ICU Admission Prediction For Patients Infected With Covid-19

Tianran Wei

Course: Data mining approaches in epidemiology

February 10, 2021

Abstract

Since the coronavirus outbreak in 2020, healthcare systems around the world have been suffering from extremely difficult situations as the epidemic has intensified. ICU, or intensive care unit, as a critical part of the healthcare system, its capacity has been reached at a critical level in many countries. Given this situation, countries are taking various measures to alleviate the ICU shortage, such as building new ICU units and urgently training new medical staff. Another way to alleviate the shortfalls in ICU resources is to explore how medical resources can be allocated more appropriately. In this paper, anonymized clinical data from Hospital Sírio-Libanês, São Paulo and Brasilia are used and a data mining approach is practiced on it. Multiple supervised machine learning models including logistic regression, decision tree, support vector machine, naive Bayes, and artificial neutral network are developed to predict admission to the ICU of confirmed COVID-19 cases. This will greatly reduce the workload of medical staff, assist in decision-making in earlier stage, and increase the accuracy of medical judgments. The result of the performance evaluation of the models showed that random forest model has the highest accuracy of 94.5% while the Multi-layer Perceptron classifier has the accuracy of 92.2%.

Contents

1 Introduction

1.1 Background

Since the novel coronavirus, SARS-CoV-2-causing Coronavirus Disease 19 (COVID-19), was declared a pandemic by the World Health Organization in March 2020, it is still spreading around the world, with more than 100 million confirmed cases and 2.2 million deaths across nearly 200 countries.

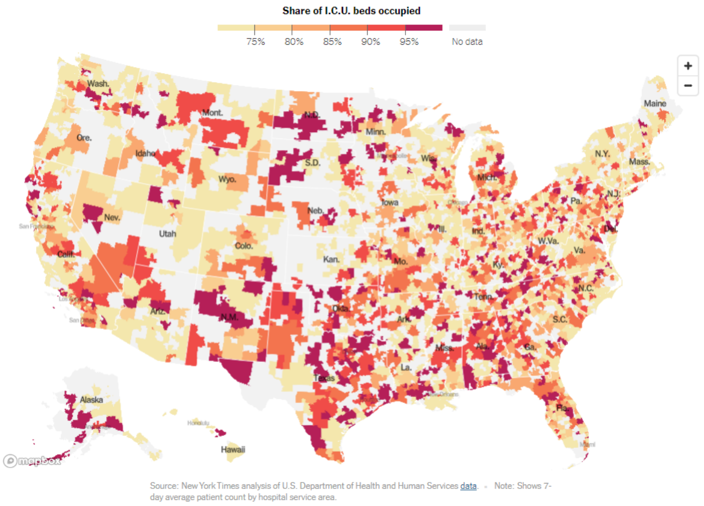
The most Significant symptom was fever. Less than half of the patients presented with respiratory systems including cough, sputum production, itchy or sore throat, shortness of breath, and chest congestion. 5.6% of patients had diarrhea [1]. At the same time, some patients are suffering from more severe symptoms. Studies from Wuhan, China reported a high incidence of acute respiratory distress syndrome (29%), RNAaemia (15%), acute cardiac injury (12%) and secondary infection (10%)[2]. Similar rates of critical illness (16%) were also reported in Lombardy, Italy [3]. These severely symptomatic infected patients go through a period of medical observation before being admitted to the ICU when their condition deteriorates.

In regions where medical resources are extremely scarce, we can observe a higher mortality rate than in other regions. Wuhan, China, was the site of the initial outbreak of the new coronavirus. In January 2020, the beginning of the outbreak, the local health system was in overload and its mortality rate was much (2%-3%) higher than the national average in China (1%) [4].

This figure is even more exaggerated in Italy. As of April 15, 2020, the case fatality rate in Lombardy reached 18,3%, according to the data given by the Civil Protection Department of the Italian Government. The mortality rate in the rest of Italy is 10.6%. Considering the various regional testing strategies and capacities in each region this figure may not be accurate. Mortality rates provide more reliable data and truly quantify. As of April 15, Lombardy had 113-1 death per 100 000 population, more than six times higher than in the rest of Italy[5]. Although the aging of the population and statistical methods affect the mortality figures of the Italian epidemic, the collapse of the national health system is still the main reason for the much higher than average mortality rate of newly crowned infected persons in areas with severe epidemics[6].

