➔ ***Regular Research Paper – NS***

Prediction of Heart Failure Exitus with Machine Learning Classification Algorithms

Usame Omer OSMANOGLU*
Department of Biostatistics, Eskisehir Osmangazi University Faculty of Medicine, Eskisehir, Turkey
usa.osm@gmail.com

Fezan MUTLU
Department of Biostatistics, Eskisehir Osmangazi University Faculty of Medicine, Eskisehir, Turkey
fsahin@ogu.edu.tr

Abstract

Aim: In this study, it is aimed to contribute to the artificial intelligence (AI) literature in the field of health and to the researchers who want to work in this field, by presenting the relevant theoretical infrastructure, with an application study on the estimation of heart failure mortality with ML classification algorithms.

Materials and Methods: In this study, ANN (artificial neural network), SVM (support vector machine), NB (naive bayes) classifier, KNN (k nearest neighbor), LR (logistic regression), DT (decision tree) and RF (random forest) algorithms, which are machine learning classification methods, were used to predict heart failure mortality. In order to increase the number of data, synthetic data derivation was applied. In addition, cross validation was applied to increase model accuracy. Model success was measured by confusion matrix and ROC AUC (Receiver Operating Characteristic, Area Under the Curve) score.

Results: In the practice study, it was determined that the risk factors for heart failure mortality were the duration of patient follow-up, ejection fraction, serum creatinine level and age of the patient. As a result of the application, 85.0% accuracy, 78.1% sensitivity, 88.2% specificity and 83.1% ROC AUC values were reached with Randım Forrest algorithm.

Discussion and Conclusion: In conclusion, it has been seen that the use of machine learning classification algorithms in the estimation of cardiac mortality has the potential to provide an important contribution to physicians as a decision support mechanism.

Keywords: *Machine Learning, Classification, Prediction, Heart Failure.*

Available Online: 02.12.2021

*This study was produced from the doctoral thesis prepared by the first author under the supervision of the second author.

1. INTRODUCTION

In the 21st century, which can be called the age of artificial intelligence, machine learning methods that become widespread and can improve themselves can provide better quality services to humanity in many areas. As a result of these developments, machine learning (ML) has started to provide services in many areas such as diagnosis and diagnostic estimation to support physicians in decision making, as in many sectors today.

Machine learning is the science of computational statistics based on making predictions using computers. Machine learning algorithms, a sub-topic of artificial intelligence, create a statistical sample data model, called training data, by making predictions or decisions without being explicitly programmed. It focuses on deriving predictions on test data from learned data based on known features. Machine learning algorithms are used in various applications such as medicine, e-mail filtering and computer vision, where it is difficult to develop traditional algorithms to achieve the needed goals. Machine learning is a subfield of computational statistics that focuses on making predictions using computers [1].

Heart failure, deterioration in filling or pump functions caused by structural or functional disorders of the heart; It is a complex clinical syndrome characterized by fatigue, shortness of breath at rest with exertion, orthopnea, paroxysmal nocturnal dyspnea, nocturia, mental status changes, anorexia, and abdominal pain [2]. The American Heart Association predicts an estimated 46% increase in acute heart failure from 2012 to 2030, and that by 2030, more than 8 million people aged 18 and over in the United States will have acute heart failure. According to the results of the Heart Failure Prevalence and Predictors in Turkey (HAPPY) study, which reflects the situation in Turkey, more than 2 million individuals in Turkey have heart failure. The survival rates of these patients are lower than those of bowel, breast or prostate cancers. However, it puts a serious burden on caregivers, healthcare workers and the healthcare system for patients with heart failure [3].

The aims of this study are to explain the ML method, algorithms, working principles, application steps, model training and outcome evaluation in an applied way to researchers who are planning an artificial intelligence study, especially for the estimation of heart failure mortality in the field of health, and to evaluate the performance of ML classification methods in estimating heart failure mortality. In this study, the heart failure Prediction [4] name was used to evaluate data preprocessing methods such as data identification, data sampling, synthetic data derivation, and machine learning classification algorithms applied while developing a Decision Support System (DSS) that can help physicians. An application study was carried out on the open data source that contains the data. Thus, it is aimed to help researchers who want to develop similar DSS in the field of health, especially in the estimation of heart failure mortality, in artificial intelligence applications.

