MLM-1909 PROJECT

COPD Exacerbation Triage

using R

By KRUTI SHAH

Kruti Shah – MLM -1909

1. Data Source /Lineage and Description

This problem is taken from the research paper **A machine learning approach to triaging patients with chronic obstructive pulmonary disease**

The dataset is extracted from the **Supporting information** section of the paper.

It contains 2 data sets

- Copd –This dataset if formed from the simulated data prepared by the 6 pulmonogists engaged on this experiment for the research.

  The final variable list is shown in Table 1, and includes

  1) patient background characteristics that are associated with COPD exacerbation risk and severity,

  2) current clinical symptoms that encompass widely accepted features of exacerbations, and

  3) physiologic measurements that are predicted to influence physician perception of exacerbation severity.

  The Copd dataset is used for training the supervised algorithms. The Copd dataset consists of 3200 observations.

Variable		Units—Type
Patient Profile	Age	years—continuous
	Weight	lb—continuous
	Height	feet + inches—continuous
	Gender	Male/Female—categorical
	COPD GOLD STAGE	1,2,3,4—categorical
	Baseline MMRC Dyspnea	1,2,3,4,5—categorical
	Recent Exacerbations & Hospitalizations	Yes/No—categorical
	Lives Alone?	Yes/No—categorical
	Smoker	Yes/No—categorical
	Long-Term Oxygen User	Yes/No—categorical
	Assisted Daily Activity	Yes/No—categorical
Comorbidities	Congestive Heart Failure	Yes/No—categorical
	High Blood Pressure	Yes/No—categorical
	Coronary Artery Disease	Yes/No—categorical
	Diabetes	Yes/No—categorical
	Anemia	Yes/No—categorical
	Pulmonary Hypertension	Yes/No—categorical
	Acid Reflux	Yes/No—categorical
Symptoms	Shortness of Breath	1,2,3—categorical
	Cough	1,2,3—categorical
	Wheezing	1,2,3—categorical
	Change in Sputum Color	Yes/No—categorical
	Increased Sputum Volume	Yes/No—categorical
	Cold/URI	Yes?NO—categorical
	Medication Compliance	1,2,3—categorical
	Sleeplessness	Yes/No—categorical
	Current MMRC Dyspnea	1,2,3,4,5—categorical
Vital Signs	Oxygen Saturation	%—continuous
	FEV1	Vol/sec—continuous
	Heart Rate	BPM—continuous
	Temperature	°F—continuous

https://doi.org/10.1371/journal.pone.0188532.t001

Table 1 – List of patient profile, comorbidity, vital sign, and symptom factors, with respective measures, used in the COPD triage and exacerbation algorithms.

- Copd_dr

This dataset is similar to Copd dataset except that it has first field as the doctor's identification field. It provides the information as to which physician provided the data. 100 records of COPD patients from each doctor and such 10 doctors are being considered for this research. This dataset is used for validation of the supervised algorithm. It consists of 1000 (10*100) observations. The use of consensus physician opinion as a validation standard and the

analysis of individual physician performance on that standard are unique for this dataset

2. Context / background

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease that over time makes it hard to breathe.

2.1 Key Points

1. COPD is chronic. In other words, you live with it every day.
2. It can cause serious long-term disability and early death.
3. There is no cure for COPD, but it is often preventable and treatable.
4. COPD is referred to as chronic bronchitis or emphysema.

With COPD, the airways in your lungs become inflamed and thicken, and the tissue where oxygen is exchanged is destroyed. The flow of air in and out of your lungs decreases. When that happens, less oxygen gets into your body tissues, and it becomes harder to get rid of the waste gas carbon dioxide. As the disease gets worse, shortness of breath makes it harder to remain active.

Sometimes referred to as either chronic bronchitis or emphysema, most people will have symptoms of both conditions, so health professionals prefer to call the disease COPD. However, some doctors think that chronic bronchitis may be present even though a person does not have the airway obstruction characteristic of COPD. It is important to remember that in many cases, COPD can be prevented and can be treated.

In this study (research paper), the authors present a machine learning-based strategy for early detection of exacerbations and subsequent triage. Their application uses physician opinion in a statistically and clinically comprehensive set of patient cases to train a supervised prediction algorithm. The accuracy of the model is assessed against a panel of physicians each triaging identical cases in a representative

patient validation set. Their results show that algorithm accuracy and safety indicators surpass all individual pulmonologists in both identifying exacerbations and predicting the consensus triage in a 101 case validation set. The algorithm is also the top performer in sensitivity, specificity, and when predicting a patient's need for emergency care.

I am using this use case and the dataset provided to demonstrate my knowledge of different machine learning methods learnt in the MLM class.

3. Business problem and question

COPD affects 12.7 million Americans and costing nearly $50 billion annually. Each year, COPD leads to more than 700,000 hospitalizations Approximately 19.6% of patients hospitalized in the United States are readmitted within 30 days, accounting for an estimated expense of $17 billion annually . In an attempt to improve clinical outcomes and control rising healthcare costs, the Centers for Medicare and Medicaid Services (CMS) developed the Hospital Readmission Reduction Program (HRRP), which imposes financial penalties on hospitals with excessively high readmission rates for specific conditions, such as heart failure (HF), acute myocardial infarction (AMI), and pneumonia.

These CMS imposed penalties are geared to begin the process of moving to a patient-centric, disease management system in which the patients are trained and supported as they learn to participate in the management of their illness. To develop a tool to triaging patients with COPD to reduce the occurrence and re-occurrence of exacerbations requiring multiple hospitalizations will help Optum develop cost-saving strategies as it relates to COPD exacerbation.

Question: Can we predict COPD exacerbations based on the provided data i.e. patient background characteristics, current clinical symptoms, and physiologic measurements using the machine learning methods?

Based on these predictions, COPD exacerbation rapid action plan from the data can be developed.

This rapid action plan can be used to

a. educate and train the patient and family to self-monitor and
b. self-manage post discharge to lower the likelihood of unnecessary hospitalization for patients with COPD

4. Value proposition

Using the results of this project, a predictive analytical tool COPD exacerbation triage tool can be developed. The tool can help reduce exacerbations and triage the exacerbations effectively. This will in turn reduce the hospital readmissions and also improve the day to day life of the COPD patients.

In one of the referenced article, it mentioned that every year $17 billion is spent on the COPD hospital readmissions. Let's assume that we can reduce partial readmissions, let's take 25% for the value calculations, that would be savings of approx. **$4.25 Billion.**

The value acquired by the improvement in the life of the COPD patient is priceless!

5. Hypothesis

Ho: There is no significant relationship between COPD exacerbation and patient profile, comorbidities, symptoms and vital signs features.

H1: COPD exacerbation can be predicted from the patient profile, comorbidities, symptoms and vital signs features.

6. Exploratory data analysis (EDA)

6.1 Data cleanup and transformation

1. The Copd dataset had duplicate records which were removed.
2. 17 new fields are created and added to the dataset. These fields will further be used to hold factorial values based on the analysis of various fields.
3. Height field is converted into inches by converting the feet into inches and adding inches in Height2 field.
4. The Risk factor fields had text data which was converted into factorial data. New fields were added and those fields being set to factorial value based on the Risk factor field value.
5. BMI field is converted into numeric.
6. Baseline Dyspnea field values are converted into factorial '1'. '2'. '3','4','5'.
7. Baseline MMRC field values are converted into factorial '1'. '2'. '3','4','5'.
8. Categorize variable (Gold Stage) into groups. 0 represents the unknown value.
9. Separate variable (Sputum) into two variables: Sputum color and Sputum production.

