

PREDICTION OF SURVIVAL RATE OF HEART FAILURE PATIENTS USING MACHINE LEARNING TECHNIQUES

NARGIZA FOZILJONOVA¹, ITO WASITO^{2*}

¹Department of Computing, Westminster International University in Tashkent,
12 Istikbol Street, Tashkent 100047, Uzbekistan

²Department of Computing, Westminster International University in Tashkent,
12 Istikbol Street, Tashkent 100047, Uzbekistan

E-mail: ¹n.foziljonova2504@gmail.com, ²wasito@wiut.uz

^{*}) Corresponding Author

ABSTRACT

Heart Failure problem as part of Cardiovascular Disease (CVD) is one of the leading causes of death taking about 17.5 million of people's lives yearly that is amounted about 32% of all deaths in the world. This study is to analyze the leading factors to the heart failure patient problem. The main goal of the paper is to determine the most appropriate machine learning technique for prediction of the survival rate of heart failure of patients. The data in the experiment was taken from a hospital in Pakistan from April to December of 2015. The algorithms were chosen for prediction model are KNN, Random Forest and Decision Tree. The results of the experiments show that Random Forest which produced ROC value 0.956 that outperforms the other machine learning techniques.

Keywords: *Prediction Model, KNN, Decision Tree, Random Forest, Heart diseases.*

1. INTRODUCTION

According to statistics from WHO, Cardiovascular Disease (CVD) is one of the leading causes of death taking about 17.5 million of people's lives yearly that is amounted about 32% of all deaths in the world. CVD is a name for a group of heart diseases, such as heart attack, coronary heart disease, rheumatic heart disease and etc., and among them heart attacks and strokes were the results of more than 80% of deaths caused by CVD (WHO, 2021). Centers for Disease Control and Prevention stay that people in United States have a heart attack in each 40 seconds.

Prediction of CVD from medical point of view is very difficult and challenging task that needs previous health records of the patients. Most researches indicate that main factors that has a big impact on the development of CVD are unhealthy diet, smoking, low physical activity and etc. The development of CVD in turn becomes a leading factor of Heart Failure that happens when the needs of the body cannot be satisfied due to inability of heart to pump enough blood. Considering this

modern life style of most people a system that can predict the possibility of getting heart disease and further impacts of it on patients is needed, since forecasting the illness and survival rate of patients in early stages can save a lot of lives by choosing of appropriate medical treatment.

The possibility of decreasing death rate among CVD patients can be reached by usage of machine learning techniques that can predict survival rate of patients using their electronic medical records that contain information about above mentioned factors that resulted in acquisition of heart diseases.

For the purpose of finding suitable models for prognosis many discussions were held between conventional statistics and machine learning algorithms and based on experimental analysis of Steele et al. (2018), machine learning techniques were found appropriate due to their focus on predictive performance and generalization of models that repeat processes in order to improve an algorithm (Handelman et al., 2018). The main idea behind the usage of machine learning techniques for

prediction is the difficulty of analyzing due to big amount of received data (R. Katayra and S. K. Meena, 2020).

Following the thought of machine learning implementation, many researchers proposed different Classification algorithms for heart disease prediction, such as Naïve Bayes, Decision Tree, K-Nearest Neighbors, Random Forest and etc. However, most of them have their own limitations and the main goal of the paper is to estimate most robust and effective machine learning technique that could help to predict the survival rate of the patients with heart failure. Moreover, the paper aims to analyze the performance of several variables that medical records of the patients contain in order to understand which of them has a big impact on CVD. For that reason, dataset with information about heart failure patients from Pakistan that is presented in Kaggle database was chosen.

The structure of the study as follows: start point is the analysis of previous literature related to the topic of the paper with small introduction to data mining and machine learning and second section describes the obtained data with identification of main exploratory variables. While the third section is about preprocessing of the data for model building, next part of the study contains information about three different machine learning techniques for building predicting model and two methods for evaluating the constrained models, and the last part summarizes results and suggests suitable model approach for the given dataset.

2. LITERATURE REVIEW

A key role in discovery process of knowledge from databases belongs to data mining that is formally described as a discovery of interesting, unexpected, or valuable structures in large datasets (Milanovic and Stamenkovic, 2011). Data mining is widely used in various fields, such as business, scientific research and others, and usually many data mining applications have temporal aspects and most frequent form among them is time series. A time series database can be explained as ordered values or observations that are collected in a specific sequence of time. Generally, despite the format of the temporal data this kind of datasets help to understand historical patterns and analyze their relationship and impact on future events based on the knowledge of the past (Han and Kamber, 2006).