2. MATERIALS AND METHODS

An application study was conducted with the heart failure dataset named “Heart Failure Prediction” [4], which was taken from an open data source, in order to evaluate data preprocessing and machine learning algorithms such as data identification, data sampling, synthetic data derivation applied while developing a decision support system that can help physicians. A data set with 13 variables (6 categorical, 7 numeric) with a sample volume of 299 was created (Table 1-2.). It is thought that this study will help researchers who want to develop DSS in the field of health, especially in artificial intelligence applications.

The codes used in this study were created using the Python Programming Language. Pandas, numpy, statmodels, matplotlib and sklearn libraries included in the Python Program were used.

Table 1. Frequency table of categorical variables in the data set.

Variable	Category	Frequency	Percentage
anemia	0 (no)	170	57.0
	1 (yes)	129	43.0
diabetes	0 (no)	174	58.2
	1 (yes)	125	41.8
high blood pressure	0 (no)	194	64.9
	1 (yes)	105	35.1
gender	0 (female)	105	35.1
	1 (male)	194	64.9
smoking	0 (no)	203	67.9
	1 (yes)	96	32.1
death event	0 (alive)	203	67.9
	1 (ex)	96	32.1
Total		299	100.00

Table 2. Frequency table of categorical variables in the data set.

Variable	Mean	Median	Standard Deviation	Minimum Value	Maximum Value
age	60.83	60.00	11.90	40	95
creatinine phosphokinase	581.84	250.00	970.29	23	7861
ejection fraction	38.08	38.00	11.84	14	80
platelets	263358.03	262000.00	97804.24	25100.00	850000.00
serum creatinine	1.39	1.10	1.035	0.50	9.40
serum sodium	136.63	137.00	4.41	113	148
time	130.26	115.00	77.61	4	285

2.1. ROC (Receiver Operating Characteristic) Curve

It allows the determination of appropriate cut-off points to determine the optimum sensitivity and optimum specificity of a medical test. A reference is needed to determine the appropriate cut-off point via the ROC curve. ROC curve method uses values such as Sensitivity, Specificity, Accuracy Rates. Quantitative data of the variable measured in clinically or pathologically ill or healthy individuals are used to determine the appropriate cut-off point [5].

2.2. Cross Validation

Cross-validation is a commonly used method for estimating the performance of a classification algorithm or comparing performance between two classification algorithms on a dataset. This procedure divides a data set into k randomly and approximately equal sized pieces, and each piece is used to test a classification algorithm and the model consisting of other pieces. The performance of the classification algorithm is evaluated by the average of the k accuracies resulting from the cross validation [6].

2.3. Synthetic Minority Oversampling Technique (SMOTE)

Smote is a data derivation method. Unbalanced data is when the number of units of the subcategories of a dependent variable is very different from each other. In artificial intelligence applications, estimations in unstable data sets are biased towards the sub-category with a large number of units. This method is used in both statistical sampling and machine learning [7]. With the Smote method, it is the process of subtracting the unit number of the subcategory of a variable with a low number of units to the number of units of the subcategory with a large number of units. With this process, synthetic data suitable for the statistical parameters of the raw data set are derived [8].

2.4. Python Programming Language

Python is an interpretative, object-oriented, highly interactive and modular programming language. Since its simple syntax based on indentations facilitates learning and memorization of the language, it provides the feature of being a programming language that can be started to be programmed easily [9].

2.5. Statistical Analysis

A machine learning and statistical classification problem in particular, a complexity matrix is a tabular layout that visualizes the performance of an algorithm. If the predicted variable (dependent variable, target, target, output, output) is in binary format, the accuracy is evaluated with the evaluation complexity matrix. Using this matrix, many performance criteria such as sensitivity, specificity, precision, negative predicted value (npv), accuracy and f1-score (f1 score) can be calculated (Table 3). The accuracy formula is generally used to evaluate the success of the prediction model [10].

Table 3. Confusion matrix.

