Prediction of Patients' Length of Stay at Hospital During COVID-19 Pandemic

Jianing Pei, Qixuan Chen, Xin Lin

Abstract

The purpose of the research is to predict Covid-19 patient length of stay in hospital. The stay length is separated into 11 classes. The first 10 classes correspond to 0-10 days, 11-20 days...91-100 days respectively, and the last class is more than 100 days. Three Machine Learning models, namely, K-nearest Neighbors Algorithm, Logistic Regression and Decision Tree are implemented with python to make prediction with the dataset. In the part of preprocessing, PCA, correlation, normalization, one-hot encoding and simple encoding are used to make preparation for the implementation of the models. The way of encoding and hyperparameters of models are adjusted, but there is no obvious change of the accuracy. The final accuracy of each model are 0.3442, 0.3524 and 0.3541, which are not very high. Because of this, the features are are not very relevant to the length of class, in our opinion. With this result, we concluded that our dataset is not good, and the features should be improved to make a better prediction. Also, the classes are not practical, as the range of stay length is too large. More detailed data sources should be taken to make the prediction more useful for the healthcare management.

Keywords: Machine Learning; Prediction; Length of stay; K-nearest Neighbor; Support Vector Machine; Random Forest

1 Introduction

The world has been exposed to the Covid-19 epidemic since December 2019(Chi Zhang, 2020). Till now, the virus is still spread rapidly, and the better management of healthcare and treatments are required to deal with continuing crisis(Junaidi, 2020). Although there are many cases of using data technology in the health system(dela Cuesta, 2018), the length of hospital stay is an important indicator to monitor and predict the health management process. Thus, if the length stay of the patients can be predicted according to their features, the management of the medical resources can be improved greatly. Many work has been done to improve the management of healthcare previously, including some ways to reduce Health Disparities for Priority Populations("Improving Cultural Competence to Reduce Health Disparities for Priority Populations", 2014), the

Quality Improvement Initiative for Nursing Facilities(Jo A. Taylor, n.d.), etc. Besides the improving healthcare for these aspects and groups of people, the treatment for Covid-19 patients also should be improved. WIth the dataset of AV: Healthcare Analytics II(Prabhavalkar, n.d.), the length of stay can be well predicted, so that the medical resources can be managed in advance according to the prediction, and more patients are able to get treatment on time.

2 Method

2.1 KNN

KNN(K-nearest neighbor) was used to predict the length of stay in hospital of each person according to their different features. To make the dataset be well trained by KNN, some preprocessing methods were used. As the dataset contains categorical data, which cannot be recognized by KNN, one-hot encoding and simple encoding were considered to be used. After normalizing the data by min-max normalization: $y = \frac{x-min}{max-min}$, the function of the different encoding was compared with same set of hyperparameters. Since the accuracy of simple encoding was higher, it was chosen finally.

As the dataset, which contained more than 400,000 samples, was too large to be run out with KNN, 50,000 samples were chosen randomly and divided into the training set and test set, and the percentage of them were 90% and 10% respectively. There were three hyperparameters for this model: distance measure(Manhattan distance: $(|x_1 - x_2|)+(|y_1 - y_2|)$ and Euclidean Metric: $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$), different ways of processing depending on correlation and K(number of neighbors). Different combinations of hyperparameters were implemented to obtain the best model, which had the highest accuracy.

2.2 Support Vector Machine

SVM was used to analyze the dataset. The data contains two types of data, numerical and categorical. Multicollinearity, however, occurs when two or more independent variables in the dataset are correlated with each other, so to avoid it, I drop the first column

after one-hot encoding. As a result, the dimensions of the train and test are 116. According to the principal component analysis, we choose 68 features to analyze.