In the modern age of technological development, interest in the time series and sequences data is growing day by day, and this comes from the possibility of using and implementing this type of data in the wide spheres of human life, such as economics, finance, health care, e-commerce, etc. (Esling and Agon, 2012). There has been increasing number of studies of modeling time series data in the community of both machine learning and statistics, since data has a key role in both of the above-mentioned fields of study where the of emphasis is put on discovering knowledge from the data without depending on predetermined equation as a model.

Mapping data into predefined classes is called classification and it is described as a supervised learning method, since it uses training data in constructing the classifier with classes that are predicted in advance. In other words, the algorithm assigns time series data objects to its appropriate class. The main aim of classification is to compare classes and identify what makes them unique from each other (Esling and Carlos, 2012). There are several algorithms that are widely examined for classification of time series data, such as k-nearest neighbors, decision tree, neural network and support vector machine. In this paper three of the above-mentioned algorithms are reviewed.

a) k-Nearest Neighbors (k-NN)

Being a straightforward function on data, k-nearest neighbors method returns an object of the data to the class that is most popular among k – neighbors of that object, classes of which are already defined. The main advantage of k – nearest neighbors' algorithm among others is its simplicity in use and understanding, however, some evidence demonstrate that it suffers from significantly loose of speed with the increase of data (Xi et al., 2006).

b) Decision Tree

One more popular classifier that is identifies different ways of dividing a dataset to branches is defined as Decision Tree (DT), main focus of which is describing the upcoming decisions, cases that are possible to occur, and the results that are part of each event and decisions. It has three nodes: root node, where the dataset can be divided into one or more branches by outgoing branches; internal node, where incoming branch can be divided into two or more outgoing branches; and end node,

where classes are represented by leaf nodes, while decisions represented by branches. At the root node decision to reach the class label is made by classifier (De Ville and Neville, 2013).

c) Random Forest

Random Forest is a method that is proposed to use ensembles of trees, each of which is grown in accordance with a random parameter, in order to achieve regression accuracy (G. Biau, 2012). According to R. Caruana and A. Mizil (2006) Random Forest method is a great statistical learning model that works well with small and medium data.

Considering the case of time series data that is related to health, there has been many deployments of machine learning techniques for prediction in electronic health data (S. Rose, 2018). The main advantage of the combination of clinical data and modern machine learning techniques is a help that they provide to rapidly generate prediction models for a number of similar health questions (J. H. Chen and S. M. Asch, 2018). These include Random Forest, Naïve Bayes, Support Vector Machine, Logistic Regression, and etc., and among them Random Forest is considered as one of the most suitable with higher accuracy score compared to others (R. Katayra and S. K. Meena, 2020; Huang et al., 2021; A. Salazar et al., 2022; S. Aqeel, 2019).

Most researches used accuracy methods, such as ROC and AUC, to find out machine learning techniques that predicts whether people have heart failure or not and according to H. Jindal et al. (2020), KNN, Random Forest and Logistic Regression are more accurate, cost efficient and less time-consuming comparing to other mentioned techniques for predicting heart failure.

Following the idea of previous researches, the paper is going to use such Classification techniques as KNN, Random Forest and Decision Tree in order to estimate the most suitable one based on ROC-AUC scores. However, in comparison with provided broad literature, the research is built in order to analyze the most accurate model among three above-mentioned ones in order to predict the survival rate of people with heart failure disease.

3. MAJOR ANALYTICAL FINDINGS AND DEVELOPMENTS

Data understanding and exploring

The dataset contains medical records of 299 patients with heart failure diagnosis at the hospital in Pakistan in the period of April – December of 2015. It has features to report body, lifestyle and clinical information amounted in 13 variables that are presented in Table 1 below and among them Death event is a target variable:

As it can be seen from the Table 1, all variables have either numerical values, or binary values with 0 and 1. The high level of CPK and Serum creatinine in the blood can indicate heart failure, while the opposite happens with Serum sodium – existence of heart failure can be a reason of low level of this feature (Johns Hopkins Rheumatology, 2019). The Central Tendency Measures of these variables presented in Table 2.