10. Categorize variable (Current MMRC) into groups.
11. Categorize feature (Recent Worsening) into "less than 24 hours", “1 to 3 days and "more than 3 days". Randomly assign the value in hours based on the group.
12. Convert the unknown value to “Same As Usual” for feature (Controller Medication).
13. Convert the unknown value to “0” for feature (Current Pulse Oxy).
14. Convert the unknown value to “0” for feature (Current Heart Rate).
15. Convert the unknown value to “0” for feature (Current Temperature).
16. Convert the unknown value to “0” for feature (Confidence2).
17. Convert the unknown value to “0” for feature (Confidence1).
18. Convert the unknown value to “0” for feature (Final Triage).
19. Convert the unknown value to “0” for feature (Final Triage 2).
20. Convert the unknown value to “0” for feature (Current FEV1).
21. Calculate the triage value based on the following function: floor(Confidence1 * Final Triage + Confidence2 * Final Triage 2 / Confidence1 + Confidence2)*100)
22. Calculate the triage levels based on the triage value.

7. Setup for analysis

7.1 Defining the baseline / control / denominator

Positive Control

The variables “Current FEV1,” “Shortness of Breath,” “Change in Sputum Color,” “Cough,” “Wheezing,” ”Recent Worsening Time,” and “Current Pulse Oxy” can form a control group that contains patients who are diagnosed as COPD exacerbation.

Negative Control

The same variable set can be used to form the negative control group for patients who are diagnosed as not COPD exacerbation. The following table summarizes the control level of variables at each control group:

	Predictors	Positive Control Group COPD Excarbation	Negative Control Group Not COPD Excarbation
Control	Current FEV1	Less Than 30	Greater Than 80
	Shortness of Breath	More Than Usual	Less Than Usual
	Change in Sputum Color	Yes	No
	Cough	More Than Usual	Less Than Usual
	Wheezing	More Than Usual	Less Than Usual
	Recent Worsening Time	More Than 3 Days	Less Than 24 hours
	Current Pulse Oxy	Less Than 85	More Than 90

Feature Selection

A. **To assess whether the candidate predictors are significantly associated with COPD triage, Pearson's Chi-squared test and Pearson's product-moment correlation is ran to examine the dependency between predictors and the predicted variable "Triage"**

The following table summarizes the dependency between predictors and the COPD triage. Since "Gender," "Age," "Height," "BMI," and "Weight" are not significant, they are removed from the model.

Predictors	P-Values	Predictors	P-Values
Congestive Heart Failure	0.005	Current MMRC Dyspnea	< 0.0001
High Blood Pressure	0.0018	COPD GOLD STAGE	< 0.0001
Coronary Artery Disease	0.0006	Baseline MMRC Dyspnea	0.0019
Current Oxygen Saturation	< 0.0001	Medication Compliance	0.0003
Current Temperature	< 0.0001	Sleeplessness	< 0.0001

Recent Exacerbations Hospitalizations	0.0015	Change in Sputum Color	< 0.0001
Long Oxygen User	0.0107	Smoker	0.0042
Lives Alone	0.0021	Diabetes	0.0074
Shortness of Breath	< 0.0001	Anemia	< 0.0001
Assisted Daily Activity	0.0018	Cough	< 0.0001
Pulmonary Hypertension	0.0017	Age	0.2093
Change in Sputum Color	< 0.0001	Weight	0.5979
Hour Recent Worsening	< 0.0001	Height	0.3401
Increased Sputum Volume	< 0.0001	Gender	0.1981
Current FEV1	< 0.0001	Infection	< 0.0001
Current Heart Rate	< 0.0001	BMI	0.8545
		Acid Reflux	0.0031

B. Multicollinearity test with Variance Inflation Factor (VIF) and Principal Component Analysis (PCA)

Multicollinearity is the occurrence of high intercorrelations among independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model. The issue of multicollinearity can be mitigated by the dimensional reduction. Here VIF and PCA is used for determining multicollinearity.

A VIF for a single explanatory variable is obtained using the r-squared value of the regression of that variable against all other explanatory variables:

$$VIF_j = \frac{1}{1 - R_j^2}$$

A VIF is calculated for each explanatory variable and those with high values are removed. The definition of 'high' is somewhat arbitrary but values in the range of 5-10 are commonly used.

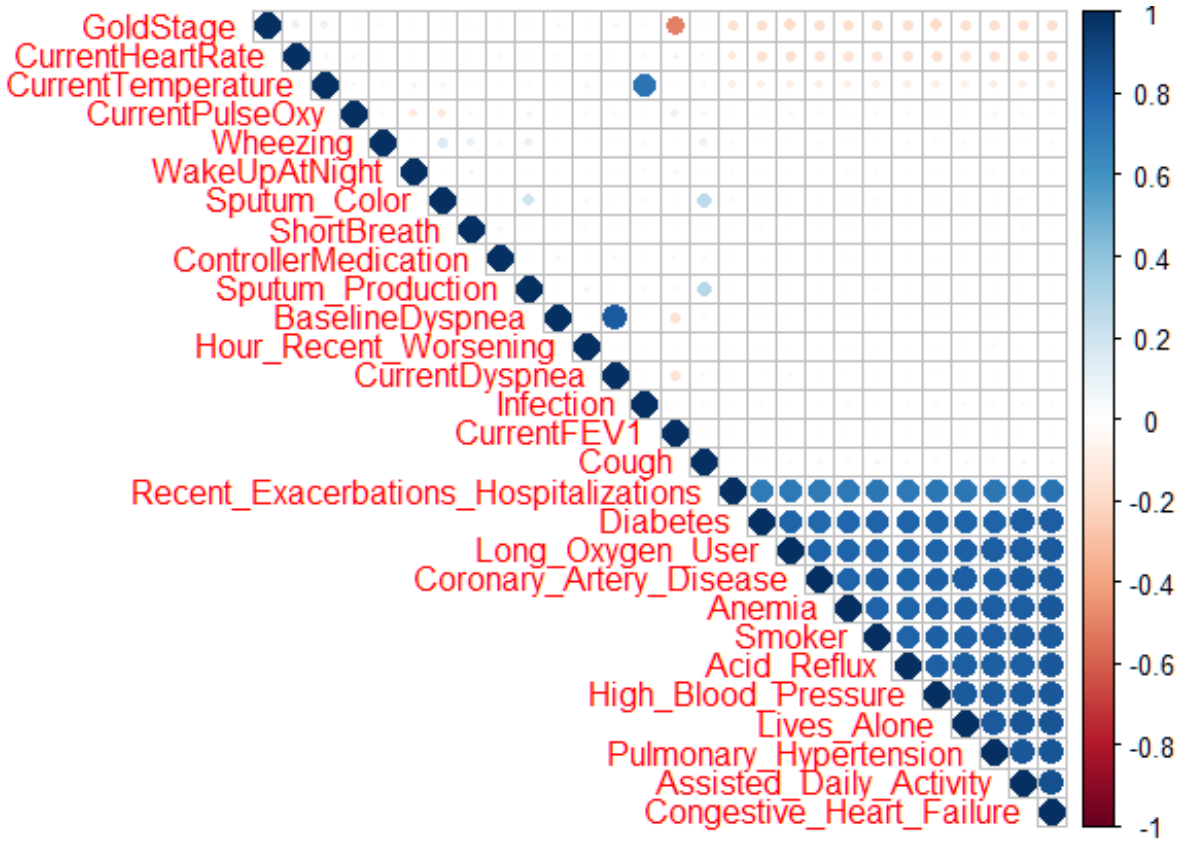
```
copd.pca <- subset(copd.pca, select = -c(Triage_Level))

fit <- lm(Triage ~ ., data = copd.pca)

sqrt(vif(fit)) > 3

copd.pca <- subset(copd.pca, select = -c(Triage))
```