Since the original training set included more than 100,000 samples, according to the official document, first I used linear SVM instead of kernel SVM.Secondly, I still want to use the kernel function, since the dataset is nonlinear. However, the original training set samples was too large to analyze. I change the sample size for training set and test set. There are 40000 samples in new training set and 10000 samples in new test set. I can't do a good job of analyzing 68 features based on 50,000 data. So I change to use simple label encoding, and then I found 11 features to consider. There are three different kernel function:Gaussian radial basis function,Polynomial function and Sigmoid function. And when we use different kernel function, we need to find the optimal value of C and gamma for each function.("Support Vector Machine", n.d.) Where parameter C controls how much you want to punish your model for each misclassified point for a given curve, and gamma defines how far the influence of a single training example reaches.

2.3 Random Forest Classifier

Random Forest Classifier is an ensemble learning method used to classify samples from different classes (Liaw, Wiener, et al., 2002). In the model, a number of decision trees are constructed and the majority class predicted by these decision trees is considered to be the classification result. In each tree node, a random subset of features are selected and among which, the best predictor is chosen according to a loss function, such as Gini impurity.

To implement this model, the healthcare dataset is first divided into training set(80%), validation set(10%), and test set(10%). During data preprocessing, different encoding methods, either one-hot encoding or simple encoding, were used to encode the categorical features in each sample. Principle component analysis(PCA) was then applied to reduce the dimentionality of the feature space and get rid of irrelevant features. In PCA, only the top n components were selected to explain 85% variance in the feature space. Therefore, the

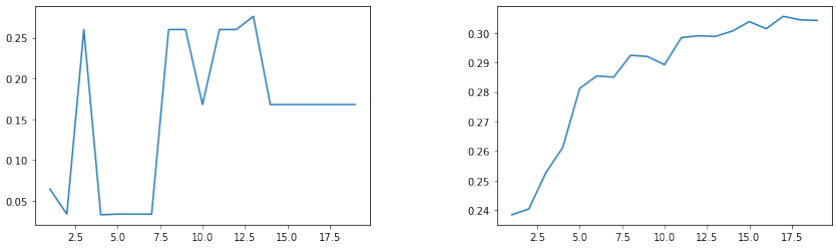
four preprocessing regimes are one-hot encoding, one-hot encoding with PCA, simple encoding, simple encoding with PCA. These four preprocessing methods were evaluated based on their corresponding random forest classification accuracy, and the one with highest classification accuracy on the validation set was chosen to be the final preprocessing regime.

Four hyperparameters in the random forest classifier are then tuned using the training set and validation set. The number of estimators is the number of decision trees in the random forest, where each tree gives a classification result and votes for the most likely class. The maximum depth of each tree is another parameter that controls the model complexity and bias-variance tradeoff. Two rest two parameters are the minimum number of samples at each leaf node and the minimum samples required to split a node. For each of the parameters, a wide range of values were experimented and the best values were selected based on model accuracy.

3 Results and Discussion

3.1 KNN

Different distance measures with same dataset: 50,000, $K \in (1, 20)$, 11 features.
Manhattan distance and Euclidean Metric figure 1



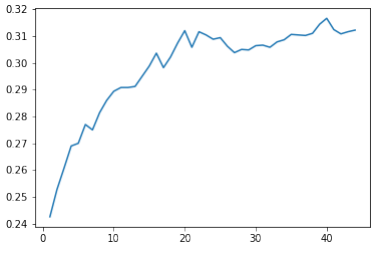
(a) Manhattan distance k=13 accuracy(max)=0.2762

(b) Euclidean Metric k=17 accuracy(max)=0.3056

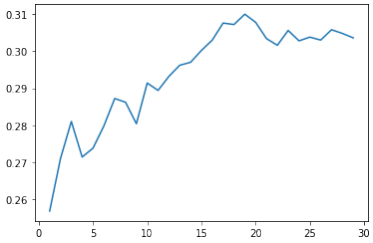
Figure (1) Hyperparameter change of distance measure

Because the accuracy of Euclidean Metric was higher than the Manhattan distance obviously, it was assumed that the models which