The number of patients is 299 and their age varies between 40 and 95 years and most of them on average is elder than 60. According to age distribution of patients that is provided below in Figure 1, the number of patients in the age of 60 is more than 30, while most patients accounting more than 75 people are in the range of 50 to 70. From this distribution, it can be assumed that people of middle and elder ages are more prone to heart failure rather than younger aged people.

Table 1: Data Explanation

FEATURE	EXPLANATION	VALUE
Age	Age of the patient	Numerical
Anemia	Does a patient have decreased level of hemoglobin?	Binary (0 for “No”, 1 for “Yes”)
High blood pressure	Does a patient have hypertension?	Binary (0 for “No”, 1 for “Yes”)
CPK	Level of CPK enzyme in the blood	Numerical
Diabetes	Does a patient have diabetes?	Binary (0 for “No”, 1 for “Yes”)
Ejection fraction	Percentage of blood leaving	Numeric
Gender	Patient is woman or man	Binary (0 for “Women”, 1 for “Men”)
Platelets	Level of platelets in the blood	Numeric
Serum creatinine	Level of creatinine in the blood	Numeric
Serum sodium	Level of sodium in the blood	Numeric
Smoking	Does a patient smoke?	Binary (0 for “No”, 1 for “Yes”)
Time	Follow-up days	Numeric
Death event	Does a patient die during follow-up period?	Binary (0 for “No”, 1 for “Yes”)

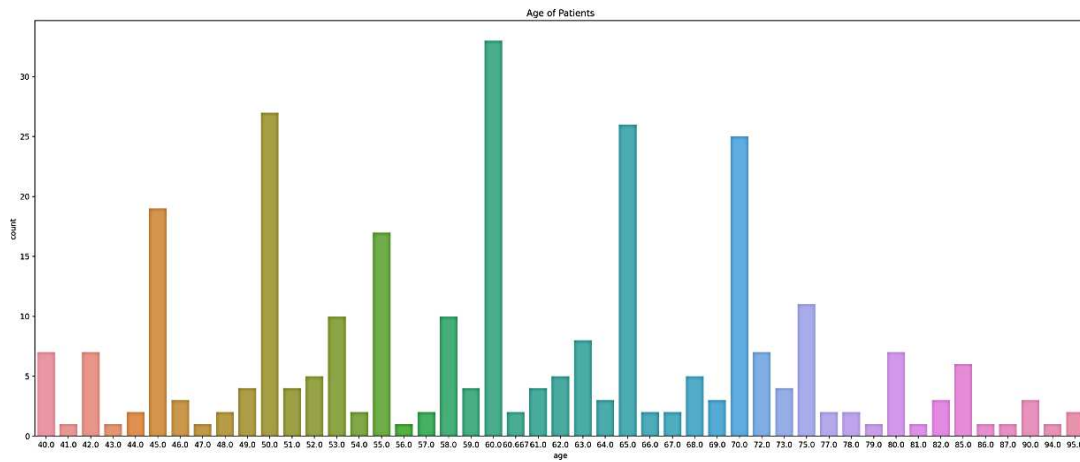


Figure 1. Age Distribution Of Patients

Pie charts describe the proportions of target variable, smoking patients, gender of patients and the ratio of patients with high blood pressure. As it can be observed from them, during follow – up period about 67.9% of patients are survived and the same amount of people had no smoking habit. The gender ratio of men to women is 64.9% to 35.1%, indicating that 105 of patients are women, while 194 of them are men. Exactly the same percentage of 64.9% is distributed for patients with lower blood pressure, meaning that the blood pressure of 105 patients is high and 194 patients have low blood pressure (provided Annex 1).

The distribution that is provided in Figure 3 describes the age of patients and if they could survive or not during follow – up period. According to it, most patients that could not survive were at the age of 60, compared to other age categories. As it was mentioned above, most patients with heart failure were at the age between 50 and 70 and the distribution below shows that the death rate among them is lower compared to survival rate.

