JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 910 - 917
https://publishoa.com
ISSN: 1309-3452

Recognition of Chronic Kidney Disease Using Machine Learning

K Prasad Babu[1], Syed Noorullah[2]

[1]Associate Professor, Dept of ECE, Ashoka Womens Engineering College, Affiliated to JNTUA, Kurnool, A.P, India
[2] Assistant Professor, Dept of ECE, Ashoka Womens Engineering College, Affiliated to JNTUA, Kurnool, A.P, India.

ABSTRACT

Chronic kidney disease (CKD) is a global prevalent ailment that causes lives in a predominant number. CKD is the 11th most deadly cause of global mortality with 1.2 million death each year and according to kidney Foundation of Bangladesh, around 40,000 CKD people experienced kidney failure annually as well as several thousand passed away in short stage of life because of CKD. Predictive analytics for healthcare using machine learning is a challenged task to help doctors decide the exact treatments for saving lives. Scientist researched collaboratively chronic kidney diseases, with the majority of their work on pure statistical models, generating numerous gaps in the development of machine-learning models. In this work we discussed the current methods and suggested improved technology based on the Correlation, which combined significant characteristics of the F scores and evaluated two pre- processing scenarios. In addition, we provided machine training methods for anticipating chronic renal disease with clinical information. Two techniques of master teaching are explored like Random Forest Classifier (RFC) and Logistic Regression (LR). The components are made from UCI dataset of chronic kidney disease and the results of these models are compared to determine the best regression model for the prediction. From this two preprocessing cases, replacing missing values with mean values of each column and choosing important features was most logical as it allows to train with more data without dropping. However, correlation gave the best outcomes in both two cases where it obtained 98% accuracy. Thus, the system can be implemented for early stage CKD prediction in a cost efficient way which will be helpful for under developed and developing countries.

Keywords: Chorionic Kidney Disease, Random Forest Classifier (RFC), Logistic Regression.

1. Introduction

Low- Chronic Kidney Disease (CKD) is considered as an important threat for the society with respect to the health in the present era. Chronic kidney disease can be detected with regular laboratory tests, and some treatments are present which can prevent development, slow disease progression, reduce complications of decreased Glomerular Filtration Rate(GFR) and risk of cardiovascular disease, and improve survival and quality of life. CKD can be caused due to lack of water consumption, smoking, improper diet, loss of sleep and many other factors.

This disease affected 753 million people globally in 2016 in which 417 million are females and 336 million are males. Majority of the time the disease is detected in its final stage and which sometimes leads to kidney failure. The existing system of diagnosis is based on the examination of urine with the help of serum creatinine level. Many medical methods are used for this purpose such as screening, ultrasound method. In screening, the patients with hypertension, history of cardiovascular disease, disease in the past, and the patients who have relatives who had kidney disease are screened. This technique includes the calculation of the estimated GFR from the serum creatinine level, and measurement of urine albumin-to-creatinine ratio (ACR) in a first morning urine specimen.

Our Contribution: So the main contribution of this work is after addressing feature reduction learning algorithm selection, tuning and comparing several machine learning techniques for CKD dataset, we found correlation method perform best. So it is then used as base model. Correlation method is used to rank the important features. After that we applied several machine learning techniques like, Random Forest Classifier model and Logistic Regression to compare the accuracy.

Machine learning is an application of artificial intelligence that involves algorithms and data that automatically analyzes and makes decision by itself without human intervention.It describes how computer perform tasks on their own by

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 910 - 917
https://publishoa.com
ISSN: 1309-3452

previous experiences. Therefore we can say in machine language artificial intelligence is generated on the basis of experience.The difference between normal computer software and machine learning is that a human developer hasn't given codes that instruct the system how to react to situation, instead it is being trained by a large number of data.

The machine Learning is not dependent on any explicit programming, but the data fed into it. Based on the data you feed into machine learning algorithm and the training given to it, an output is delivered.

Supervised Machine Learning:

Supervised learning is the type of machine learning in which machines are trained using well "labeled" training data, and on basis of that data, machines predict the output. The labeled data means some input data is already tagged with the correct output. We can say like, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

There are two types of supervised machine learning:

Classification - Classification predicts the category to which a new observation belongs.

Regression - Predicting a value from a continuous data set.

Here in our work we are dealing with a classification machine learning problem.

Unsupervised Machine Learning

As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset. It means, in unsupervised machine learning models are trained using unlabeled dataset and are allowed to act on that data without any supervision. In unsupervised learning, the models are trained with the data that is neither classified nor labeled, and the model acts on that data without any supervision.