		Actual		**Total**
		1 (Positive)	**0 (Negative)**	
Prediction	**1 (Positive)**	True Positive (TP)	False Positive (FP)	Precision, Positive Prediction Value
	0 (Negative)	False Negative (FN)	True Negative (TN)	Negative Predicted Value (NPV)
Total		Recall, Sensitivity	Specificity	Accuracy

Precision =TP/(TP+FP)
NPV =TN/(FN+TN)
Sensitivity =TP/(TP+FN)
Specificity =TN/(TN+FP)
Accuracy =(TP+TN)/(TP+FP+FN+TN)

3. RESULTS

n the variable selection stage, stepwise and forward methods chose the same variables according to the output obtained. Operations were continued with stepwise variables. According to this result, time, ejection fraction, serum creatinine and age variables affect heart failure mortality at a statistically significant level. The resulting output is shown below. The plot showing the importance percentages of the selected variables is shown in Figure 2. According to this plot, the factor affecting heart failure mortality at the highest level among the four selected variables is the

time variable showing the follow-up period of the disease. Other factors are serum creatinine level, ejection fraction and age variables, respectively.

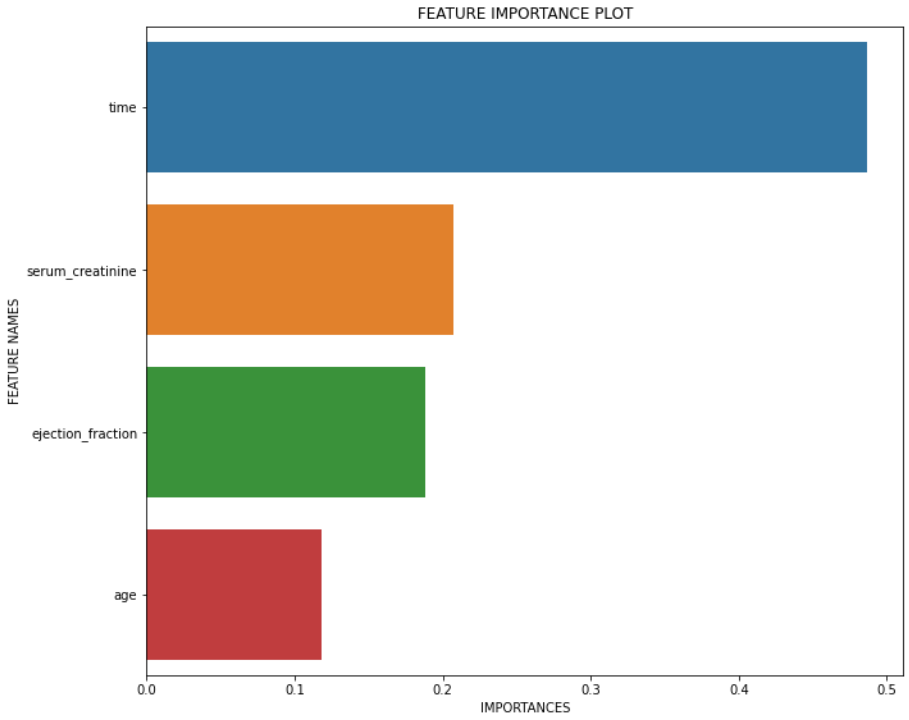

Figure 1. Importance plot of the features selected according to the stepwise method.

According to the output of the Smote application, before the smote process, the data set was divided into 239 (n_{alive} =163, n_{ex} =76) training data and 60 (n_{alive}=40, n_{ex}=20) test data with 60 units. After the Smote process, the training data set reached 326 (n_{alive} =163, n_{ex} =163) units. The total number of data increased from 299 to 386. Stratified sampling method was preferred while separating the data into training and test data. In this way, the distribution of the dependent variable in the raw data was kept constant in the two separated data sets. Thus, the estimation results ensured their reliability in terms of reflecting the real situation. The ROC AUC graph of the algorithms used is given in Figure 2.