However, with the rapid increase in the number of infections, medical systems in various countries are not prepared for pandemics. And the ICU, as the last line of defense for infected patients, has been an extremely scarce medical resource around the world until now due to its high construction costs and long construction and medical staff training cycles.

The heatmap below[7] illustrates the ICU bed's capacity in the U.S. on Dec. 9, 2020. The data is self-reported to the U.S. Department of Health and Human Services by individual hospitals. As we can see, in more than a third of areas hospitals are running critically short of intensive care beds. Hospitals serving more than 100 million Americans reported having fewer than 15 percent of intensive care beds still available as of last week, according to a Times analysis of data reported by hospitals and released by the Department of Health and Human Services.



Source: New York Times analysis of U.S. Department of Health and Human Services data. • Note: Shows 7-day average patient count by hospital service area.

Figure 1: I.C.U. Beds Near Capacity Across U.S.

In Germany, at 12.12.2020, an intensive care doctor has warned that only five to 10% of intensive care beds are still available in the most parts of Germany. Meanwhile, 28,438 new cases have been reported over the past 24 hours[8]. In some countries with severe outbreaks, the shortage of beds has even led doctors to "screen" critically ill patients by "age and chance of survival".

Facing the crisis of medical resources and the resulting increase in mortality due to the new crown epidemic, this paper aims to use Data Mining Technic to allocate medical resources more accurately and rationally, and to use models to assist in the determination of ICU admissions in earlier stage to rationalize the allocation of ICU beds in a more efficient way and reduce the pressure on physicians' workload. This will alleviate the high mortality rate in pandemic settings where medical resources are scarce and buy more time for vaccination.

1.2 Dataset

The Dataset used in this paper contains anonymized data from Hospital Sírio-Libanês, São Paulo and Brasilia. All data were anonymized following the best international practices and recommendations and all of the patients in this dataset are COVID-19 virus infected. The dataset contains four types of patient information:

- Patient demographic information (03)
- atient previous grouped diseases (09)
- Blood results (36)

- Vital signs (06)

The number in parentheses represents how many features the information is represented by. In total there are 54 features, expanded when pertinent to the mean, median, max, min, diff and relative diff. Each patient's medical observation is divided into five stages, which is named as "window" in the dataset. At each stage the patient's information is recorded in one row, so that each patient occupies five rows in the dataset. For 385 patients, we have in total 1925 rows and 231 columns in the original dataset.

At any one window the patient may be admitted to the ICU due to deterioration of their condition. The target is to predict if the infected patient should be admitted to ICU in the next window, which means that our problem is defined as a supervised classification problem. Note that if a patient has been admitted to the ICU at a certain window, the data in this window and the windows after that will not be allowed to be used for modeling, because the patients after ICU admission stay already in ICU, which has no practical meaning for the target we are trying to predict.

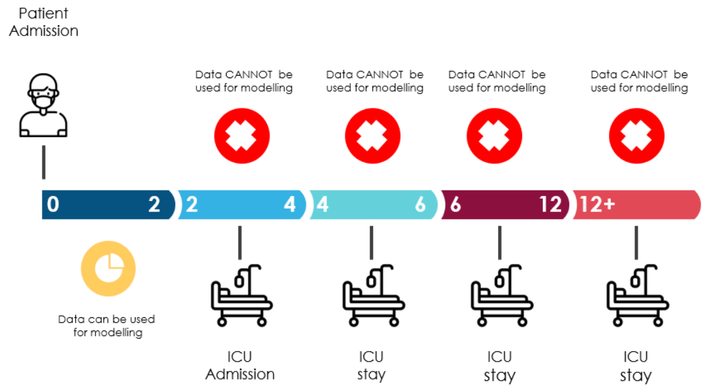

Patient Admission
Data CANNOT be used for modelling
Data CANNOT be used for modelling
Data CANNOT be used for modelling
Data CANNOT be used for modelling
0
2
2
4
4
6
6
12
12+
Data can be used for modelling
ICU Admission
ICU stay
ICU stay
ICU stay

Figure 2: ICU window concept

2 Methodology

The entire study will be guided by the Cross-industry standard process for data mining theory. The theory is also known as CRISP-DM, which is an open standard process model that describes common approaches used by data mining experts. it is the most widely-used analytics model[10]. According to CRISP-DM, The life cycle of a data mining project is broken down into six phases:

- Business Understanding
- Data Understanding
- Data Preparation

- Modeling
- Evaluation
- Deployment

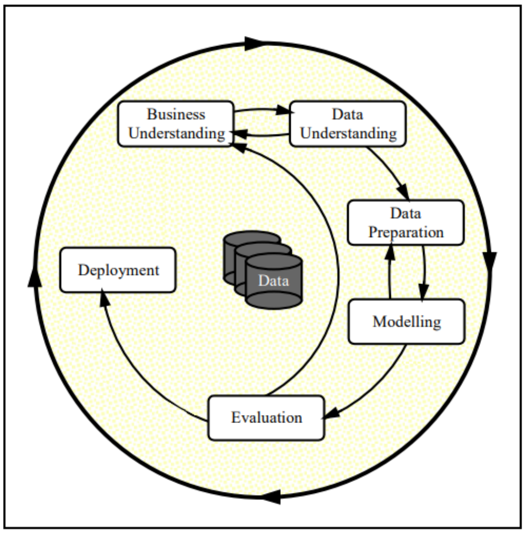


Figure 3: Phases of the Current CRISP-DM Process Model for Data Mining[11]

The Business Understanding phase focuses on understanding the objectives and requirements of the project, which is already presented in the Background section.

The data understanding phase starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information [11]. In section 1.2, a rough introduction has already been made to the dataset. In subsequent sections, we will explore the various properties of the data in more detail.

Data Preparation includes tasks like data cleaning, handling the missing values, remove redundant information and generate informative features.

In the phase modeling and evaluation, the models are selected, built, optimized, and tested. As we have a classification problem, supervised machine learning technic will be used in this thesis. There are many kinds of machine learning models for classification, such as Linear Classifiers, Logistic regression, Support Vector Machines (SVMs), K-means, Neural Networks, Decision Trees, and other tree based models like XGBoost and Random Forest [12]. However, for efficiency reasons, three models are built and tested in this thesis: Random Forest, K-Nearest-Neighboors, and Neuro Network. These models are generally considered to be effective in more complex classification problems and will be introduced in more detail in the section 4.

3 Data Preparation

Firstly, let's have a over view to the dataset.

P_ID	AGE	GENDER	DISEASE_1	...	OXY_S	WINDOW	ICU
0	60th	0	0	...	-1.00	0-2	0
0	60th	0	0	...	-1.00	2-4	0
0	60th	0	0	...	NaN	4-6	0
0	60th	0	0	...	-1.00	6-12	0
0	60th	0	0	...	-0.81	ABOVE_12	1
1	90th	1	0	...	-1.00	0-2	1
1	90th	1	0	...	-1.00	2-4	1
1	90th	1	0	...	-1.00	4-6	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
384	50th	1	0	...	-1.00	4-6	0
384	50th	1	0	...	-1.00	6-12	0
384	50th	1	0	...	-0.84	ABOV_12	0

As mentioned before, for 385 patients, we have in total 1925 rows and 231 columns in the original dataset. There are 224 features omitted in the middle of the Table for space reasons. The omitted features are, for example, 'IMMUNOCOMPROMISED', DISEASE GROUPING 2', 'CALCIUM_MEDIAN', 'CALCIUM_MEAN', 'BLOODPRESSURE_MEAN', 'HEART_RATE_MIN', 'TEMPERATURE_MIN' and so on.

3.1 Data Exploration

Next step is to do some more detailed exploration. To get a deeper understanding, the first thing is to explore the basic demographic information buried in the dataset. The figures below illustrates the Age and Gender information of all patients.

As we can see in Figure 4, the age distribution is relatively even. In terms of gender, there were significantly more infected male patients than female patients. It will be interesting to explore the correlation between age, gender and the ICU admission.
For the purpose of this, What we are going to do is to extract a very important information showed in the table below from the the data: Which patients are at which point admitted to the ICU.

P_ID	ICU	WINDOW	Age	Gender
0	1	ABOVE_12	60th	0
1	1	0-2	60th	1
2	1	ABOVE_12	10th	0
⋮	⋮	⋮	⋮	⋮
383	0	NaN	60th	0
384	0	NaN	50th	1

With the information above, we illustrate the relation between gender, age and the ICU admission below.