2. Existing Methods

A large [J. Snegha, 2020][10] proposed a system that uses various data mining techniques like Random Forest algorithm and Back propagation neural Network. Here they compare both of the algorithm and found that Back Propagation algorithm gives the best result as it uses the supervised learning network called feed forward neural network.[Mohammed Elhoseny, 2019] described a system for CKD in which it uses Density based feature selection with ACO. The system uses wrapper methods for feature selection.[Baisakhi Chakraborty, 2019][9] proposed development of CKD prediction system using machine learning techniques such as K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine and Multi-Layer Perceptron Algorithm. These are applied and their performances are compared to the accuracy, precision, and recall results. Finally, Random forest is chosen to implement this system. [Arif-Ul-Islam, 2019] proposed a system in which prediction of disease is done using Boosting Classifiers, Ant-Miner and J48 Decision Tree. The aim of this paper is twofold that is, analyzing the performance of boosting algorithms for detecting CKD and deriving rules illustrating relationships among the attributes of CKD. Experimental results prove that the performance of AdaBoost was less that of LogitBoost by a fraction. [S.Belina V, 2018] proposed a system that uses extreme learning machine and ACO for CKD prediction. Classification is done using MATLAB tool and ELM has few constraints in the optimization. This technique is an improvement under the Sigmoid additive type of SLFNs. [Siddheshwar Tekale, 2018][8] described a system using machine learning which uses Decision tree SVM techniques. By comparing two techniques finally concluded that SVM gives the best result. Its prediction process is less time consuming so that doctors can analyze the patients within a less time period. [Nilesh Borisagar, 2017] described a system which uses Back Propagation Neural Network algorithm for prediction. Here Levenberg, Bayesian regularization, Scaled Conjugate and resilient back propagation algorithm are discussed. Matlab R2013a is used for the implementation purpose. Based on the training time, scaled conjugate gradient and resilient back propagation are found more efficient than Levenberg and Bayesian regularization. [Guneet Kaur, 2017][7] proposed a system for predicting the CKD using Data Mining Algorithms in Hadoop. They use two data mining classifiers like KNN and SVM. Here the predictive analysis is performed based upon the manually selected data columns. SVM classifier gives the best accuracy than KNN in this system. [Neha Sharma, 2016] proposed a system in which the kidney disease of a patient is analyzed and the results are to compute automatically using the data set of the patient. Here Rule based prediction method is used. This system uses neuro-fuzzy method and obtained the outcome by

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 910 - 917
https://publishoa.com
ISSN: 1309-3452

mathematical computation.[Kai-Cheng Hu, 2015][6] proposed a system which uses a multiple pheromone table based on ACO for clustering. Here they divided the problem into a set of several different patterns based on their features. Two pheromone tables are used here one for keeping the track of the promising information and the other to hold the details of unpromising information which in turn increases the probability of searching directions.

3. Design Methodology

Acc-A, There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms that you can choose according to the objective that you might have it may be Classification algorithms or Regression algorithms.

Linear Regression.

Logistic Regression.

Random Forest Regression / Classification.

Decision Tree Regression / Classification.

As the prediction for model is classification type, we apply a logistic regression algorithm and random forest classifier algorithm on our dataset. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Random forest classifier model is based onthe concept ofensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Once the model is trained, it's readyto make predictions. Finally, There is a need to check to see how well our model is performing on the test data. There are many evaluation techniques are there. For this, we evaluate the accuracy score produced by the model. Confusions Matrix for the model is been used.

THEORITICALANALYSIS

Following are the steps that are used in this work.

Download the dataset

Preprocess or clean the data

Import the libraries

Read the dataset

analyze the dataset

drop unnecessary columns

change the column names

remove the randomness in the columns

find the missing values

handle the missing values

split the data into independent and dependent variables

split the data to train and test

Train the machine with preprocessed data with an Appropriate Machine learning algorithm to build a model

Save the model and its dependencies

Build a Web application using flask that integrates with Model built.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 910 - 917
https://publishoa.com
ISSN: 1309-3452

Model is based on logistic regression, and it obtains the weight of each predictor and a bias.

If the sum of the effects of all predictors exceeds a threshold, the category of the sample will be classified as CKD or not CKD.