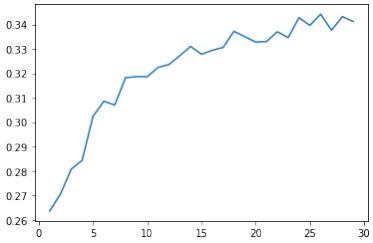
used Euclidean Metric were higher than those used Manhattan Distance.
With fixed simple encoding and Euclidean Metric, the features were adjusted according to their correlation figure 2



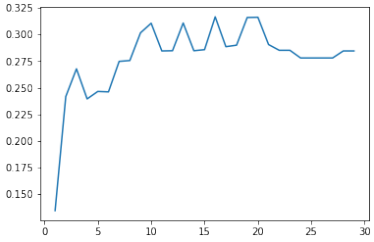
(a) 17 features k=40 accuracy(max)=0.3166



(b) 12 features k=19 accuracy(max)=0.31



(c) 6 features k=26 accuracy(max)=0.3442



(d) 1 features k=16 accuracy(max)=0.3166

Figure (2) Hyperparameter change of features

The final model uses simple encoding, Euclidean Metric, 6 features with correlation¿0.05, and had its highest accuracy 0.3442 at K=26.

3.2 Support Vector Machine

Firstly I used Linear SVM. As the dataset has 11 classes, we use one-versus-one method. Finally, the accuracy is 0.35003. Then I tried to use different kernel function. For Gaussian radial basis function (rbf):According to the Figure 3, C=1 and gamma=scale. Then the accuracy is 0.352. For,Polynomial function: According to the Figure 4, C=0.1 and gamma=scale. Then the accuracy is 0.3516. For Sigmoid function: According to the Figure 5, C=0.001 and gamma=0.001. Then the accuracy is 0.3509.

Gaussian radial basis function is the best,according to Figure 6, since the accuracy is largest.