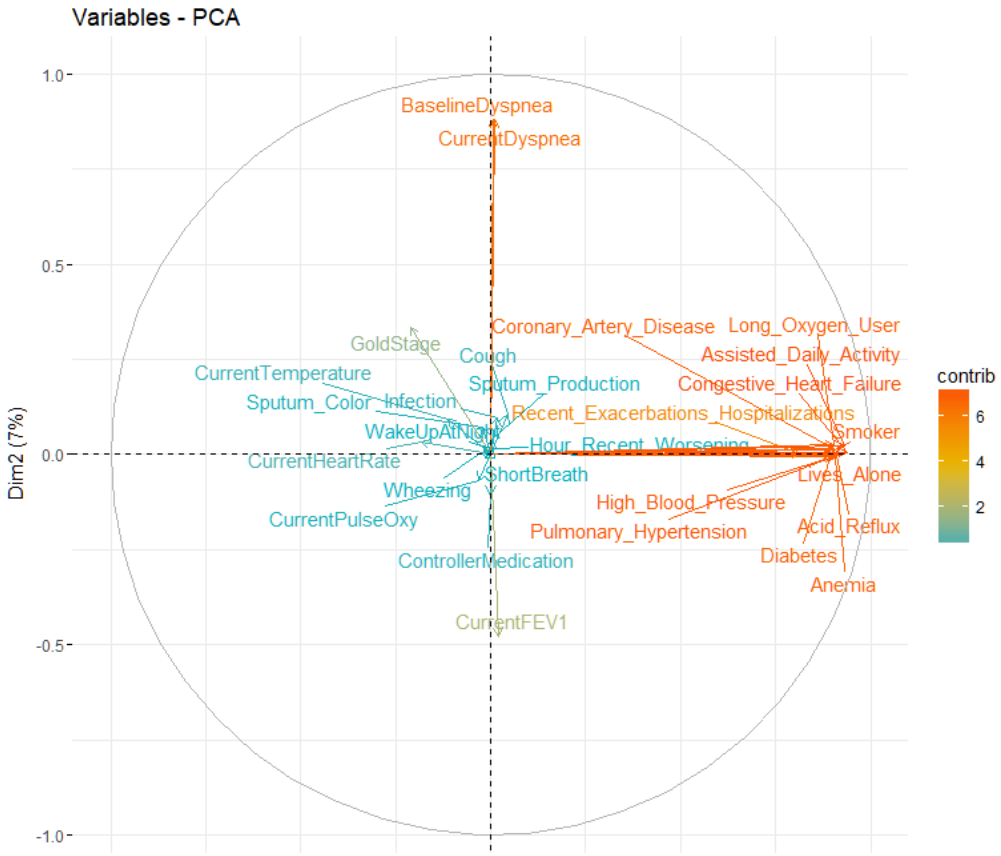
PCA is a type of linear transformation on a given data set that has values for a certain number of variables (coordinates) for a certain amount of spaces. This linear transformation fits this dataset to a new coordinate system in such a way that the most significant variance is found on the first coordinate, and each subsequent coordinate is orthogonal to the last and has a lesser variance. In this way, you transform a set of x correlated variables over y samples to a set of p uncorrelated principal components over the same samples

The Primary Component Analysis is conducted to get a picture of the variable correlations. The PCA results show the following predictors have "multicollinearity." However, **according to VIF scores, there are no more than 9.**

```
pca.res <- PCA(copd.pca, scale.unit = TRUE, graph = FALSE)
get_eigenvalue(pca.res)
fviz_eig(pca.res, addlabels = TRUE, ylim = c(0, 20))

var <- get_pca_var(pca.res)
var.dim1 <- var$contrib[order(var$contrib[,1], decreasing = TRUE),]
var.dim1[, 1:5]

fviz_pca_var(pca.res,col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel =
TRUE)

corrplot(cor(copd.pca), type="upper", order = "FPC" )

highlyCorrelated <- findCorrelation(cor(copd.pca), cutoff=0.4)
highlyCorCol <- colnames(copd.pca)[highlyCorrelated]
highlyCorCol
```

Independent Features	
Congestive Heart Failure	Diabetes
Pulmonary Hypertension	Long Oxygen User
Smoker	Coronary Artery Disease
Current Temperature	Lives Alone
High Blood Pressure	Anemia
Assisted Daily Activity	Acid Reflux