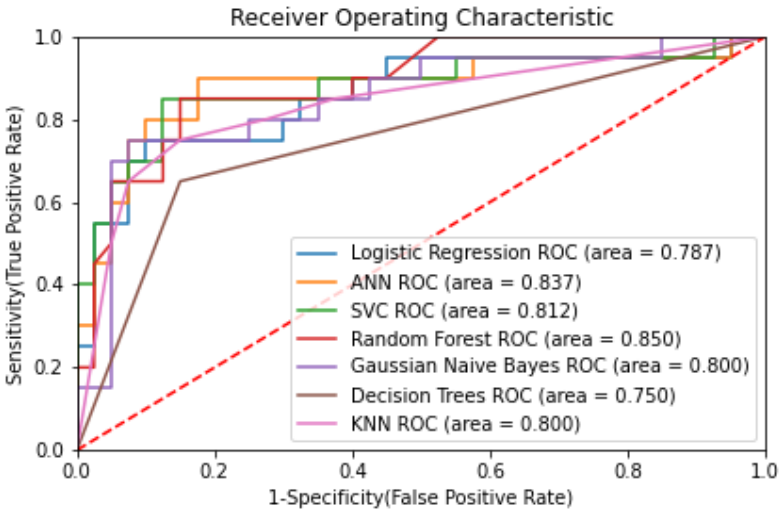

Figure 2. ROC AUC graph.

Table 4 presents the accuracy, ROC AUC score, sensitivity, specificity, negative predictive value, and precision measures of the LR, ANN, SVM, RF, NB, DT, and KNN algorithms at each fold of cross-validation.

Table 4. Test results by cross validation folds.

Fold	Algorith	Accuracy	ROC AUC	Sensitivity	Specificity	Negative Predictive Value	Precision
1	LR	0.800	0.787	0.750	0.825	0.868	0.682
	ANN	0.833	0.837	**0.850**	0.825	0.917	0.708
	SVM	0.833	0.813	0.750	**0.875**	0.875	**0.**750
	RF	**0.850**	**0.850**	**0.850**	0.850	**0.919**	0.739
	NB	0.817	0.800	0.750	0.850	0.872	0.714
	DT	0.783	0.750	0.650	0.850	0.829	0.684
	KNN	0.817	0.800	0.750	0.850	0.872	0.714
2	LR	0.733	0.720	0.684	0.756	0.838	0.565
	ANN	0.783	0.757	0.684	0.829	0.850	0.650
	SVM	**0.833**	**0.807**	**0.737**	**0.878**	**0.878**	**0.737**
	RF	0.733	0.706	0.632	0.781	0.821	0.571
	NB	0.767	0.759	**0.737**	0.781	0.865	0.609
	DT	0.750	0.746	**0.737**	0.756	0.861	0.583
	KNN	0.767	0.759	**0.737**	0.781	0.865	0.609
3	LR	0.783	0.785	**0.790**	0.781	0.889	0.625
	ANN	0.767	0.702	0.526	0.878	0.800	0.667
	SVM	**0.833**	0.765	0.579	**0.951**	0.830	**0.846**
	RF	**0.833**	**0.822**	0.789	0.854	**0.897**	0.714
	NB	0.783	0.757	0.684	0.829	0.850	0.650
	DT	0.700	0.696	0.684	0.707	0.829	0.520
	KNN	0.800	0.769	0.684	0.854	0.854	0.684
4	LR	0.800	0.811	0.842	0.781	0.914	0.640
	ANN	**0.900**	**0.913**	**0.947**	0.878	0.973	0.783
	SVM	0.800	0.755	0.632	0.878	0.837	0.706
	RF	0.883	0.844	0.737	**0.951**	0.886	**0.875**
	NB	0.817	0.809	0.789	0.829	0.895	0.682
	DT	0.867	0.832	0.737	0.927	0.884	0.824
	KNN	0.783	0.771	0.737	0.805	0.868	0.636
5	LR	0.831	0.834	0.842	0.825	0.917	0.696
	ANN	0.915	0.896	0.842	0.950	0.927	0.889
	SVM	0.932	0.909	0.842	**0.975**	0.929	0.941
	RF	**0.949**	**0.935**	**0.895**	**0.975**	**0.951**	**0.944**
	NB	0.949	0.935	0.895	0.975	0.951	0.944
	DT	0.831	0.792	0.684	0.900	0.857	0.765
	KNN	0.848	0.846	0.842	0.850	0.919	0.727

* LR: Logistic Regression, ANN: Artificial Neural Network, SVM: Support Vector Machine, RF: Random Forest, NB: Naive Bayes, DT: Decision Tree, KNN: k Nearest Neighborhood; ROC: Receiver Operating Characteristic, AUC: Area Under the Curve

The test results according to the mean of the cross-validation multiples are given in Table 5. According to this table, the highest rates of 85.0% accuracy and 83.1% ROC AUC score were achieved with the RF algorithm.