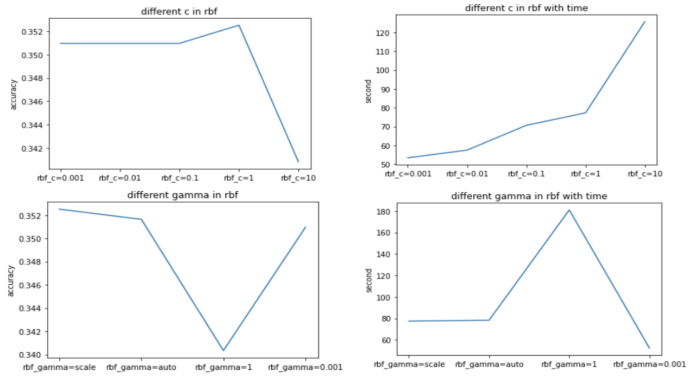


Figure (3)　An image of C and γ in rbf

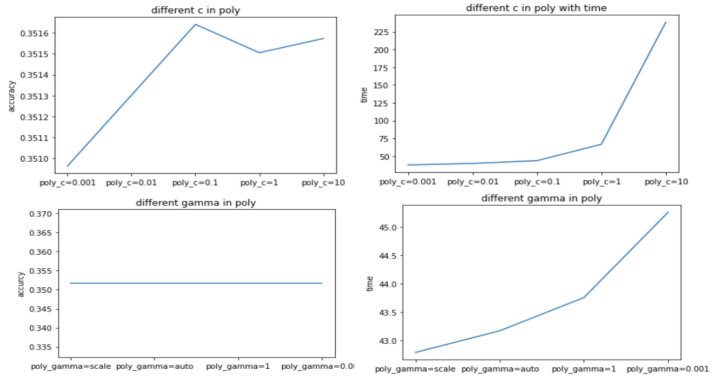


Figure (4)　An image of C and γ in ploy

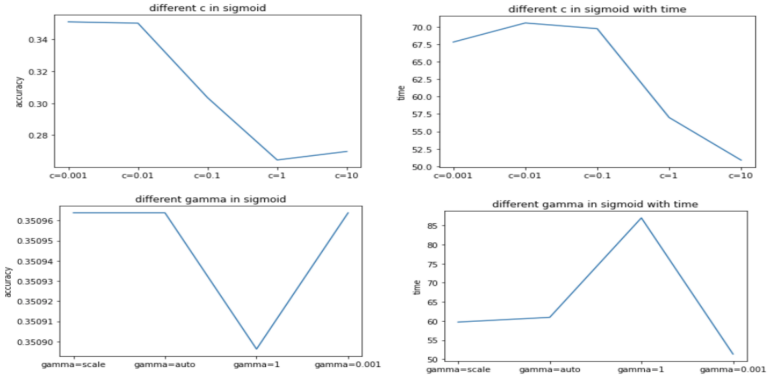


Figure (5)　An image of C and γ in sigmoid

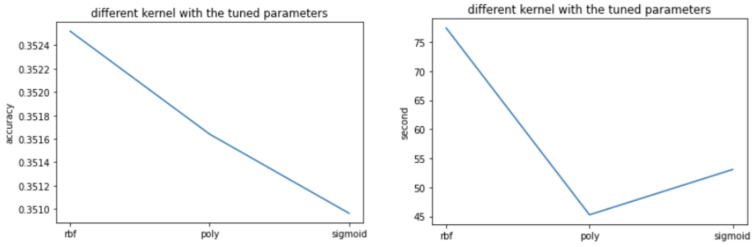


Figure (6) An image to make kernel comparison

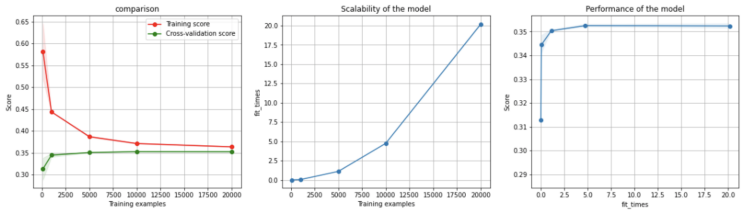


Figure (7) An image to check over-fitting

And then we need to check weather it is over-fitting.According to the Figure 7,the model does not overfit because as the sample size of data increase, the performance of the model doesn't change. Overall, the best model is use Gaussian radial basis function with c=1 and gamma=scale. And the final accuracy is 0.3525

3.3 Random Forest

The classification accuracies of random forest classifier using different preprocessing techniques are very similar. As shown in Figure. 8, while the classification accuracy for each preprocessing method increased with the number of decision trees, there was less than 2% accuracy difference across encoding methods. Simple encoder with PCA was chosen to be the preprocessing regime because it yields the highest classification accuracy even when the number of estimators is low.

During hyperparamter tuning, the model is found to be insensitive to different parameter values. Figure. 9 shows that the amount of variation in classification accuracy was lower than 3% for all four hyperparameters tuned. In the final model, number of estimator = 100, max depth = 10, min sample split = 20, and minimum samples

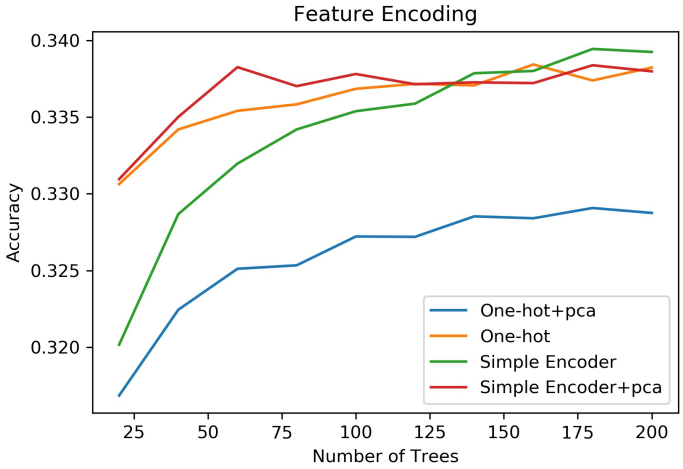


Figure (8) Comparison of different preprocessing regimes.

leaf = 20. The test accuracy is 0.354 and the train accuracy was 0.356. Therefore, there is no overfitting observed.

