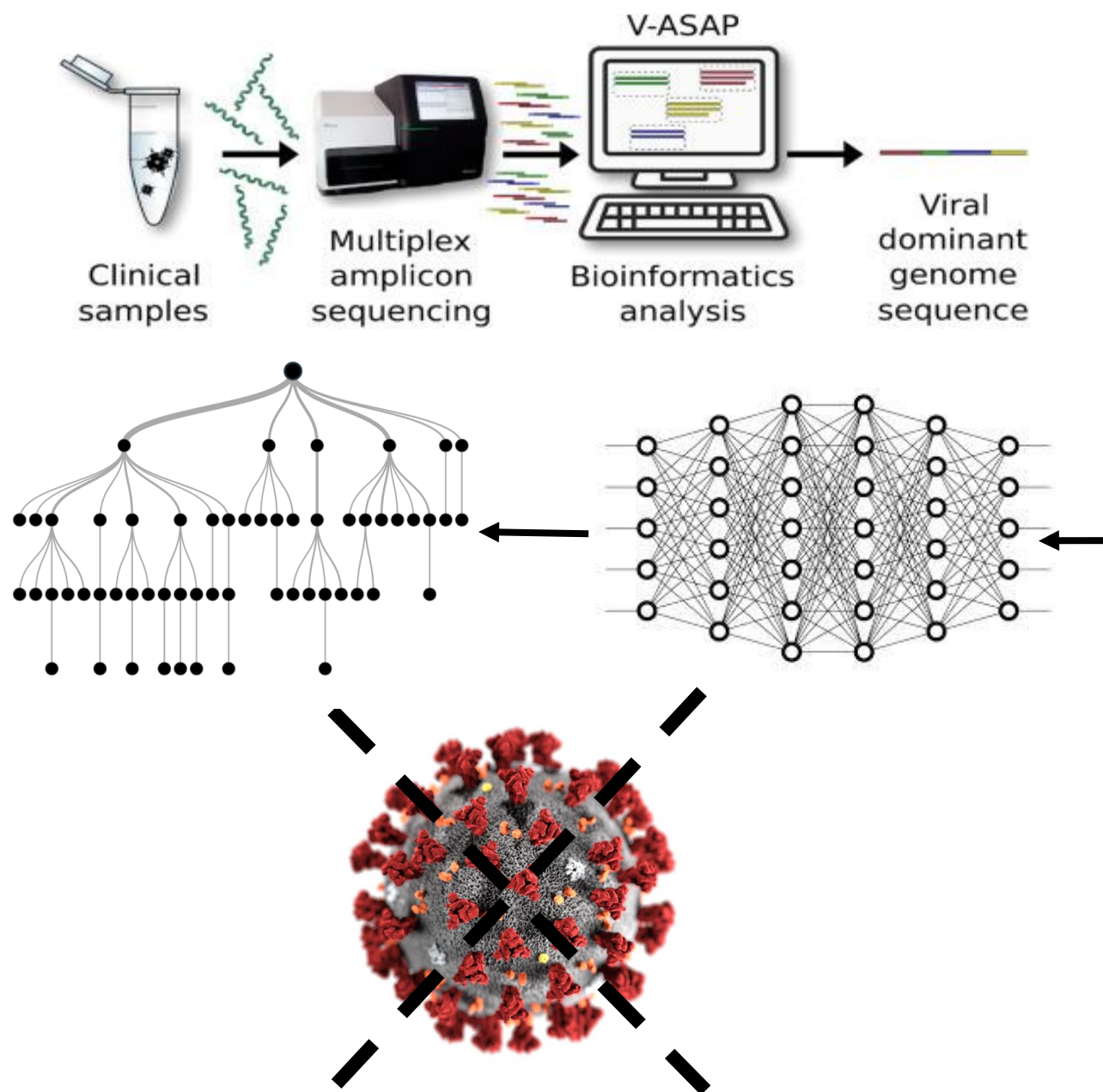
Random Forest



matplotlib Seaborn









**Azadeh Kamali Tafreshi**

PhD in Electrical & Electronic Eng.



Electrical and Electronics Eng. Dep.