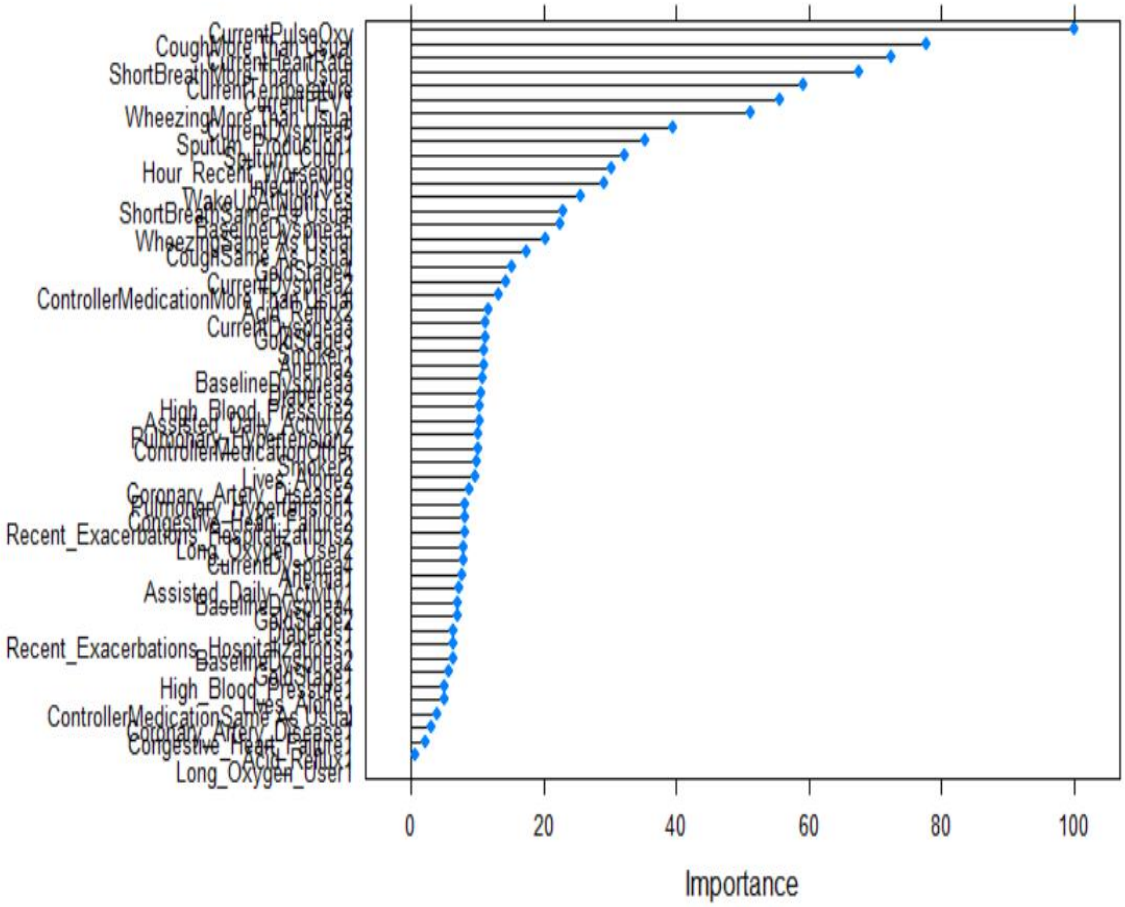


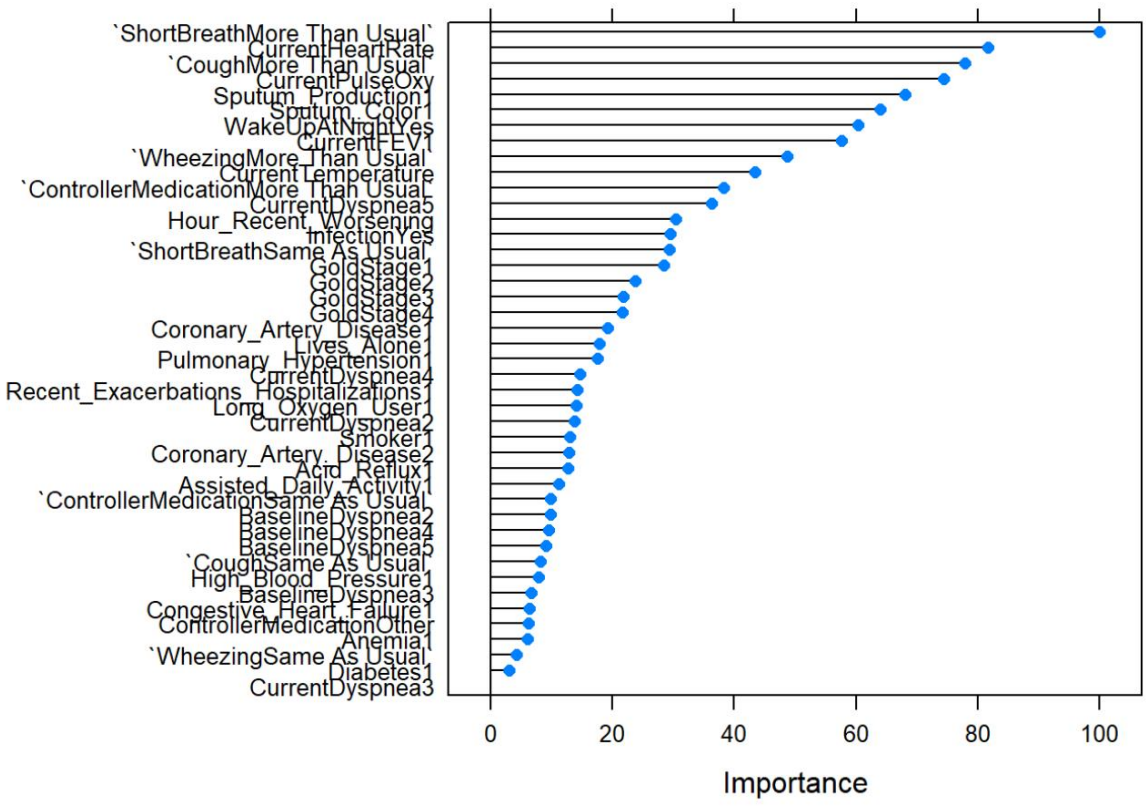
The correlation plot shows that the correlations between predictors were insignificant (correlation < 0.4 and Variance Inflation Factor < 3). Hence it is concluded that dimension reduction is not required.

C. To further check the correlated predictors, the importance of the predictors were ranked by fitting them into the general regression model and random forest model. The rankings are summarized in the following table.

	General Regression	Random Forest
Importance ranking from high to low	Short Breath	Current Pulse Oxy
	Current Heart Rate	Cough
	Cough	Current Heart Rate
	Current Pulse Oxy	Short Breath
	Sputum	Current

	Production	Temperature
	Sputum Color	Wheezing
	Sleeplessness	Current FEV1
	Current FEV1	Current Dyspnea
	Wheezing	Sputum Color
	Current Temperature	Sputum Production





Events/Non-events/Not Events

Since we are primarily interested in factors that lead to appropriate COPD triage assessment,

Event - An event is defined as "valid" if the patient's outpatient encounters were diagnosed as COPD or if the patient's inpatient encounters were diagnosed as COPD exacerbation.

Non-Event An event is considered as Non Event if the patient's outpatient were not diagnosed as COPD or if the patient's inpatient encounters were not diagnosed as COPD exacerbation.

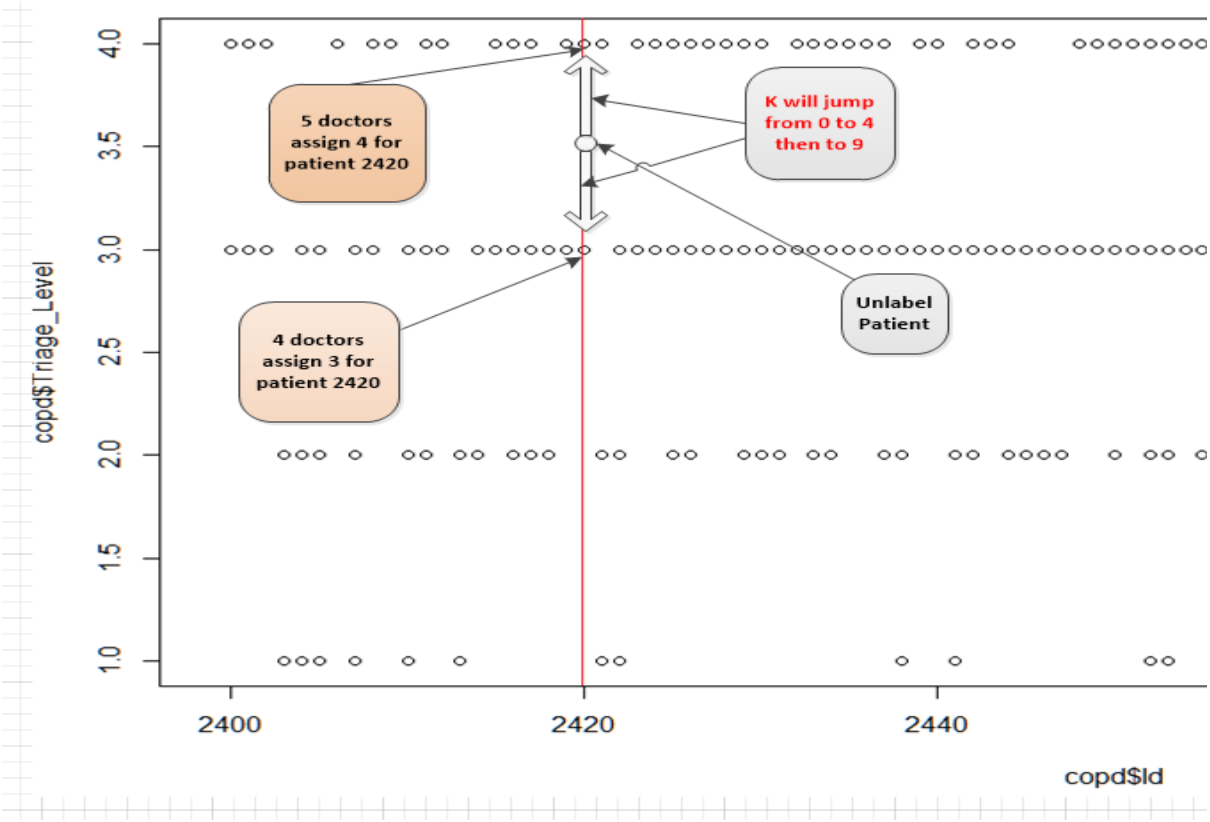
Not-Event Any other event outside of this database is considered not-event.

8. Model – Creation and Validation

8.1 KNN Model and Predictions

This algorithm is a supervised learning algorithm, where the destination is known, but the path to the destination is not. Understanding nearest neighbors forms the quintessence of machine learning.
KNN, K nearest neighborhood, is a supervised algorithm that classifies an unlabeled sample based on the distances to the labeled samples.
Three features sets were prepared to compare the performances of COPD triage classification.
A. Complete list – to compare KNN to other algorithm
B. Top 16 important features – to understand the performance impacted by importance of features
C. 4 most important categorical features plus 8 most important continuous features – to understand the performance impacted by continuous features

The COPD exacerbation triage levels were encoded to 4 levels: **"Ok," "Plan," "Doc," and "ER."** The features were scaled. The repeated CV was used to perform the train process.
One problem was found for implementing KNN to this case. Since each patient has several labeled triages based on the diagnosis from different doctors, it is hard to find the optimal K. To improve the model, the numerical triage was introduced to the model as a feature. This additional dimension will transform the same patient into different cases depending on the numerical triage.