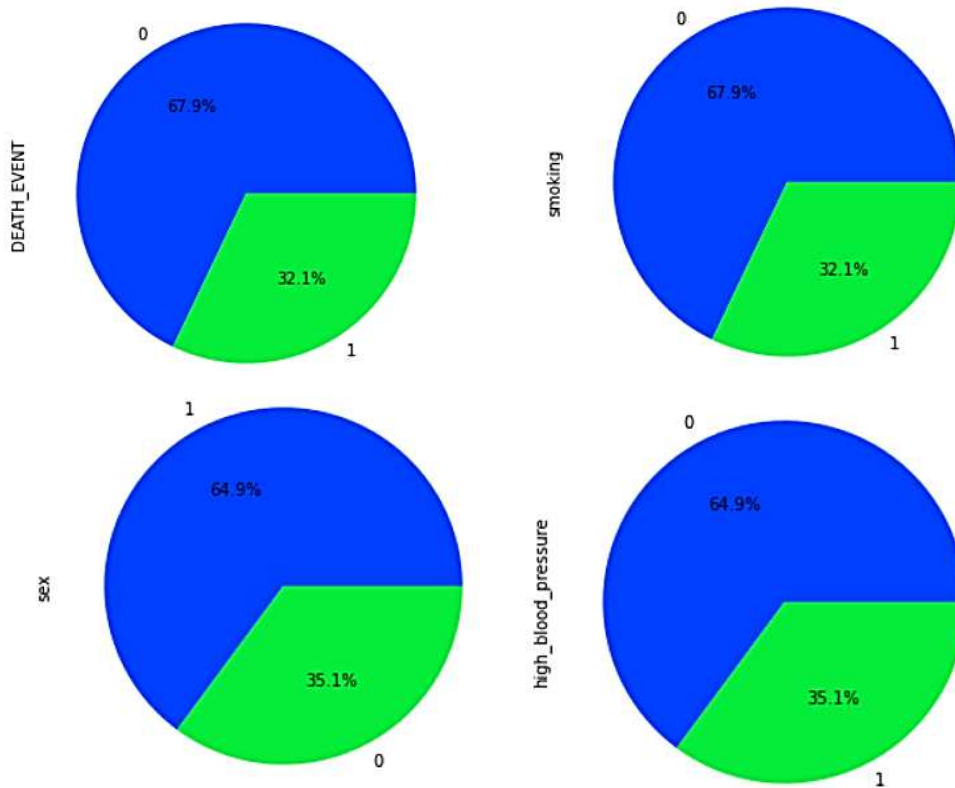
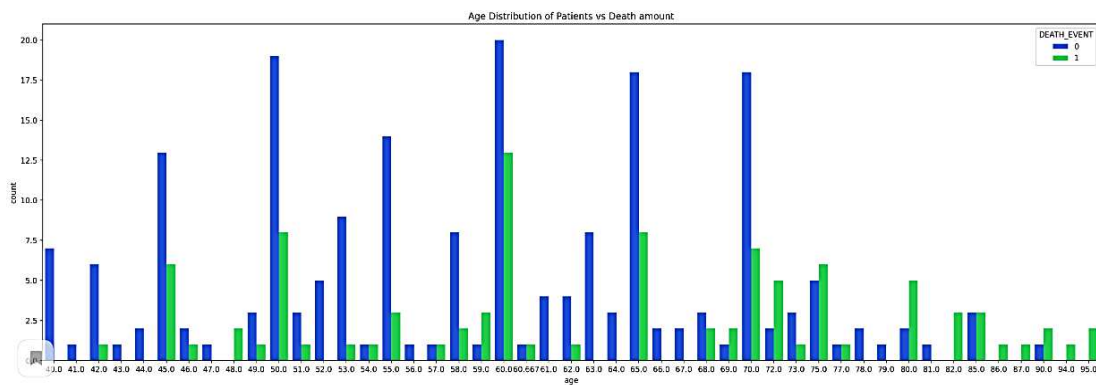


Figure 3. Age Distribution Of Patients' Vs Death Number

Data Preprocessing



The data was preprocessed for missing values and duplicates detection, the results suggest that the dataset has not any missing value and duplicate. Based on the outcomes of detection for unique values, it can be seen that numerical variables that were discussed in Table 1 have several unique

values, while binary features have only two values – 0 or 1.

After the detection for missing values and duplicates, the data was observed for outliers and Table 5 shows 10 outliers in the data. In order to

predict the survival rate of heart failure patients then our main task is to build models using KNN, Random Forest and Decision tree machine learning techniques. Since for these techniques the data should be scaled and normalized, the outliers should be dropped.