Table 5. Test results relative to the mean of cross-validation folds.

	LR	ANN	SVM	RF	NB	DT	KNN
Accuracy	0.789	0.840	0.846	**0.850**	0.827	0.786	0.803
ROC AUC	0.787	0.821	0.810	**0.831**	0.812	0.763	0.789
Sensitivity	**0.782**	0.770	0.708	0.781	0.771	0.698	0.750
Specificity	0.794	0.872	**0.911**	0.882	0.853	0.828	0.828
Negative Predictive Value	0.885	0.893	0.870	**0.895**	0.887	0.852	0.876
Precision	0.642	0.739	**0.796**	0.769	0.720	0.675	0.674

* LR: Logistic Regression, ANN: Artificial Neural Network, SVM: Support Vector Machine, RF: Random Forest, NB: Naive Bayes, DT: Decision Tree, KNN: k Nearest Neighborhood; ROC: Receiver Operating Characteristic, AUC: Area Under the Curve

The complexity matrix calculated in the 1st layer of the RF algorithm, which gives the highest success and ROC AUC score, is given in Table 6.

Table 6. Test results relative to the mean of cross-validation folds.

		Actual	
		1 (Positive)	**0 (Negative)**
Prediction	**1 (Positive)**	TP=17	FP=6
	0 (Negative)	FN=3	TN=34

* TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative

4. DISCUSSION AND CONCLUSION

Although the history of machine learning goes back several decades, it was initially possible due to the huge computational requirements and the limitations of the computing power available at that time. However, with the explosion of information in machine learning and the advancement of computer technology, a great movement has started in recent years. As a result of this mobility, serious progress has been made in a short time in the field of artificial intelligence with the knowledge unit of statistics. Although the contribution of statistical methods to machine learning is so great, there are some differences between these two disciplines. The difference between machine learning and statistics is their purpose. Machine learning models are designed to make the most accurate predictions possible. Statistical models, on the other hand, are designed to make inferences about the relationships between variables. But machine learning is built on statistical science [11].

In the literature, it has been observed that artificial intelligence methods have been used for the predictions of heart failure since 2003, with 50-8200 units and 8-236 variable data sets, success rates between 61-98% and ROC AUC scores [12-18].

According to the application results in this study, the RF algorithm gave the highest accuracy rate with 85.0%. It was observed that the lowest accuracy rate was 78.6% with the DT algorithm. Even in this data set, which consists of a small number of units and variables with approximately 300 people, and has an imbalance situation, the fact that the estimation success is close to 80% is thought to be promising for future studies. It is thought that it is quite possible to achieve higher estimation success by increasing the number of units and variables.

Data set creation is the first and most important step in a machine learning study, as the entire system is built on the data set. Data pre-processing steps should be applied carefully in accordance with the purpose of the study. Apart from the data preprocessing methods applied in this study, there are also methods that can be used for different purposes.

The Smote method should be applied to the training data set [19]. If the whole dataset is divided into training and test datasets after smote is applied, synthetic data will also be included in the test dataset. Therefore, data similar to the data in the training will be used for the test. In other words, data similar to learned data were used for testing and may lead to unrealistically high success rates.

In this study, the prediction performances of machine learning classification algorithms in estimating heart failure exitus were compared. It is very important to minimize the error in areas where the error is fatal, such as the health area. Despite the fact that physicians are very experienced due to their humanitarian situation, there is always the possibility of making mistakes in diagnosis and diagnosis. The use of artificial intelligence in the field of health; It is thought and recommended that it can provide vital assistance to physicians in decision support with its advantages such as making diagnosis and diagnosis with high success rates, processing large data sets quickly and without errors. It is thought that the use of artificial intelligence studies, especially in the field of health, will increase in the future. Therefore, it is thought that this study will make an important contribution to the digital transformation in health and to the literature.

REFERENCES

[1] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill, New York. 2-3. ISBN: 0070428077.