8.1.1 Summaries for different KNN models Model

Accuracy	Complete list (30 features)	Top 16 important features	4 most important categorical features + 8 most important continuous features	4 most important categorical features + 8 most important continuous features + numerical triage value
Accuracy	41	50	48	62
Accuracy OK	48	86	87	100
Accuracy Plan	57	62	63	73
Accuracy Doctor	49	58	50	64
Accuracy ER	65	70	70	78
K	21	21	23	23

KNN works well for numerical features since the numerical feature sets will make each sample a unique case. Removing less important features from feature sets will improve the performance as well.

8.2 Naïve Bayes and Random Forest Model and Predictions

Naïve Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Although there are correlated features in the final dataset, the result from linear regression shows better performance compared to the dataset without correlated features in the dataset. To compare how sensitive the Naïve Bayes classifier is to the independence among predictors, the following models were prepared:

A. The final dataset – the dataset with the 30 features

B. The "positive" dataset – the dataset without the correlated features

- ShortBreath
- Cough
- Hour_Recent_Worsening
- Wheezing
- Sputum_Color
- Sputum_Production
- CurrentDyspnea
- CurrentPulseOxy,
- CurrentFEV1
- CurrentHeartRate
- CurrentTemperature
- Triage_Level

Then the Random Forest algorithm was used to fit dataset A. The main reasons for this process are:

1. For applications in classification problems, the Random Forest algorithm will avoid the overfitting problem

2. The Random Forest algorithm can be used for identifying the most important features from the training dataset. This will ensure the "positive" dataset contains all the important features.

8.2.1 Summaries for different Naïve Bayes and Random Forest models

	Model			
	(A) 30 features Naïve Bayes	(B) 12 uncorrelated, important features Naïve Bayes	(A) 30 features Random Forest	(B) 12 uncorrelated, important features Random Forest
Accuracy	55%	60%	69%	68%
Accuracy Ok	50%	96%	98%	82%
Accuracy Plan	57%	68%	80%	87%
Accuracy Doctor	62%	67%	68%	72%
Accuracy ER	67%	78%	83%	70%

When fitting dataset "A" with Naïve Bayes, it was found that the performance was impacted by correlated features. The result can be confirmed by dataset "B," which had better performance than dataset A. On the other hand, the Random Forest algorithm was not shown to be sensitive to correlated features. The performance for dataset "A" was slightly better.

8.3 SVM Predictions

SVM, support vector machine, is a supervised algorithm to classify samples based upon the idea of maximizing the margin i.e. maximizing the minimum distance from the separating hyperplane to the nearest example. Since the SVM algorithm is highly configurable, we can't tell which SVM model is better till the different parameters are applied. The following models were prepared to compare the performance of COPD triage classification.

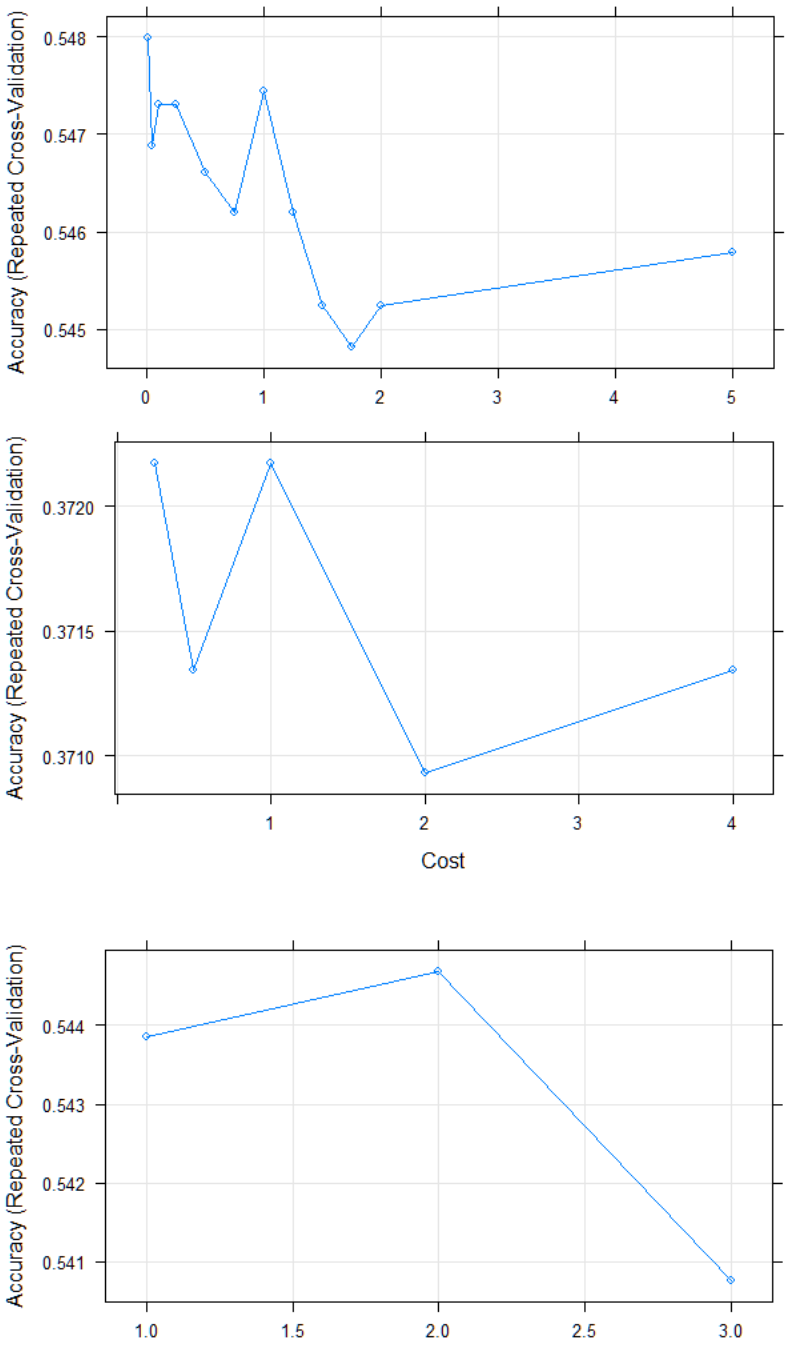
A. All features plus linear Kernel – Find the best margin based on linear kernel

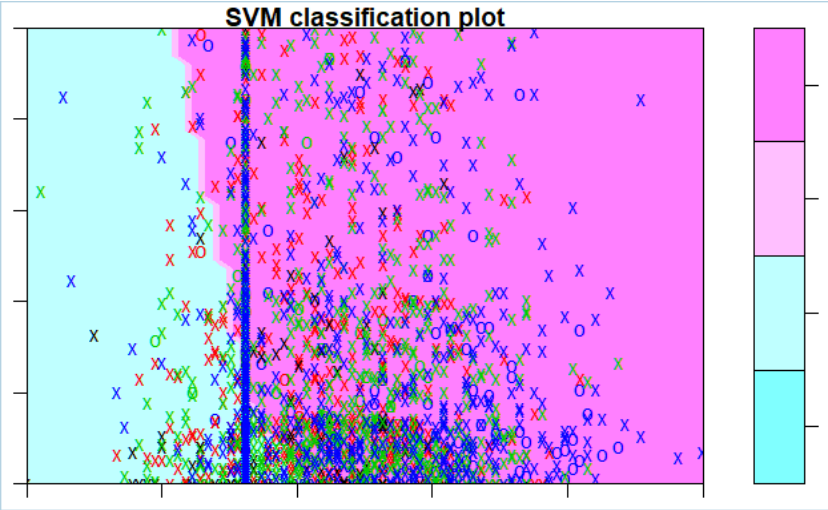
B. All features plus radial Kernel – Find the performance based on the assumption of non-linear dataset.