For the above-mentioned reason, the features of the dataset should be verified for skewness also and the results of the detection illustrate four variables, such as level of CPK, platelets and serum in the blood, in line with the ejection fraction are skewed. In order to continue the work with them, these features were normalized by the usage of standard deviation and mean values.

After all data preprocessing steps, the correlation matrix was constructed aiming to analyze what variables have any relationship on a target variable or on each other. The results show that death of patients is positively correlated with their age and the level of serum in the blood, while time variable affects target variables in a negative way.

Considering the relationship between features, the level of serum in the blood and age of patients are highly correlated, while the correlation between time and ejection fraction is negative.

TABLE 4: Missing Values & Unique Values

	Missing_Number	Missing_Percent		
DEATH_EVENT	0	0.0	age	47
time	0	0.0	anaemia	2
smoking	0	0.0	creatinine_phosphokinase	208
sex	0	0.0	diabetes	2
serum_sodium	0	0.0	ejection_fraction	17
serum_creatinine	0	0.0	high_blood_pressure	2
platelets	0	0.0	platelets	176
high_blood_pressure	0	0.0	serum_creatinine	40
ejection_fraction	0	0.0	serum_sodium	27
diabetes	0	0.0	sex	2
creatinine_phosphokinase	0	0.0	smoking	2
anaemia	0	0.0	time	148
age	0	0.0	DEATH_EVENT	2
			dtype: int64	

Table 5. Number Of Outliers

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	tim
38	60.0	0	2656	1	30	0	305000.00	2.3	137	1	0	3
52	60.0	0	3964	1	62	0	263358.03	6.8	146	0	0	4
163	50.0	1	2334	1	35	0	75000.00	0.9	142	0	0	12
200	63.0	1	1767	0	45	0	73000.00	0.7	137	1	0	18
296	45.0	0	2060	1	60	0	742000.00	0.8	138	0	0	27
217	54.0	1	427	0	70	1	151000.00	9.0	137	0	0	19
117	85.0	1	102	0	60	0	507000.00	3.2	138	0	0	9
167	59.0	0	66	1	20	0	70000.00	2.4	134	1	0	13
281	70.0	0	582	0	40	0	51000.00	2.7	136	1	1	25
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	