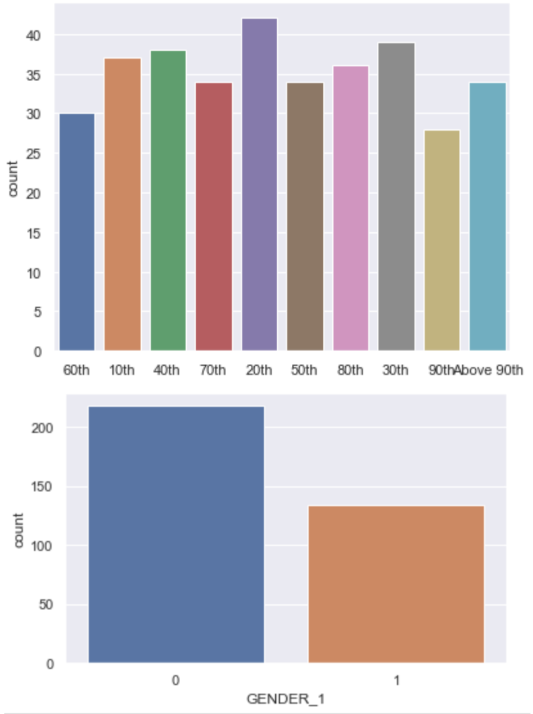


Figure 4: Age and gender distribution of 385 patients

According to the Figure 5, There was no significant difference between males and females in the relationship between gender and ICU admissions. Female patients were admitted to the ICU at a few higher rate. A clear correlation was observed in different age groups, but it was a negative one contrary to the general perception. The common perception is that the older people are, the more likely they are to be admitted to the ICU, but what we observed an opposite result.

3.2 Data Transformation

Transforming the data into a format that can be modeled effectively is the trickiest part of the Data Preparation session.

First restate that our goal is to predict whether the patient will be admitted to the ICU in the next window stage. As mentioned above, in the original data, each patient has five rows for five windows of the observation period. If the ICU is modeled directly as a target label without any feature transformation, the resulting model will be used to predict whether the patient should be admitted to the ICU at each window stage.

Therefore, we should take all the data before the patient gets ICU admission as input, and use whether the patient gets ICU admission in the next stage as

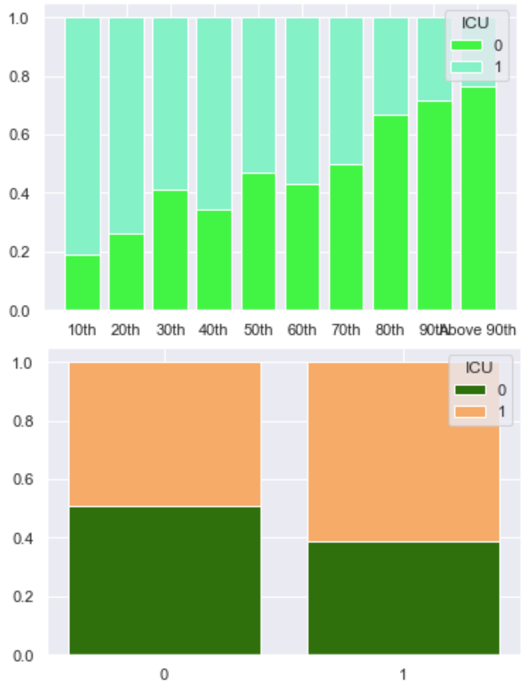


Figure 5: Age, gender and ICU admission

Target feature for model building. To achieve this, we need to transform the data as follows:

- Combine five rows of data for each patient in five observation windows into one row.
- Handling the missing values and correlation
- For patients who received ICU admission, determine the window in which they received the admission, and set all data for that window period and the periods after to NaN.
- For patients who did not receive ICU admission, keep the data from all window stages.

After the transformation is complete, we now have a new dataset where for each patient, the information for all of five windows is stored in one row, so our data has now 385 rows.

But this is not enough, there are still a lot of NaN values in our data. And because the data of the patient after ICU permission cannot be used for modeling, which means, the existing NaN values they cannot be filled.

Imagine a scenario where we Suppose a couple of ICU beds just became vacant, and the medical staff wants to determine which of the current patients is more likely to take these spots. In this scenario, it would be wise to look at the most recent available data to do the assessment. For example, for the patients who was admitted to the ICU at the window 6-12, we select the data from the last 2 available time windows to predict ICU admission.

A test has been done, in which the last 1 available time window was selected to build the model. But the performance was worse than using the last 2 time windows.

After the extraction is complete, we obtain a new, neat data set with 352 rows and 93 columns for modeling in the next section.

3.3 Correlation

Feature engineering is the task of improving predictive modelling performance on a dataset by transforming its feature space[13]. As a crucial part of feature engineering, feature selection is an effective way to reduce computational cost and enhance model performance.