C. All features plus polynomial Kernel – to test the model by different degree of polynomial: 1, 2, or 3.

D. All features plus linear Kernel based on e1701 – to find the number of support vectors

E. 16 features plus linear Kernel based on e1701 – to see the impact of feature importance.





8.3.1 Summaries for different SVM models

	Model				
	30 Features Linear Kernel	**30 features Radial Kernel**	**30 features polynomial kernel**	**30 features linear kernel e1071**	**16 features linear kernel e1071**
Accuracy	75%	41%	73%	72%	67%
Accuracy Ok	66%	50%	65%	66%	50%
Accuracy Plan	80%	50%	77%	82%	84%
Accuracy Doc	79%	51%	77%	75%	70%
Accuracy ER	84%	51%	88%	84%	78%
Configurations	Cost: 0.25	Cost: 1 Sigma: 0.011	Cost: 0.25 Degree : 2 Scale : 0.01	Gamma: 0.018 Cost : 0.25	Gamma: 0.055 Cost : 0.25
Number of Support Vectors				Ok : 615 Plan : 196 Doc : 826 ER : 521	OK : 634 Plan : 208 Doc : 838 ER : 530

For this case, the linear kernel has better performance. The range of cost values from 0.05 to 5 was used to find the best C value. For the liner kernel, the best C is 0.25. Then, the radial kernel was used to check if the data set if non-linear. The poor performance (41% accuracy) was found from the results. Final, the polynomial kernel with degree of 1, 2, and 3 was used to verify the previously results. The result showed the best performance is based on 2 degree. To find out the number of support vectors, the "e1701" library used. The constant cost (0.25) and linear kernel was entered to compare two models: all features and the top most import features. The results showed that the model with less importance features removed did not produce the better performance. The number of support vectors also increased.

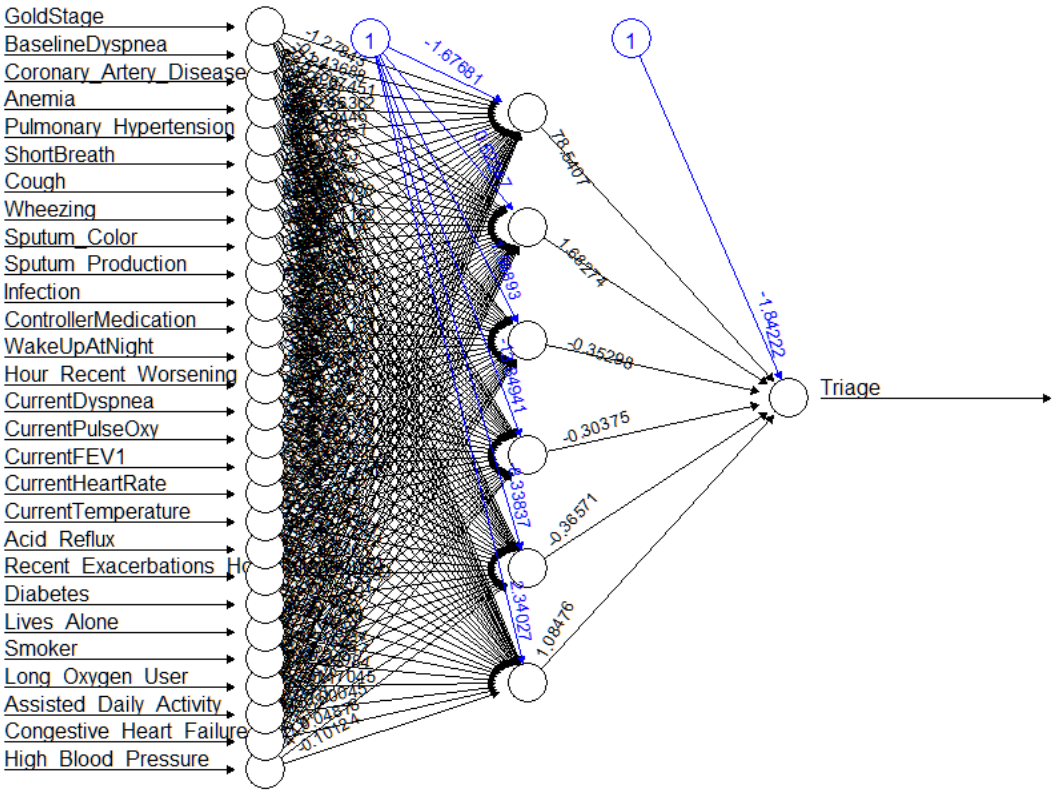
8.4 Artificial Neural Network Model and Predictions

A simple predictive algorithm tries to mimic the relationship between the input and the output variables. Such relationships can easily be predicted using simple regression algorithms. But it becomes difficult to make predictions in case of complex non-linear relationships and significant covariate terms. In such cases, we need more sophisticated machine learning tools. The artificial neural network (ANN) is the one of the tool that breaks the problem into multiple steps and solves for each step.
The ANN is generally used in cases where the past is repeated in almost exactly the same way. ANN will be able to memorize every single call. Hence, it is a kind of machine learning technique that has enormous memory. Moreover, for each observation, ANN does multiple re-calibrations for each linkage weight. Hence, the time taken by the algorithm rises much faster than other traditional algorithm for the same increase in data volume

There are two reasons to select the ANN as one of algorithms for this project:
1. Most of the time, decisions of clinical diagnoses follow a certain guideline. Let's assume the past experience can be applied to future patients.

2. It is able to do multi class classifications and able to calculate the possibility for each class.

The linear model was tested with different combinations of layers and hidden nodes. The model which has the best performance (the smallest **R**oot**M**ean**S**quare**E**rror) will be used as the base model for the classification model.

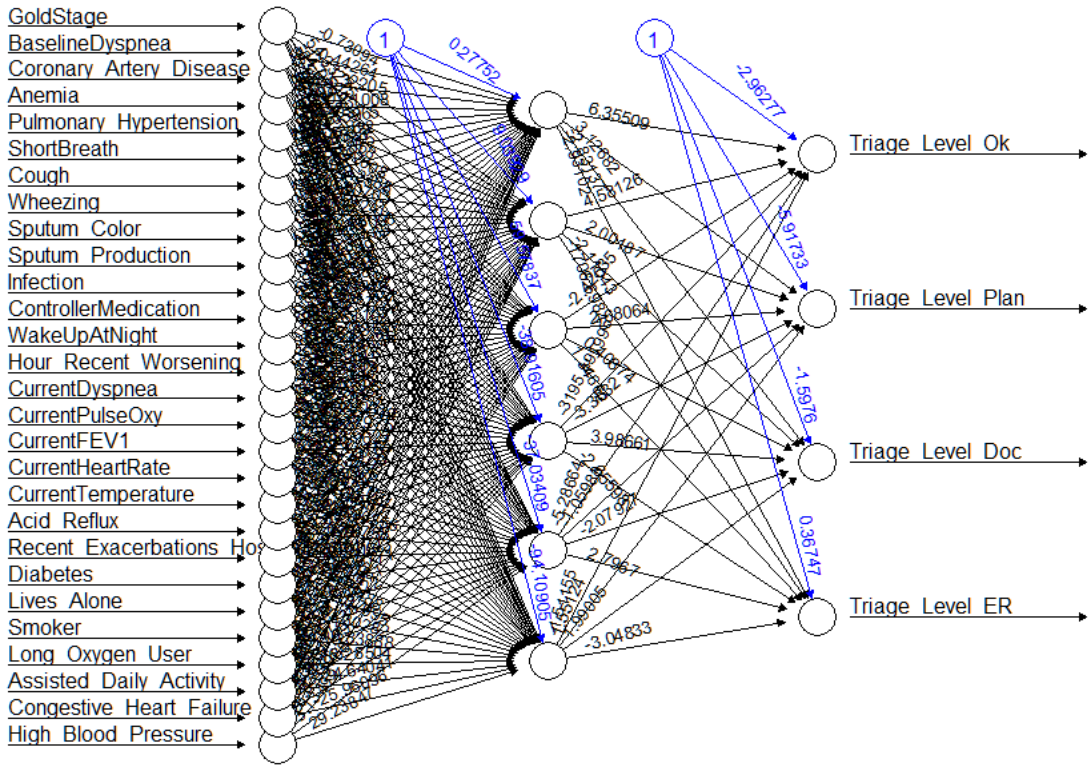


8.4.1 Summaries for different ANN models

Layer 1 Number the Hidden nodes	Layer 2 Number the Hidden nodes	RMSE
0	0	0.17
1	0	0.169
2	0	0.161
3	0	0.134