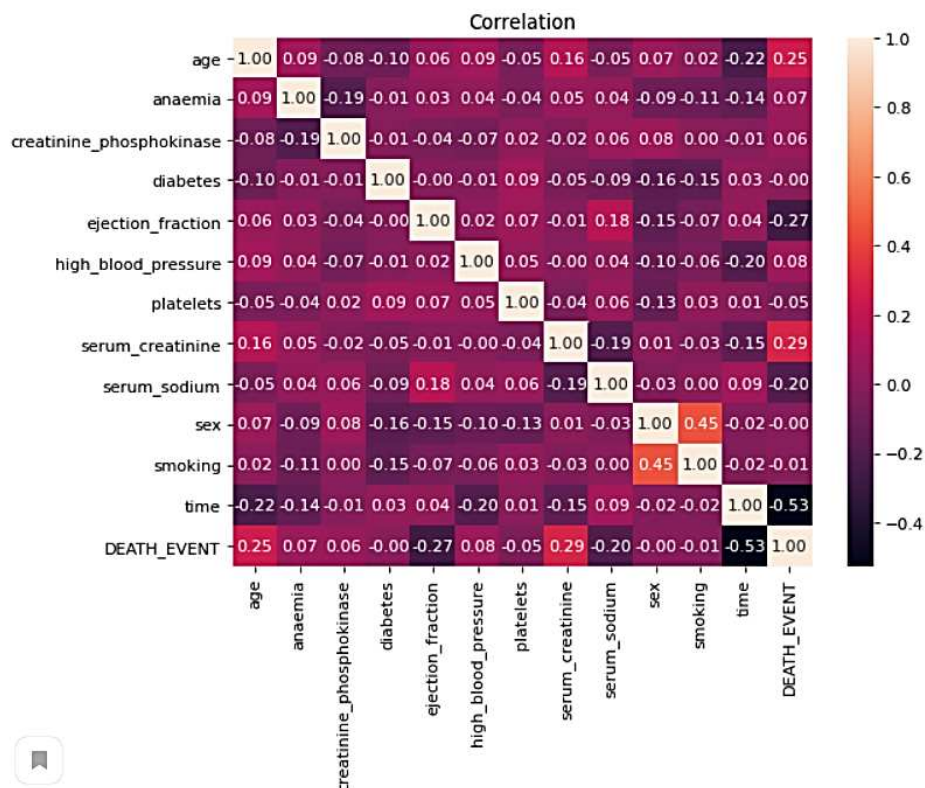


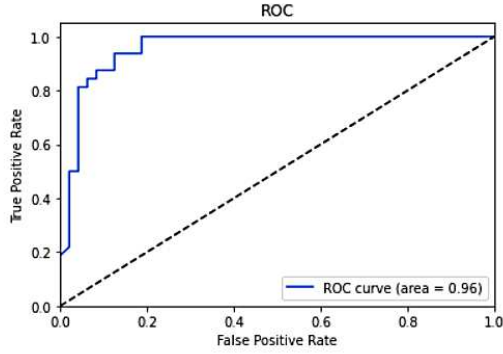
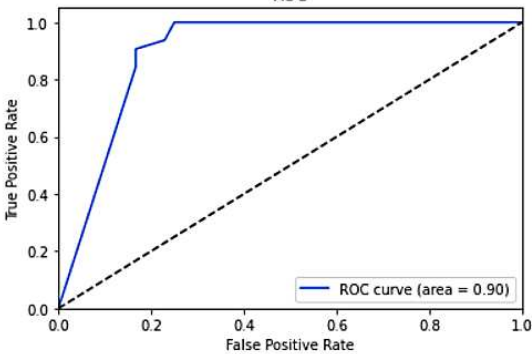
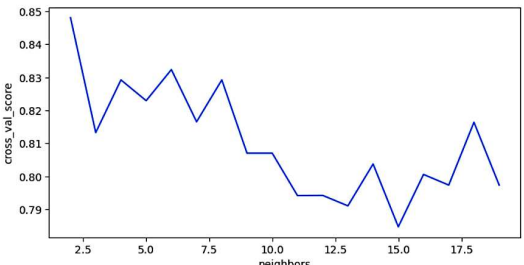
Figure 2. Correlation Matrix

Modelling

As it was stated earlier, machine learning algorithms to predict the survival rate heart failure patients are K – Nearest Neighbor (KNN), Random Forest and Decision Tree. All of three methods deal with classification problem and while both KNN and Decision Tree are considered as supervised machine learning techniques, Random Forest is unsupervised algorithm. For these models, Death Event was chosen as a label feature for classification.

Before the building the model, the dataset was divided into train and test sets with the proportion of 80% to 20%, respectively. Train set helps to prepare models that are going to be used in the analysis and test set is for making new predictions that aimed to use for evaluation of the models’ performance. In constructing Decision Tree and Random Forest models the depth was in range from 1 to 10 that resulted in higher accuracy of train and test sets.

The outputs of the models are provided below and according to them, it can be assumed that Random Forest and Decision Tree are more suited to the dataset rather than K – nearest Neighbors based on the accuracy and average scores.

Model	Score	Graph	
Random Forest	F1 – score and Support		
	Accuracy		0.88 and 80
	Macro average		0.87 and 80
	Weighted average		0.88 and 80
Decision Tree	F1 – score and Support		
	Accuracy		0.86 and 80
	Macro average		0.86 and 80
	Weighted average		0.86 and 80
KNN	F1 – score and Support		
	Accuracy		0.82 and 80
	Macro average		0.81 and 80
	Weighted average		0.82 and 80

For evaluating models, the F1 ROC – AUC scores that are provided in Table 8 below were used and according to them, Random Forest is more appropriate and more suitable machine learning algorithm for this dataset, since both its accuracy score are higher than other models’.

Model	F1 - score	ROC-AUC Score
Random Forest	0.861111	0.956706
Decision Tree	0.840580	0.898112
K – Nearest Neighbor	0.766667	0.861003

4. CONCLUSION

This paper studied the dataset about patients in Pakistan with heart failure disease and focused on the defining more suitable predicting machine learning model with target label of death event feature. The main advantages of implementing machine learning techniques for predicting CVD are its fast speed, better accuracy and cost efficiency.

