**Data Exploration and Preprocessing**

Given the progress of the project, the main Machine Learning Analysis is based on 3.5 million rows of data – a sample of the full database.

Our hypothesis is that, patients who have medical underlying conditions are more likely to run into serious situations, even death from COVID infection, and are supposed to be classified as high risk group.

After eliminating the negative cases, we got more than 1.4 million rows of data and 41 columns.

Furthermore, we eliminated the columns with irrelevant info, and encoded all the medical conditions, and target column [“Date of Deceased”], removed the outliers and scale down the [‘age’] column into following groups:

*(# Child(0-16) : 1*

*# Young\_Adults(17-32) : 2*

*# Middle\_age\_Adults(33-48) : 3*

*# Old\_age\_Adults(49-64) : 4*

*# Senior\_Adults(65-90) : 5)*

**Preliminary Feature Selection**

Before fitting the model, we used R to perform a preliminary feature analysis. The result shows almost all the features/medical conditions are significantly relevant to the target ['Death']. More details will be disclosed by Safaa.

*(Even though ['pregnant'] and ['cardiovascular'] doesn’t show its significance, we still decided to include all the initial features as all of them are considered valuable info for doctor’s diagnosis, allocating medical resources. )*

**Machine Learning Analysis**

As we found only 11% of the data is in ‘High risk’ class, we are focused on the following models and different resampling method to handle the imbalanced dataset.

***(Models for comparison:***

1. ***Logistic Regression*** *– baseline for classification problem;* ***(Drawback is assumption of linearity between the dependent and independent variables)***
2. ***SVM*** *– handle non-linear problem as well; Handle Outlier well;* ***(Quite time consuming)***
3. ***Decision Tree*** *– Handle non-linear problem effectively; Cases of missing values,* [*outliers*](https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm)*, feature scaling have less significance on the decision tree’s data; Easy to understand and interpret; (****tend to get overfitting)***
4. ***Ensemble learning****: including:*

* *Random Forest*
* *Gradient Boosting*
* *Balanced Random Forest*

*(The classifier is an ensemble method in which each tree of the forest will be provided a balanced bootstrap sample. The balancing is achieved by random under-sampling)*

* *Easy Ensemble (The classifier is an ensemble of AdaBoost learners trained on different balanced bootstrap samples. The balancing is achieved by random under-sampling.)*

***Resampling:***

*SMOTE, SMOTEENN, Undersampling, original*

*SMOTEENN: Class to perform over-sampling using SMOTE and cleaning using ENN. Combine over- and under-sampling using SMOTE and Edited Nearest Neighbours.)*

**Evaluation Metrics:**

Given the use case of the project, we would rather be more aggressive in classifying patients than miss any potential high risk patients. So Recall is the metrics we are focusing on.

**Initial training and validation**

Based on the listed models, we started with the K fold(n\_splits=10) cross validation on 30,000 rows of data, and compared Recall of different models with resampling methods.

The result shows both SMOTE and Undersampling performed well across all models while SMOTEENN was showing large amount of variance in Recall. Comparing with the validation without any resampling, we can confirm that Resampling significantly improves the model performance.

We also found that the Easy Ensemble and Balanced Random Forest stay consistent across different resampling methods due to natural integration of Undersampling.

**Further Training and Validation**

Furthermore, we increased the sample data to 50,000 rows, and used StratifiedKFold(n\_splits = 5) validation and test datasets to evaluate (25% of the sample data) all 7 models with both SMOTE and Undersampling.

Based on the best recall score, we decided to use the Balanced Random Forest as our prediction model.

**Conclusion**

Finally, we ran the 1.4 million rows of data into the Balanced Random Forest, and got the 88% of recall and 86.57% of the balanced accuracy score in almost 1 min.

Before we reach the conclusion, we also did Feature engineering and other optimization attempt. I will hand it over to Safaa for further explanation.

**Side notes:**

**Further improvement –**

* Further evaluate the models with 6.5 million rows of full database
* Fine tune parameters, including max\_depth, n\_estimators using GridSearchCV
* Evaluate the time used for training and predicting process of each model for multi-dimensional comparison

**Other Experiment and Conclusion:**

We also tried the Deep Learning with 2 hidden layers, and relu activation function, but got the similar recall and balanced accuracy score as the rest of the Machine Learning Models.So we didn’t include the DL details in the comparison. **(DL: robust in figuring out the classification, especially in detecting images and language.** **(Drawbacks: time, need more fine-tuning)**

However, we noticed with SMOTE, the time spent on resampling the dataset is significantly more than the Undersampling method. Resample 1.4 million rows of data using SMOTE takes 10mins.

**The reason to use bagging over boosting:**

It effectively reduced the model variance, and avoid overfitting in front of noise data; Ensures consistency and improves the accuracy of machine learning algorithms used in statistical classification and regression. It’s often used in HealthCare Industry like disease prediction using the patient’s medical history.

**Healthcare:**Bagging has been used to form medical data predictions. Ensemble methods are used for gene and/or protein selection to identify a specific trait of interest. Random forests are used for various purposes in the healthcare domain like disease prediction using the patient’s medical history.

**DISADVANTAGES**: Bagging consumes more computational power than boosting (becomes computationally expensive) since we must utilize numerous models, and it may not be acceptable in certain use scenarios. If it is not adequately modeled, it may result in high bias and therefore underfitting.

As the scope of the project keeps expanding, we will need to evaluate the timer of each models handling millions of rows of the datasets. The model will change over time to address further challenges including computational power constraints.