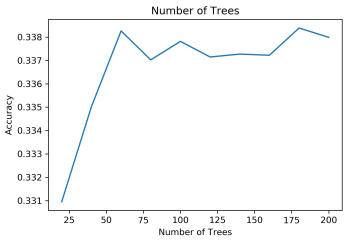
3.4 Discussion

According to Table 1, we find that the accuracy of the three models is about 0.35.
When we tuned the hyperparameters, we found that the adjustment of the hyperparameters could not improve the accuracy very well. At best we can only improve accuracy by four percent. Also we tried different way to encode, one-hot encoding and simple label encoding, there was no significant change in accuracy.

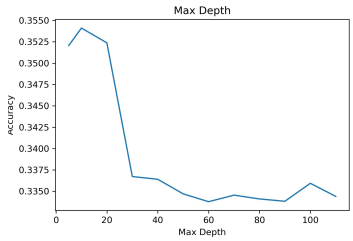
Model	KNN	SVM	Random Forest
Accuracy	0.3442	0.3525	0.3541

Table (1) Model accuracy

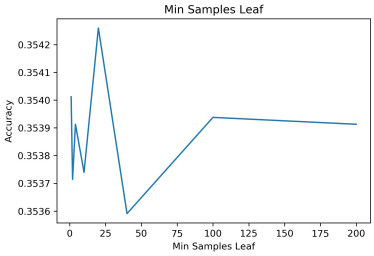
We think that the accuracy of this dataset can only reach about 0.35, and we cannot significantly improve the accuracy. We think this is due to some features of the dataset.
According to the Figure 10, we can find only one factor that has a strong correlation with stay and other factors have very little relevance to it. Also we know this dataset is unbalanced, based on the



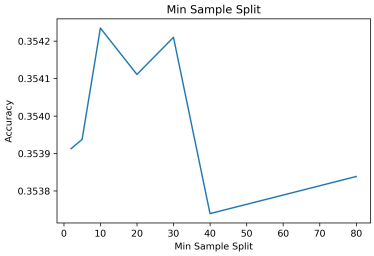
(a) Number of estimators



(b) Maximum depth



(c) Minimum samples at each leaf node



(d) Minimum samples to split

Figure (9) Hyperparameter tuning.

Class	0	1	2	3	4
Size	160661	78139	87491	55159	11743

5	6	7	8	9	10
35018	2744	10254	4838	2765	6683

Table (2) Class size

table 2. Class 0 has a huge amount of data, while class 6 and class 9 contain very few samples. From what has been discussed above, we believe that the factors contained in this dataset are not the main factors. And the classification of stay is not reasonable, we should not divide stay into 11 classes.

