Predicting Titanic Survivors by Using Machine Learning

Yufan Ai *

College of Engineering, Michigan StateUniversity, East Lansing, USA

* Corresponding Author Email: aiyufan@msu.edu

Abstract. About a century ago, one memorable night in April 1912, a world-shattering event happened. The Titanic, the 2,240-passenger luxury cruise ship, sank forever off the coast of Newfoundland in the North Atlantic after extensive damage to its hull by an iceberg on its maiden voyage. Only 705 people survived this disaster. Although nearly a century has passed, the research on Titanic has never stopped, and there are still many studies on it. This study was supposed to predict the survival of passengers on Titanic using different methods based on data from the Kaggle competition "Titanic: Machine Learning from Disaster." It predicted each passenger in the test set who would survive the sinking. The result was the percentage of correct prediction. In the Machine Learning study, the task is to achieve 80% accuracy in predicting the survival distribution of the Titanic disaster based on the demographic data testing notebook by different algorithms models. Using classification is the main point to calculate the efficiency achieved by those models through the test environment. The f-measurement scores obtained from the machine learning technology were in comparison with the f-measurement scores obtained by Kaggle.

Keywords: Machine Learning, Kaggle, Method, Data Analysis, Model EvaluationIntroduction.

1. Introduction

On April 14, 1912, at 11 p.m., the Titanic experienced a horrific disaster, and nearly full of its parts got damaged by the catastrophe [1]. Sadly, there were not enough lifeboats to rescue all 2,224 passengers on board [2]. By introducing survivors and officially collected data, such as ferry tickets and corresponding numbers of people records, it can be known that in this disaster, the Age and gender distribution of the dead were uneven [3]. Data research on the disaster has continued to this day. The approach is a death distribution budget based on public data on Kaggle. Using different models will calculate the efficiency achieved through the test environment provided by Kaggle [2].

The requirement for this research is to build a series of machine learning models with an f-measurement accuracy greater than 80 percent on a given set of demographic information on Kaggle. Predictive modeling is an essential component of predictive analysis [4]. Through the mathematical process, the input data can predict future results. The different attributes of people, such as gender, Age, the category they belong to, and their social class, build the predictive models by providing a good data analysis database [5]. In machine learning, the data has two parts, one is for training, and the other is for testing. The training models are used to fit machine learning models, and the testing models are used to evaluate equipped machine learning models. The train-test split procedure estimates the performance of a machine learning algorithm when it makes predictions based on data that cannot use to train the model [6]. To reach the aim of this research project, the use the construction of Scikit-Learn and related machine learning algorithms to perform data analysis, build models and evaluate algorithm performance on data used to train models in Kaggle [2].

2. Methodology

2.1. Data

There are two groups in the historical data, one is the training set, and the other is the testing set. For the training set, using the testing dataset, which Kaggle provides, can determine whether the passenger survived to build the model for generating prediction patterns.

Loading data from a package requires importing the data into a library and executing instructions. The original library covers a lot of related data, so there is little data to show. After loading the test data content, the study can approach categorizing and recording the data by getting the first ten rows of the dataset to [7]. Table 1 shows the training dataset.

Table 1. Training dataset

	Passengers	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund,Mr.Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cummings,Mrs.JohnBradley(Florence Briggs)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen,Miss.Laina	female	26.0	0	0	STON/O2.3101282	7.9250	NaN	S
3	4	1	1	Futrelle,Mrs.Jacques Heath(Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen,Mrs.William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran,Mr.James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy,Mr.Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson,Master.Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson,Mrs.Oscar	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser,Mrs.Nichola	female	14.0	1	0	237736	30.0708	NaN	C

It is helpful to record each row of passengers on board after receiving the dataset's count of rows and columns. On the other hand, these columns are considered the distribution data for each commuter. Theoretically, the dataset has 891 passengers in the rows and 15 data points in the columns [8]. Mean value, count, standard deviation, and other statistics can also display on the dataset. Subsequently, passengers are classified and analyzed in different categories.

2.2. Section snippets

For this part, an essential process of building machine learning can be implemented, as Figure 1 shown, to fit the collected training data into a data model that can be tested. Under this data model, more accurate data will be determined to compare and analyze the obtained results using the different models. These are essential steps that need to be performed in performing classification and regression analysis, which are summarized in Figure 1.