During analysis it was found that number of patients with CVD is higher among people at the age of 50 – 70 and the mean age in the dataset is 60 years. However, the death rate among these 60 aged patients is higher compared to 50- and 70-years old people and correlation matrix demonstrated that age feature in line with serum level in the blood are highly correlated with death rate, meaning that 1 year and 1 level increase in age and serum level, respectively, will have a positive impact on death rate among patients with heart failure. In other words, it can be concluded that more elder people with heart failure become, chances to prolong the life are getting lower.

To predict the survival rate for the dataset three machine learning techniques were presented – KNN, Random Forest and Decision Tree. The accuracy scores of them are 86.1%, 95.67% and 89.91%, respectively. Estimated experiments show that Random Forest technique is the most accurate technique in prediction of survival rate of heart failure patients.

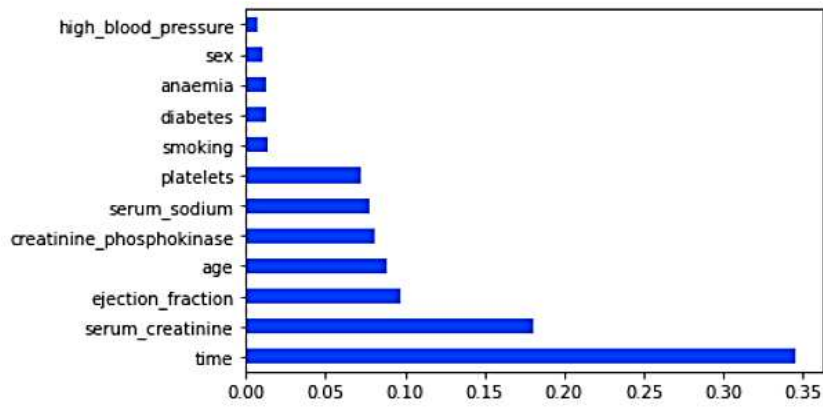
REFERENCE

- [1] Who.int. Cardiovascular diseases (CVDs). 2021 [online] Available at: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)) [Accessed 12 December 2021];
- [2] Centers for Disease Control and Prevention. 2021. Heart Disease Facts | cdc.gov. [online] Available at: <https://www.cdc.gov/heartdisease/facts.htm> [Accessed 6 December 2021];
- [3] Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H. & Luscombe, N. M. (2018). Machine Learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease, PLoS One, 13(8), e0202344;
- [4] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Lee, M. J. & Asadi, H. (2018). eDoctor: machine learning and the future of medicine. J Intern Med, 284(6), 603-619;
- [5] Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021). <https://doi.org/10.1007/s12553-020-00505-7>;
- [6] Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P., 2021. Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 1022(1), p.012072;
- [7] Kaggle.com. Heart Failure Prediction. 2020. [online] Available at: <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data> [Accessed 4 December 2021];
- [8] Milanovic, M. and Stamenkovic, M., 2011. Data mining in time series. 13th ed. [ebook] EKONOMSKI HORIZONTI. Available at: https://www.researchgate.net/publication/317099157_Data_mining_in_time_series [Accessed 4 November 2021];
- [9] Han, J. and Kamber, M., 2006. Data Mining: Concepts and Techniques. 2nd ed. Amsterdam: Elsevier/Morgan Kaufmann, pp.1-68;
- [10] Philippe Esling, Carlos Agon. Time-series data mining. ACM Computing Surveys, Association for Computing Machinery, 2012, 45 (1), pp.12. [ff10.1145/2379776.2379788](https://doi.org/10.1145/2379776.2379788). [ffhal-01577883](https://doi.org/10.1145/2379776.2379788);
- [11] P. Esling and A. Carlos, "Time-series data mining," ACM Comput. Surv., vol. 45, no. 1, p. 12, 2012;
- [12] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast time series classification using numerosity reduction," in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 1033–1040;
- [13] B. De Ville and P. Neville, Decision Trees for Analytics Using SAS Enterprise Miner. Cary, NC, USA: SAS Institute, 2013;
- [14] Biau G. Analysis of a Random Forests Model. Journal of Machine Learning Research [Internet]. 2012 [cited 9 January 2022];13. Available from: <https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf>;
- [15] Caruana R, Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms [Internet]. USA: Department of Computer Science, Cornell University, Ithaca; [cited 17 January 2022]. Available from:

- <https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf>;
- [16] Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open*. 2018;1(4): e181404.;
- [17] Chen J, Asch S. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *New England Journal of Medicine*. 2017;376(26):2507-2509.;
- [18] Mansur Huang N, Ibrahim Z, Mat Diah N. Machine Learning Techniques for Heart Failure Prediction. *MALAYSIAN JOURNAL OF COMPUTING*. 2021;6(2):872.;
- [19] The International Conference on Advanced Machine Learning Technologies and Applications (AMLT2019). *Advances in Intelligent Systems and Computing*, 2020;
- [20] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* 20, 16 (2020); <https://doi.org/10.1186/s12911-020-1023-5>;
- [21] Johns Hopkins Rheumatology. Creatine Phosphokinase (CPK). <https://www.hopkinslupus.org/lupus-tests/clinical-tests/creatinine-phosphokinase-cpk/>. Accessed 25 Jan 2019;
- [22] O. Simeone, “A very brief introduction to machine learning with applications to communication systems,” *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 648–664, Dec. 2018.

ANNEX

Random Forest



Decision tree

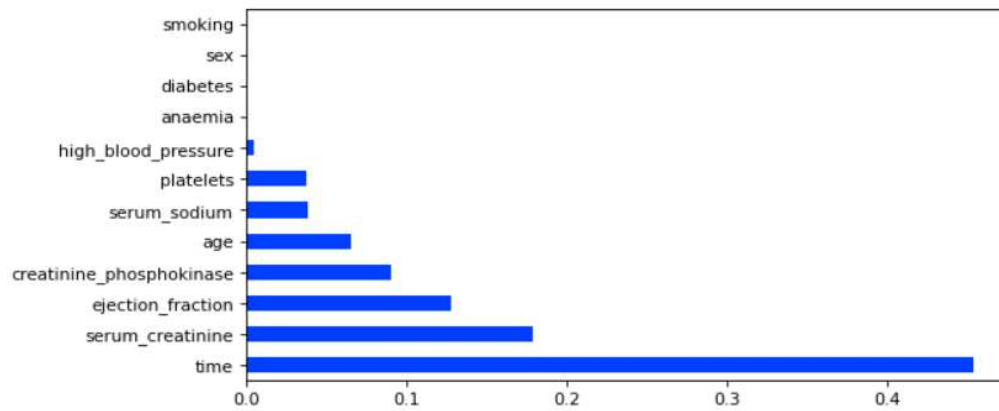


TABLE 2. CTM of variables

	count	mean	std	min	25%	50%	75%	max
age	299.0	60.833893	11.894809	40.0	51.0	60.0	70.0	95.0
anaemia	299.0	0.431438	0.496107	0.0	0.0	0.0	1.0	1.0
creatinine_phosphokinase	299.0	581.839465	970.287881	23.0	116.5	250.0	582.0	7861.0
diabetes	299.0	0.418060	0.494067	0.0	0.0	0.0	1.0	1.0
ejection_fraction	299.0	38.083612	11.834841	14.0	30.0	38.0	45.0	80.0
high_blood_pressure	299.0	0.351171	0.478136	0.0	0.0	0.0	1.0	1.0
platelets	299.0	263358.029264	97804.236869	25100.0	212500.0	262000.0	303500.0	850000.0
serum_creatinine	299.0	1.393880	1.034510	0.5	0.9	1.1	1.4	9.4
serum_sodium	299.0	136.625418	4.412477	113.0	134.0	137.0	140.0	148.0
sex	299.0	0.648829	0.478136	0.0	0.0	1.0	1.0	1.0
smoking	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0
time	299.0	130.260870	77.614208	4.0	73.0	115.0	203.0	285.0
DEATH_EVENT	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0

Table 6:

Before

After

	Skewed Values
creatinine_phosphokinase	4.827396
serum_creatinine	4.605615
platelets	1.429547
DEATH_EVENT	0.797132
smoking	0.729243
high_blood_pressure	0.583959
ejection_fraction	0.546146
age	0.413840
diabetes	0.358569
anaemia	0.286636
time	0.125442
sex	-0.663509
serum_sodium	-0.855445

	Skewed Values
DEATH_EVENT	0.797132
smoking	0.729243
high_blood_pressure	0.583959
age	0.413840
diabetes	0.358569
anaemia	0.286636
platelets	0.153154
time	0.125442
creatinine_phosphokinase	0.038332
serum_creatinine	-0.005096
ejection_fraction	-0.005676
sex	-0.663509
serum_sodium	-0.855445