Since we have a dataset with 93 columns, the dimension is relatively high. The resulting computational cost may not be significant in this small volume dataset, but if it is used in a real world scenario, the efficiency of the model is an important consideration.

Correlation analysis serves as a common but efficient way to do feature selection. Here the Pearson correlation coefficient[14] was used. The table blow shows the Pearson correlation between all of the features in descending order.

Feature-1	Feature-2	p-correlation
OXYGEN_begin	OXYGEN_REL_begin	ABOVE.999805
OXYGEN_end	OXYGEN_REL_end	0.999786
HEART_RATE_begin	HEART_RATE_MEDIAN_begin	0.993496
⋮	⋮	⋮
IMMU_end	OXYGEN_end	0.000079

we see some features are extremely correlated. Since features with a Pearson correlation score bigger than 0.9 are highly dependent on each other, discarding one of the both will not lead significant information lost.

After dealing with the correlation problem, the data set has now 352 rows and 52 columns.

4 Modeling

Machine Learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions[15]. The ML algorithms are able to read and modify its structure based on a set of observed data with adaptation done by optimizing over a cost function or an objective [16].

Supervised Machine Learning (SML) is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. Supervised classification is one of the tasks most frequently carried out by the intelligent systems[17].

In this thesis, we apply four of the multiple classification models for ICU admission prediction. They are:

- Random Forest Classifier
- KNN Classifier
- Neural Network Classifier

The random forest classifier consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector (Breiman)[18]. We use randomly selected features or a combination of features at each node to grow a tree and use them to build the random forest classifier.

K-Nearest Neighbor is a widely used text classifier because of its simplicity and efficiency. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors[19].

MLPClassifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. It relies on an underlying Neural Network to perform the task of classification. What we use is the MLPClassifier that ia implemented by Scikit-Learn in python language[20].

In the modeling training and evaluating process, the Leave-one-out k-fold-cross-validation was used. Cross-validation is also called out-of-sample testing. It is a model validation technique where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. In our modeling process, 5-fold-cross-validation was used.

Since In the chapter 3 we have done a lot of work on data preprocessing and transformation, the modeling can be done directly on the processed data. As an example, we use the code below to build the Random Forst Model.

```
#Split dataset into training and validation
x_train, x_validation, y_train, y_validation =
    train_test_split(data_2[feature_cols], data_2['ICU'],
            test_size = 0.1,shuffle = True)

#Cross-validate RF baseline model
baseline_model = RandomForestClassifier()
fitted_baseline_model = score_model(estimator =
      baseline_model,
              train_data = (x_train, y_train),
                validation_data = (x_validation ,
                y_validation), cv = 5)
```

With the open source machine learning toolkit scikit-learn in python, the modeling process becomes very simple and smooth.

5 Evaluation

To estimate how well a model will generalize to out-of-sample data and quantify the model performance, we need to evaluate the models that we built.

There lots of metrics that can be used to evaluate classification model performance. The most common metric is the accuracy.

$$accuracy = \frac{Number - of - correct - predictions}{Number - oftotal - number - of - prediction}$$

In addition to that, Area under the ROC Curve, also called AUC, is also a very popular metric. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

TP, FP, TN, FN stand for “true positive”, “false positive”, “true negative”, “false negative” and refer to the result of a test and the correctness of the classification.

One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0 and if the predictions are 100% correct has an AUC of 1. The advantages of AUC is that it s scale-invariant. It measures how well predictions are ranked, rather than their absolute values.

We will use these two metrics for evaluation.

As we can see in figure 6, the RF model achieved the highest accuracy score of 0.945 and auc score of 0.981. The MLPClassifier performed a little worse than the RF, but the scores were also above 0.9. The performance of KNN is not good, perhaps because the dataset has a relatively high dimension, which results in ”curse of dimensionality”. As the dimension increases, the indegree distribution of the k-NN digraph becomes skewed and the prediction becomes less precise.

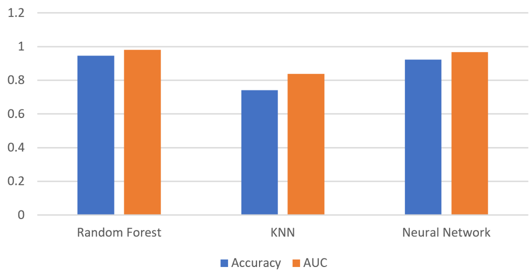


Figure 6: Model evaluation for ICU admission prediction

6 Future Work

In summary, we applied a data mining approach on ICU admission prediction for infected patients. With the model built, doctors can predict in an relatively accurate and practical way whether a patient needs to be sent to ICU in the next observation window. This helps to promote the rational allocation of ICU beds in the early stage, thus alleviating the problem of resource constraint in ICUs in severe epidemic areas.