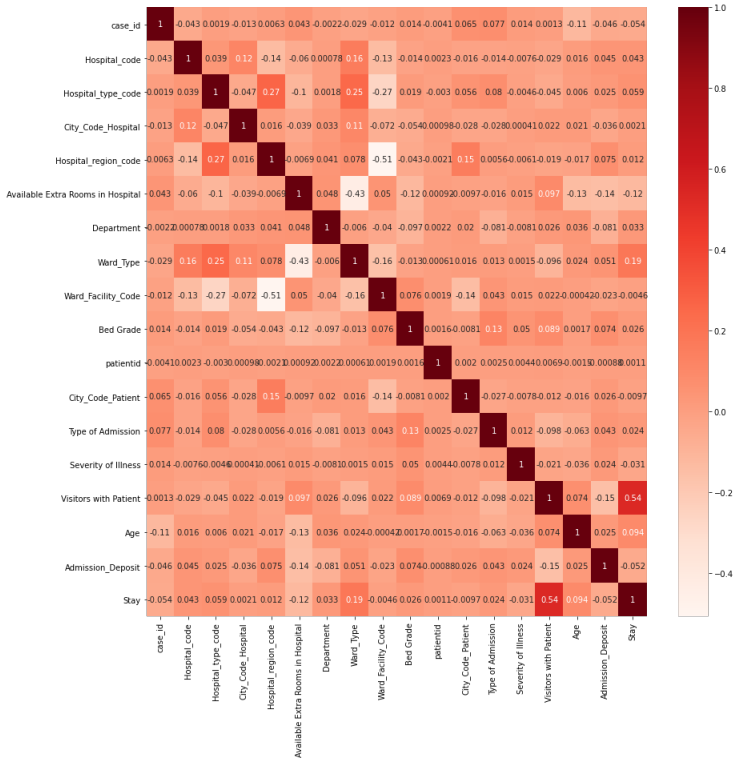


Figure (10) correlation matrix

4 Conclusion

In this project, machine learning is utilized to solve real-world medical resource allocation problem during COVID-19 pandemic. By predicting each patient's estimated stay at hospital as a small 10-day interval, doctors and governors could make plans about medical equipment and human resources accordingly.

However, the highest classification accuracy of 35.41% given by our models is a bit too low to be useful as a diagnosis tool in practice. We propose several ways optimize the prediction model in the future. Firstly, since the classification classes are highly imbalanced, with most of the patients spending less than 40 days in hospital, the classification labels could be reformatted into fewer classes concentrating on patients' stays less than 40 days. Secondly, the feature space in the raw dataset could be further expanded to include more relevant patient features such as body temperature. Extra features could potentially be better predictors for patients' length of stay, since Coronavirus infection symptoms are still understudied due to the novelty of the disease.

5 Acknowledgement

We thank Dr. Victor Adamchik for giving insightful guidance on our project. We also acknowledge Siyi Wu for generous help and useful discussion.

6 Contribution

Jianing Pei: SVM model implementation, Report(SVM parts and discussion), slides
Qixuan Chen: KNN model implementation, Report(KNN parts and introduction, abstract), slides
Xin (Jack) Lin: Randorm forest implementation, Report(Random Forest parts, conclusion, acknowledgement), slides

References

Chi Zhang, J. L. K. T. W. Y. H. Z.-G. W., Zhao Wu. (2020). Discharge may not be the end of treatment: pay attention to pulmonary fibrosis caused by severe covid-19. *DAWN*. Retrieved from https://www.dawn.com/news/1585323

dela Cuesta, J. B. (2018). Research proposal form example: Health research. *NursingAnswers*. Retrieved from https://nursinganswers.net/essays/research-proposal-form-example-health-5513.php

Improving cultural competence to reduce health disparities for priority populations. (2014). *AHRQ*. Retrieved from https://effectivehealthcare.ahrq.gov/products/cultural-competence/research-protocol/

Jo A. Taylor, M. P. P. P. H. B. M. A.-B. J. O. M., R.N. (n.d.). The falls management program: A quality improvement initiative for nursing facilities. *AHRQ*. Retrieved from https://www.ahrq.gov/patient-safety/settings/long-term-care/resource/injuries/fallspx.html

Junaidi, I. (2020). Minister warns of increase in covid-19 cases. *DAWN*. Retrieved from https://doi.org/10.1002/jmv.26634

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

Prabhavalkar, N. (n.d.). *Av : Healthcare analytics ii.* Retrieved from https://www.kaggle.com/nehaprabhavalkar/av-healthcare-analytics-ii

Support vector machine. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Support_vector_machine