[2] Kepez, A., & Kabakçı, G. (2004). Kalp yetersizliği tedavisi. Acta Medica, 35(2), 69-81. (https://actamedica.org/index.php/actamedica/article/view/187)

[3] Değertekin, M., Erol, Ç., Ergene, O., Tokgözoğlu, L., Aksoy, M., Erol, M. K., ... & Kozan, O. (2012). Heart failure prevalence and predictors in Turkey: HAPPY study. Turk Kardiyoloji Dernegi arsivi: Turk Kardiyoloji Derneginin yayin organidir, 40(4), 298-308. Doi: 10.5543/tkda.2012.65031 (https://europepmc.org/article/med/22951845)

[4] Chicco, D., Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16. Doi: 10.1186/s12911-020-1023-5

[5] Osmanoglu, U. O. (2021). Prediction of heart failure mortality by machine learning classification algorithms. Ph.D. thesis, Eskisehir Osmangazi University.

[6] Çelik, Ö., Aslan, A. F., Osmanoğlu, U. Ö., Çetin, N., & Tokar, B. (2020). Estimation of renal scarring in children with lower urinary tract dysfunction by utilizing resampling technique and machine learning algorithms. Journal of Surgery and Medicine, 4(7), 573-577. Doi: 10.28982/josam.691768 (http://jsurgmed.com/en/download/article-file/976184)

[7] Github. (2020). https://github.com/scikit-learn-contrib/imbalanced-learn (Erişim Tarihi: 21/10/2020)

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357. Doi: 10.1613/jair.953 (https://www.jair.org/index.php/jair/article/view/10302/24590)

[9] Python, (2012). General Python FAQ. Python Software Foundation. Licensed material. https://docs.python.org/3/faq/general.html (Erişim Tarihi: 13.12.2014)

[10] Osmanoğlu, U. Ö., Atak, O. N., Çağlar, K., Kayhan, H., & Can, T. C. (2020). Sentiment analysis for distance education course materials: A machine learning approach. Journal of Educational Technology and Online Learning, 3(1), 31-48. Doi: 10.31681/jetol.663733 (https://dergipark.org.tr/en/download/article-file/920037)

[11] Stewart, M. (2019). The Actual Difference Between Statistics and Machine Learning. https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3 (Erişim Tarihi: 11.05.2021)

[12] Lee, D. S., Stitt, A., Austin, P. C., Stukel, T. A., Schull, M. J., Chong, A., ... & Tu, J. V. (2012). Prediction of heart failure mortality in emergent care: a cohort study. Annals of Internal Medicine, 156(11), 767-775. Doi: 10.7326/0003-4819-156-11-201206050-00003 (10.7326/0003-4819-156-11-201206050-00003)

[13] Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology, 66(4), 398-407. Doi: 10.1016/j.jclinepi.2012.11.008

[14] Singh, J. S., Mordi, I. R., Vickneson, K., Fathi, A., Donnan, P. T., Mohan, M., ... & Lang, C. C. (2020). Dapagliflozin versus placebo on left ventricular remodeling in patients with diabetes and heart failure: the REFORM trial. Diabetes Care, 43(6), 1356-1359. Doi: 10.2337/dc19-2187

[15] Asyalı, M. H. (2003, September). Discrimination power of long-term heart rate variability measures. In Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439) (Vol. 1, pp. 200-203). IEEE.

[16] Jovic, A., & Bogunovic, N. (2011). Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. Artificial Intelligence in Medicine, 51(3), 175-186. Doi: 10.1016/j.artmed.2010.09.005

[17] Mortazavi, B. J., Downing, N. S., Bucholz, E. M., Dharmarajan, K., Manhapra, A., Li, S. X., Negahban, S.N., Krumholz, H. M. (2016). Analysis of machine learning techniques for heart failure readmissions. Circulation: Cardiovascular Quality and Outcomes, 9(6), 629-640. Doi:10.1161/CIRCOUTCOMES.116.003039

[18] Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., Zhu, W., Sama, I., Tadel, M., Campagnari, C., Greenberg, B., Yagil, A. (2020). Improving risk prediction in heart failure using machine learning. European journal of heart failure, 22(1), 139-147. Doi: 10.1002/ejhf.1628

[19] Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Information Fusion, 54, 128-144. Doi: 10.1016/j.inffus.2019.07.006