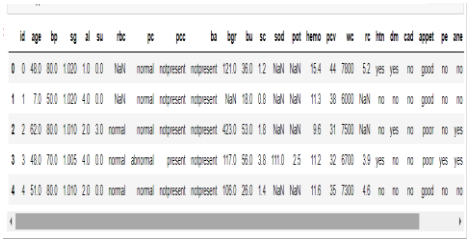
Dataset and Methods

Dataset:

The Dataset here we use is the publically available CKD Dataset from UCI repository. It contains 400 samples of two different classes. Out of 25 attributes, 11 are numeric and 13 are nominal and one is class attribute. The data set contains number of missing values. Here the information of dataset uses the patient's data like age, blood pressure, specific gravity, albumin, sugar, red blood cells etc. CKD is caused due to diabetes and high blood pressure. Dueto Diabetes our many organs get affected and it will be followed by high blood sugar. So it is important to predict the disease as early as possible. This study improvises some of the machinelearning techniques to predict the disease.

Table.1 List of attributes present in the CKD dataset

Attributes	Type
Age	Numeric
Blood Pressure	Numeric
Specific Gravity	Numeric
Albumin	Numeric
Sugar	Numeric
Red Blood Cells	Nominal
Pus Cell	Nominal
Pus Cell clumps	Nominal
Bacteria	Nominal
Blood Glucose Random	Numeric
Blood Urea	Numeric
Serum Creatinine	Numeric
Sodium	Numeric
Potassium	Numeric
Hemoglobin	Numeric
Packed Cell Volume	Numeric
Red Blood Cell count	Numeric
White Blood Cell Count	Numeric
Hypertension	Nominal
Diabetes Mellitus	Nominal
Coronary Artery Disease	Nominal
Appetite	Nominal
Pedal Edema	Nominal
Anemia	Nominal
Class	Class



id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	36.0	1.2	NaN	NaN	15.4	44	7800	5.2	yes	yes	no	good	no	no
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	18.0	0.8	NaN	NaN	11.3	38	6000	NaN	no	no	no	good	no	no
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8	NaN	NaN	9.6	31	7500	NaN	no	yes	no	poor	no	yes
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4	NaN	NaN	11.6	35	7300	4.6	no	no	no	good	no	no

Classification
CKD CKD CKDCKD CKD

Table1.1: Dataset of 400 patients

Steps:

Data Pre-Processing

Data Pre-Processing is that stage where the data that is distorted or encoded is brought to such a state that the machine can easily analyze it. A dataset can be observed as a group of data objects. Data objects are labeled by a number of features, that ensures the basic features of an object, such as the mass of a physical object or the time at which an event ensured. In the dataset there may be missing values, they can either eliminated or estimated. The most common method of dealing with missing values is filling them in with mean, median or mode value of respective feature. As object values cannot be used for the analysis we have to convert the numeric values with type as object to float64 type. Null values in the categorical attributes are changed with the most recurrent occurring value current in that attribute column. Label encoding is done to translate categorical attributes into numeric attribute by conveying each unique attribute

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 910 - 917
https://publishoa.com
ISSN: 1309-3452

value to an integer. This automatically changes the attributes to int type. The mean value is premeditated from each column and is used to replace all the missing values in that attribute column. For this function we are using a function called imputer which is used to find the mean value in each column. After the replacing and encoding is done, the data should be trained, validated and tested. Training the data is the part on which our algorithms are actually trained to build a model. Validation is the part ofthe dataset which is used to validate our various model fits or improve the model. Testing the data is used to test our model hypothesis.

Here in my dataset there are so many missing values. Prior to model building, data preprocessing is required to remove unwanted noise and outliers from the dataset that might cause the model to diverge from the proper training set. This stage tackles anything that is impeding the model's efficiency. After collecting the necessary data, it must be cleaned and prepared for the model construction. The dataset is next searched for the null values.

Feature Selection

Feature Selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling. Before implementing any technique, it is really important to understand, need for the technique and so for the Feature Selection. As we know, in machine learning, it is necessary to provide a pre-processed and good input dataset in order to get better outcomes. We collect a huge amount of data to train our model and help it to learn better. Generally, the dataset consists of noisy data, irrelevant data, and some part of useful data. Moreover, the huge amount of data also slows down the training process of the model, and with noise and irrelevant data, the model may not predict and perform well. So, it is very necessary to remove such noises and less- important data fromthe dataset and to do this, Feature selection techniques are used.