4	0	0.187
5	0	0.158
6	0	0.134
7	0	0.169
8	0	0.184
9	0	0.2
6	3	0.22

	neuralnet	nnet
	multi class classification	multi class classification
Accuracy	64%	63%
Accuracy Ok	63%	63%
Accuracy Plan	74%	74%
Accuracy Doctor	70%	68%
Accuracy ER	82%	79%



Unlike other machine learning algorithms, the ANN model classified lots of "OK" labels. Those misclassifies will put patients in risky situations. It's not sure if this is a hyper parameter tuning issue or if it needs to set up a threshold.

8.5 Extreme Gradient Boosting (xgboost) Model and Predictions

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. The following models were prepared to compare the performances of the COPD triage classifications.

A. All features with default caret configuration – Find the best hyper parameters

B. All features with configuration – Find the performance based on the different parameters.
C. The 16 most important features of configuration – to see the impact of feature importance

8.5.1 Summaries for different XGBoost models

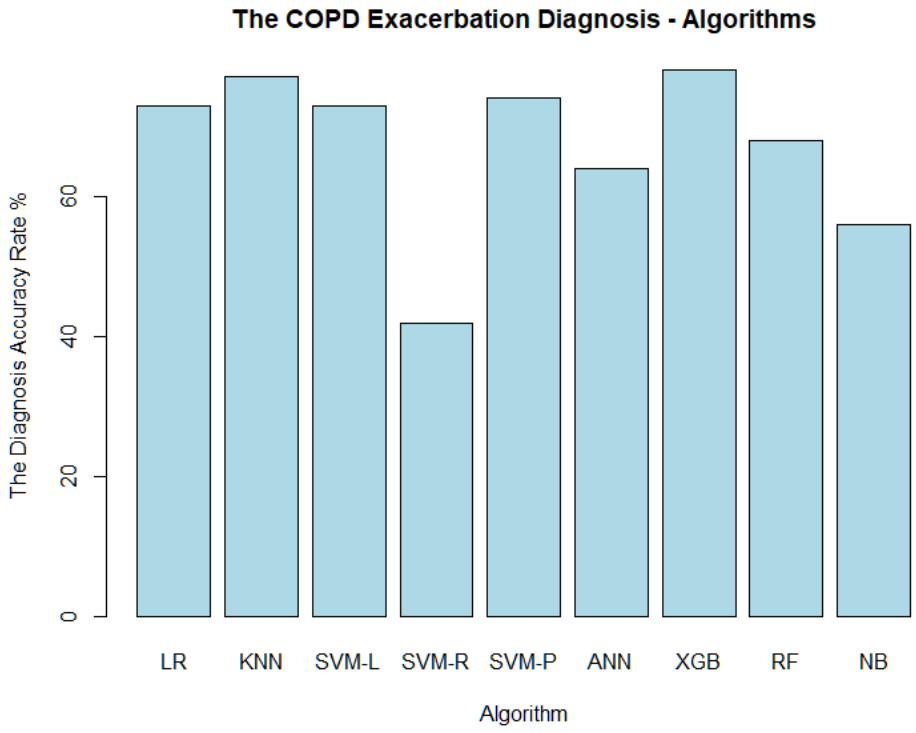
	Model		
	A: All features with grid search	B: All features with the following parameters	C: Top 16 features with the following parameters.
ETA	(0.3,0.4)	0.3	0.3
Max Depth	(1,2,3)	2	2
Gamma	0	0	0
Col Sample by tree	(0.6. 0.8)	0.6	0.6
Min child weight	1	1	1
Rounds	(100, 150)	50	100
Sub sample	(0.5, 0.6, 1.0)	0.5	0.5
Accuracy	78%	78%	68%
Accuracy OK	82%	82%	82%
Accuracy Plan	86%	87%	79%
Accuracy Doctor	80%	79%	79%
Accuracy ER	81%	86%	81%

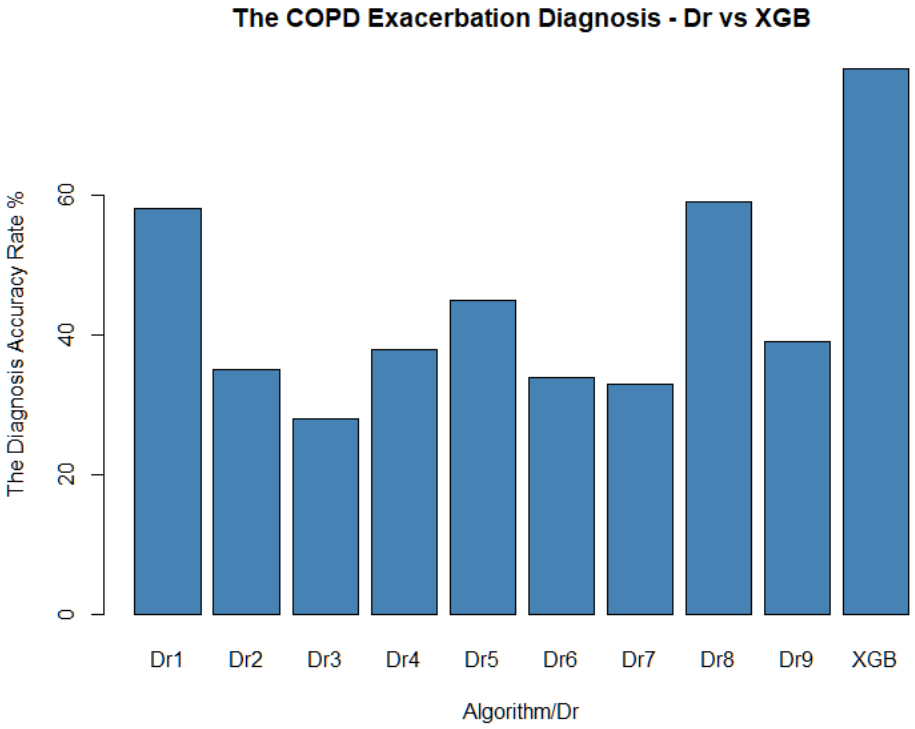
The Xgboost algorithm has shown far better results and has outperformed previous algorithms. The result showed that the feature importance did not impact the performance of the algorithm. The algorithm not only showed the highest accuracy of predictions but also performed well from safety analysis. For the patients who have a triage of ER, the algorithm showed a triage that was 10% under. The overall under triaged rate is less than 10%.