Of course, there are many constraints in this paper. For example, the size of data used is small, and the generalizability of the trained model will be limited as a result. In addition, spending more time on parameter optimization during the model training process should further improve the performance of the model.

In the future, if possible, a long-term collaboration with a hospital and obtaining data on patients with corona virus infection before they are sent to the ICU would be very beneficial for this project.

References

1 Min Cao, Dandan Zhang, View ORCID ProfileYouhua Wang, View ORCID ProfileYunfei Lu, Xiangdong Zhu, Ying Li, Honghao Xue, Yunxiao Lin, Min Zhang, Yiguo Sun, View ORCID ProfileZongguo Yang, Jia Shi, Yi Wang, Chang Zhou, Yidan Dong, View ORCID ProfileLongping Peng, Ping Liu, Steven M. Dudek, Zhen Xiao, Hongzhou Lu, *Clinical Features of Patients Infected with the 2019 Novel Coronavirus (COVID-19) in Shanghai, China*, abstract

2 Chaolin Huang*, Yeming Wang*, Xingwang Li*, Lili Ren*, Jianping Zhao*, Yi Hu*, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rongmeng Jiang, Zhancheng Gao, Qi Jin, Jianwei Wang†, Bin Cao† *2. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China*, https://doi.org/10.1016/

3 Grasselli G, Pesenti A, Cecconi M.r*Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response* JAMA. 2020.

4 C Raina MacIntyre *Wuhan novel coronavirus 2019nCoV – update January 27th 2020*, Clinical features

5 Anna Odone, Davide Delmonte,Thea Scognamiglio, Carlo Signorellier *COVID-19 deaths in Lombardy, Italy: data in context*, https://doi.org/10.1016/S2468-2667(20)30099-2

6 GrazianoOnder, GiovanniRezza, SilvioBrusaferro *Case-Fatality Rateand Characteristics of Patients Dying in Relation to COVID-19 in Italy*, https://jamanetwork.com/ on 02/07/2021

7 Lauren Leatherby, John Keefe, Lucy Tompkins, Charlie Smart and Matthew Conlen *'There's No Place for Them to Go': I.C.U. Beds Near Capacity Across U.S.*, The New York Times

8 *Coronavirus digest: Germany ICU capacity at 'critical' level*, DW news

9 Stefan Knerr, Léon Personnaz, Gérard Dreyfus, *Handwritten digit recognition by neural networks with single-layer training*, IEEE Transactions on Neural Networks 3 (6) (1992) 962–968.

10 *What IT Needs To Know About The Data Mining Process*, Published by Forbes, 29 July 2015, retrieved June 24, 2018

11 Rüdiger Wirth, Jochen Hipp, *CRISP-DM: Towards a Standard Process Model for Data Mining*, p5

12 Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana Elias B. Khalil, Deepak Turaga *Learning Feature Engineering for Classification*, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)

13 Davidov, Ori; Ailon, Nir; Oliveira, Ivo F. D. (2018), *A New and Flexible Approach to the Analysis of Paired Comparison Data*, ournal of Machine Learning Research. 19 (60): 1–29.

14 *wikipedia-Pearson correlation coefficient*

15 Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar *Foundations of Machine Learning*, second edition

16 Jebara T. *Machine learning: discriminative and generative*, Norwell: Springer; 2003.

17 Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O.,Akinjobi J. *Supervised Machine Learning Algorithms:Classification and Comparison*, International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017

18 M. Pal *Random forest classifier for remote sensing classification*, International Journal of Remote Sensing, 26:1, 217-222

19 *wikipedia-k-nearest neighbors algorithm*, https://en.wikipedia.org/wiki/K-nearest$_{n}eighbors_{a}lgorithm$

20 *Scikit-learn: 1.17. Neural network models (supervised)*, https://scikit-learn.org/stable/modules/neural$_{n}etworks_{s}upervised.html$

21 *wikipedia-Curse of dimensionality*, https://en.wikipedia.org/wiki/Curse$_{o}f_{d}imensionality$

7 Erklärung

Ich versichere hiermit wahrheitsgemäß, die Arbeit selbständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderung entnommen wurde.

Datum, Unterschrift

10.02.2021