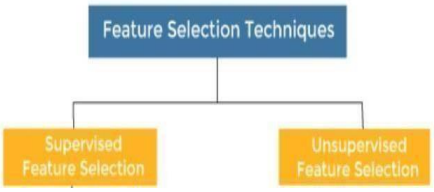


Figure2: Types of feature selection techniques

Supervised Feature selection techniques consider the target variable and can be used for the labeled dataset.

Unsupervised Feature Selection Technique:

Unsupervised Feature selection techniques ignore the target variable and can be used for the unlabelled dataset. In some cases, to improve the performance of the model Statistical-based feature selection methods involve in evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.

Correlation is a statistical term which in common usage refers to how close two variables are to having a linear relationship with each other. For example, two variables which are linearly dependent (say, x and y which depend on each other as x = 2y) will have a higher correlation than two variables which are non-linearly dependent (say, u and v which depend on each other as u = v2).

4. Simulations

Above table shows the scaling operation of features. "Equalizing the Magnitude of data is Scaling". Here Magnitude of every column will change only for independent variables and it gives importance to the feature containing large values in every column.

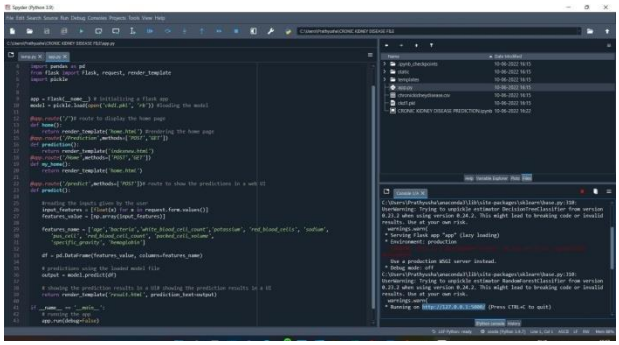
JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 910 - 917
https://publishoa.com
ISSN: 1309-3452

```
In [127]: #Here 0 indicates disease and 1 indicates no_disease
          new_df.loc[new_df['class']==0, 'class'] = 'disease'
          new_df.loc[new_df['class']==1, 'class'] = 'no_disease'

In [128]: y=new_df['class']
          y.head()

Out[128]: 0    disease
          1    disease
          2    disease
          3    disease
          4    disease
          Name: class, dtype: object

In [129]: #Target containing 250 disease and 150 no_disease
          y.value_counts()

Out[129]: disease       250
          no_disease    150
          Name: class, dtype: int64
```

```
In [136]: print(classification_report(y_test,y_predict1))
```

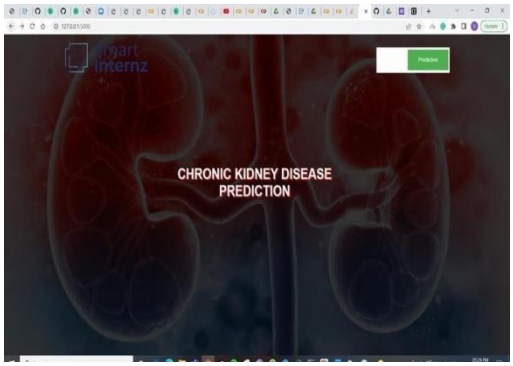
	precision	recall	f1-score	support
disease	1.00	0.98	0.99	52
no_disease	0.97	1.00	0.98	28
accuracy			0.99	80
macro avg	0.98	0.99	0.99	80
weighted avg	0.99	0.99	0.99	80

The above table visualizes and summarizes the performance of a RFC classification algorithm. Here out of 52 positive samples 51 are correctly predicted, and out of 28 negative samples all are correctly predicted. Here there is no miss classification for negative class.



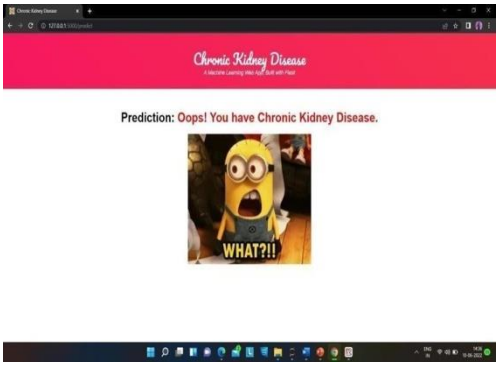
When the above code is executed the below webpage is displayed.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 910 - 917
https://publishoa.com
ISSN: 1309-3452



When we click on prediction available on home page, it get opens like the image shown below.

If the person is having Chronic Disease the screenshot is as shown below.