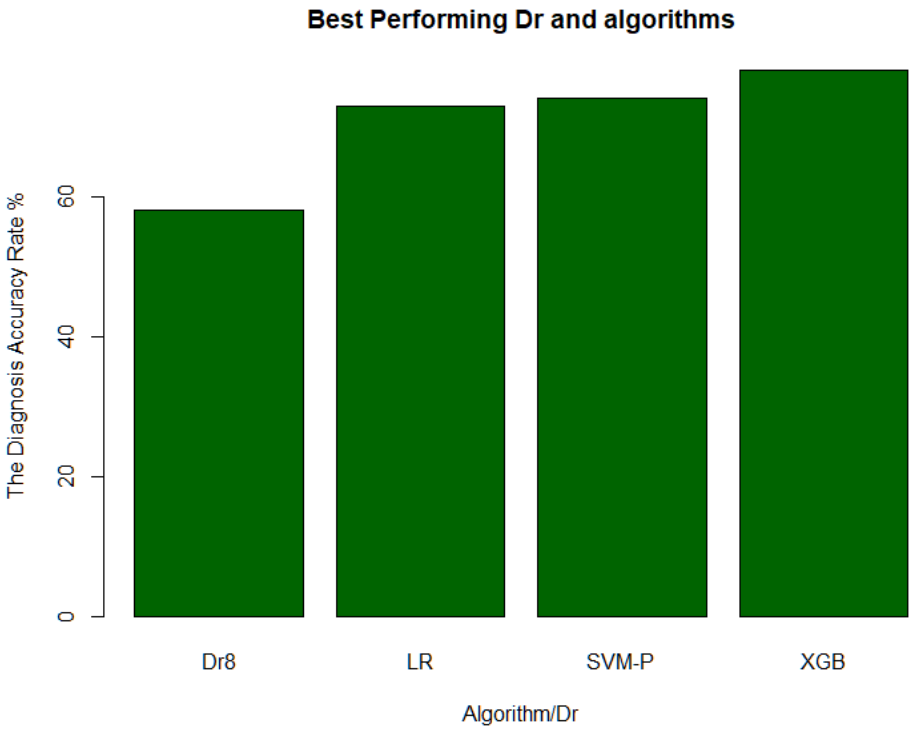
9. Summary of analysis result and explanation

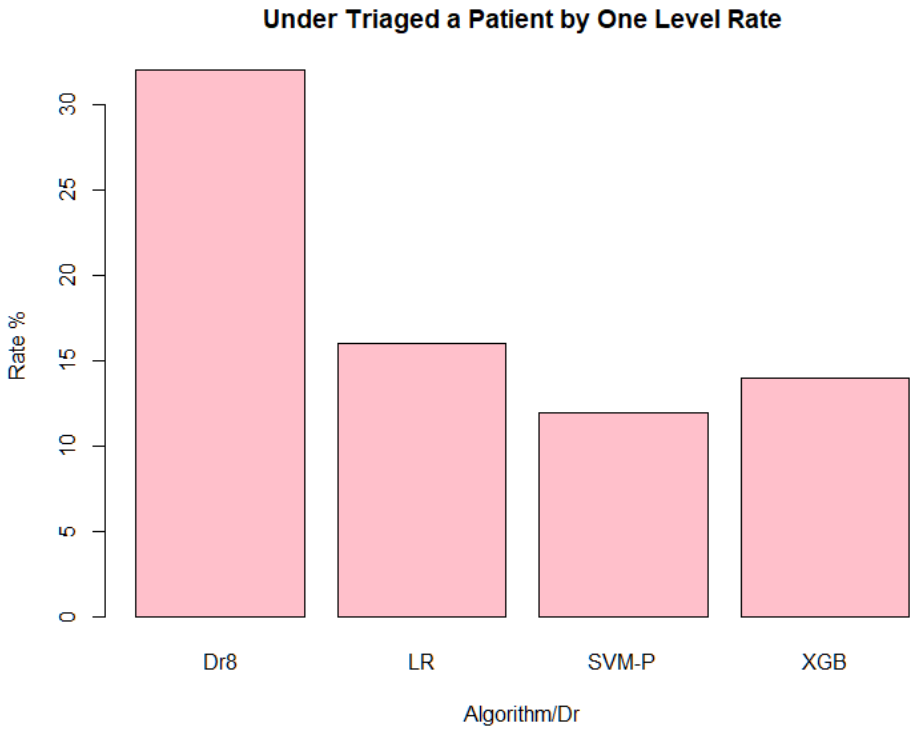
Based on performance and safety considerations, the two-step elimination approach was applied to find the best-fitting algorithm. The confusion matrix showed that accuracies of the support vector machine (with radial kernel) and Naïve Bayes models were 42% and 56% respectively. Since both models have a lower performance than Doctor 8 (the top performing physician), they were dropped from the final list. Although the KNN model had a good performance (77% accuracy), it's not a good model for this study since the patients have more than one label. Because the different cases were diagnosed by different doctors, it is hard to find the optimal K. To differentiate between the cases for a patient, I have to add the numerical triage to the predictor list. Unfortunately the numerical triage has a strong correlation with the label. Therefore, the model was dropped from the final list.
When considering safety, the confusion matrix showed that the ANN, the support vector machine (with linear kernel), and the Random Forest model under or over triaged a patient by more than one level. Specially, they under triaged a patient who should be sent to the doctor and ER. Furthermore, the black-box nature of the decision process is difficult to predict the results. They were removed from the final list as well.
Based on the approach above, the following algorithms were selected to fit this study:
1. Extreme Gradient Boosting (xgboost) – 78% accuracy with 14% under triaged a patient by one level.
2. Multiple-classification linear Regression (LR) – 73% accuracy with 16% under triaged a patient by one level.
3. Support vector machine with Polynomial kernel (SVM-P) – 74% accuracy with 12% under triaged a patient by one level.

Performance from Algorithm SVM-P 74				
Prediction	Reference			
	OK	Plan	Doc	ER
Ok	1	3	0	0
Plan	2	15	3	0
Doc	0	6	32	6
ER	0	0	6	27




The COPD Exacerbation Diagnosis - Dr vs XGB
The Diagnosis Accuracy Rate %
0
20
40
60
Dr1
Dr2
Dr3
Dr4
Dr5
Dr6
Dr7
Dr8
Dr9
XGB
Algorithm/Dr


Best Performing Dr and algorithms
The Diagnosis Accuracy Rate %
0
20
40
60
Dr8
LR
SVM-P
XGB
Algorithm/Dr



10. Conclusion

In this study, the hypothesis test showed that the diagnosis of variance for triaging COPD exacerbation exists among physicians. Therefore, using machine learning algorithms to develop an automatic triage system might potentially increase the diagnosis accuracy. The system can be a useful tool for COPD patients to reduce the cost of care and increase their quality of life by self-managing their symptoms. Instead of doing Google searches or going to the ER directly, the tool will provide patients a quick action plan based on their profile, symptoms, and vital sign severity. The results showed the best classification accuracy is 78% while the top performing physician's accuracy was 60%.

11. Ethical Considerations

The data provided in the research paper was simulated so we were safe from any HIPAA compliance issue. In case of real data, the data needs to be scrubbed and the identity to be protected.

12. Opportunity for improvement and further investigation

In the extended analysis, the different classification algorithms were compared. The result demonstrated that the top three performing algorithms are XGB, SVM-P, and LR. Since this is a multi-class classification case and there are lots of categorical predictors, the top performing algorithm XGB Tree is selected as the best fit algorithm for this study. The most important benefit of XGB is that it has the best accuracy in triaging the cases, but other reasons include:

1. Decision tree can be visualized, which makes healthcare providers establish the guideline to review and explain the results.
2. Can capture high-order interactions between inputs.
3. Training time is relatively fast.

These algorithms allow COPD patients to monitor and manage their health condition by themselves through a low cost rapid platform. For future work, it can be tried to increase the accuracy of prediction by using deep learning methods.

13. References

1. A machine learning approach to triaging patients with chronic obstructive pulmonary disease
2. https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/copd/learn-about-copd/what-is-copd.html
3. https://stackoverflow.com/questions/2907896/how-to-assign-to-the-names-attribute-of-the-value-of-a-variable-in-r
4. http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/
5. Multicollinearity with R
6. Principal Component Analysis with R
7. https://dataaspirant.com/2017/01/09/knn-implementation-r-using-caret-package/
8. Support Vector Machine Tutorial Using R | SVM Algorithm Explained
9. https://stackoverflow.com/questions/50376411/neural-network-error-in-plot-nn-weights-were-not-calculated