5. Conclusion & Discussion

The sensitivities of the classifier and word-count methods were 95.4% and 99.8%, respectively. The specificity of both was 99.8%. Categorization of individual patients as appropriately documented was96.9% accurate. Of 107 patients with manually verified moderate CKD, 32 (22%) lacked appropriate documentation. Patients whose CKD had not been appropriately documented were significantly less likely to be on renin-angiotensin system inhibitors or have urine protein quantified, and had the illness for half as long (15.1 vs 30.7 months; $p<0.01$) compared to patients with documentation. The result of each classifier has been evaluated using Confusion Matrix and saved in the form of pickle form in order to use it in web application which is been made with the help of FLASK.

work examines the ability to detect CKD using machine learning algorithms while considering the least number of tests or features. We approach this aim by applying two machine learning classifiers like Random Forest Classifier model and Logistic Regression. In order to reduce the number of features and remove redundancy, the association between

variables have been studied. In the end we showed the results in the form of Accuracy scores, recall, precision, F1 Score. Based on the above analysis, we found that Random Forest Classifier Model have the best accuracy and best F1 Score among all classification algorithms. The feature selection by correlation method found that there are hemoglobin, albumin have the most impact to predict the CKD. Also, we found that hemoglobin has the highest contribution in detecting CKD. For deployment purpose we used Machine Learning concept and we tried to create an app using python IDE called spyder. Through this app by entering some feature fields related to chronic kidney disease we predicted that, the person is diseased or not

References

[1]. V. Jha , G. Garcia-Garcia , K. Iseki , Z. Li , S. Naicker, B. Plattner, R. Saran, A. Y. Wang, C.W. Yang (2013),"Chronic kidneydisease: global dimension and perspectives", The Lancet.

[2]. Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016). Chronic Kidney Disease analysis using data mining classification techniques. In 2016 6th International Conference- Cloud System and Big Data Engineering (Confluence) (pp. 300-305)

[3]. L. Xun, Wu Xiaoming, Li Ningshan and Lou Tanqi, "Application of radial basis function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, 2010, pp. V15-332-V15-335.

[4]. A. Salekin and J. Stankovic, (2016) "Detection of chronic kidney disease and selecting important predictive attributes," IEEE International Conference on Healthcare Informatics.

[5]. D. Gupta, S. Khare, and A. Aggarwal, (2016)"A method to predict diagnostic codes for chronic diseases using machine learning techniques," 2016 International Conference on Computing, Communication and Automation (ICCCA), pp. 281–287.

[6]. A. Y. Al-Hyari, A. M. Al-Taee and M. A. Al-Taee, (2013) "Clinical decision support system for diagnosis and management of Chronic Renal Failure," 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, pp. 1-6.

[7]. Q. Zheng, G. Tasian, and Y. Fan, "Transfer learning for diagnosis of congenital abnormalities of the kidney and urinary tract in children based on Ultrasound imaging data," International Symposium on Biomedical Imaging, vol. abs/1801.0, no. Isbi, pp. 1487–1490, 2018.

[8]. S. P. Deng, S. Cao, D.-S. Huang, and Y.-P. Wang, (2017) "Identifying Stages of Kidney Renal Cell Carcinoma by Combining Gene Expression and DNA Methylation Data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 5, pp. 1147–1153.

[9]. Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR), 6(2).

[10]. Good, David M., Petra Zürbig, Angel Argiles, HartwigW. Bauer, Georg Behrens, Joshua J. Coon, Mohammed Dakna et al. "Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease." Molecular & cellular proteomics 9, no. 11 (2010): 2424- 2437.

[11]. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[12]. Ogunleye, A. A., & Qing-Guo, W. (2019). XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM transactions on computational biology and bioinformatics.

[13]. Ogunleye, A., & Wang, Q. G. (2018, June). Enhanced XGBoost-Based Automatic Diagnosis System for Chronic Kidney Disease. In 2018 IEEE 14th International Conference on Control and Automation (ICCA) (pp. 805-810).

[14]. T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[15]. R. A. Abbasi, N. Javaid, M. N. J. Ghuman, Z. A. Khan, Z. S. U.Rehman & Amanullah. (2019), "Short Term Load Forecasting Using XGBoost." In: Barolli L., TakizawaM., XhafaF., Enokido T. (eds) Web, Artificial Intelligence and Network Applications. WAINA2019.Advances in Intelligent Systems and Computing, vol 927. Springer, Cham.