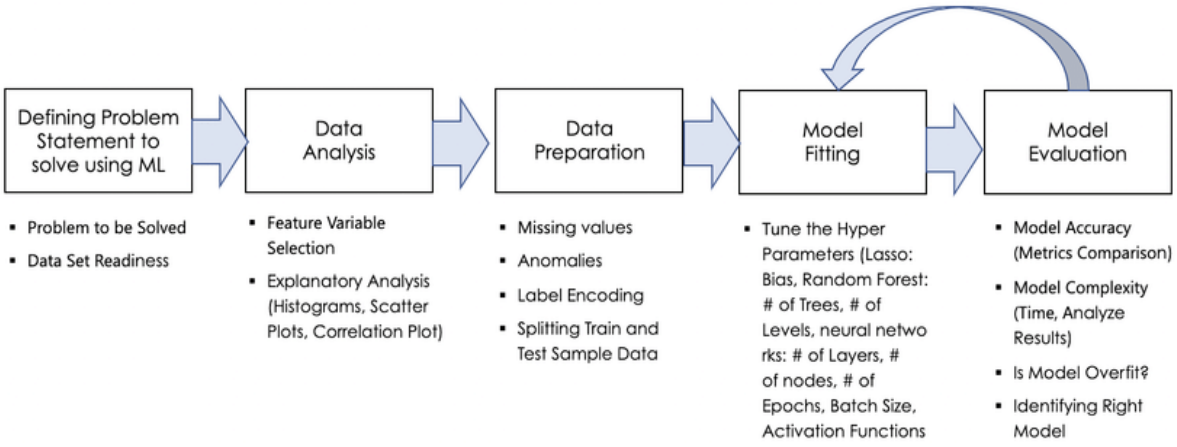


Figure 1. The process of ML

2.3. Data cleaning

The purpose of data cleaning is to prevent similar errors in data collection. Data cleanup can be divided into these steps identifying the mistakes and either correcting or deleting the data that needs to be modified or manually processing the data, which is an essential part of machine learning to build models. Correct data deletion can save us time and cost and improve data efficiency, and the previous dataset after cleaning is summarized in Table.2.

Table 2. dataset after cleaning

	Survived	Pclass	Sex	Age	Fare	Embarked	IsAlone	Title
0	0	3	1	1	0.0	0	0	1
1	1	1	0	2	3.0	1	0	3
2	1	3	0	1	1.0	0	1	2
3	1	1	0	2	3.0	0	0	3
4	0	3	1	2	1.0	0	1	1
5	0	3	1	1	1.0	2	1	1
6	0	1	1	3	3.0	0	1	1
7	0	3	1	0	2.0	0	0	4
8	1	3	0	1	1.0	0	0	3
9	1	2	0	0	2.0	1	0	3

2.4. Titanic prediction—artificial neural network

An ANN neural network is an extensive parallel interconnected network of simple adaptive units whose organization can simulate the interactive responses of biological nervous systems to real-world objects [9]. The ANN model initially learns to recognize patterns in the data, visual or auditory, through a training phase where it compares the actual output with the original intent to produce the desired outcome.

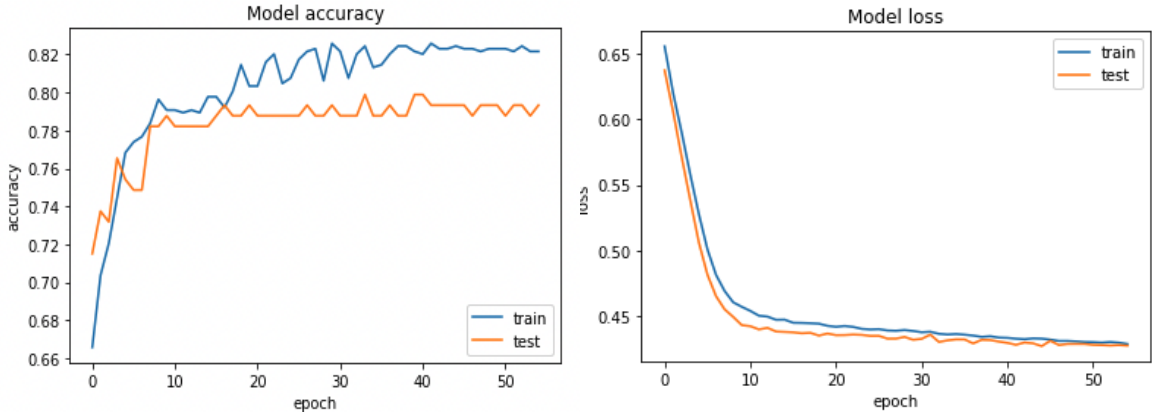


Figure 2. ANN validation

As shown in Figure 2, the ANN model can predict Titanic survivors with an accuracy of 82.16% and an error of 0.45. In addition, the model shows that the factors that impact Titanic survivors most are gender, P-class, and cabin and the different variable names, as shown in Table.3 [7].

Table 3. Different variable names

Variable	Definition	key
Survival	Survival	0=No,1=Yes
Pclass	Ticket class	1=1st,2=2nd,3=3rd
Sex	Sex	Male and Female
Age	Age in years	-
Sibsp	#of siblings/spouses aboard the Titanic	-
Parch	#of Parents/Children aboard the Titanic	-
Ticket	Ticket number	-
Fare	Passenger fare	-
Cabin	Cabin number	-
Embarked	Port of Embarkation	C=Cherbourg,Q=Queenstown,S=Southampton

2.5. Exploratory data analysis

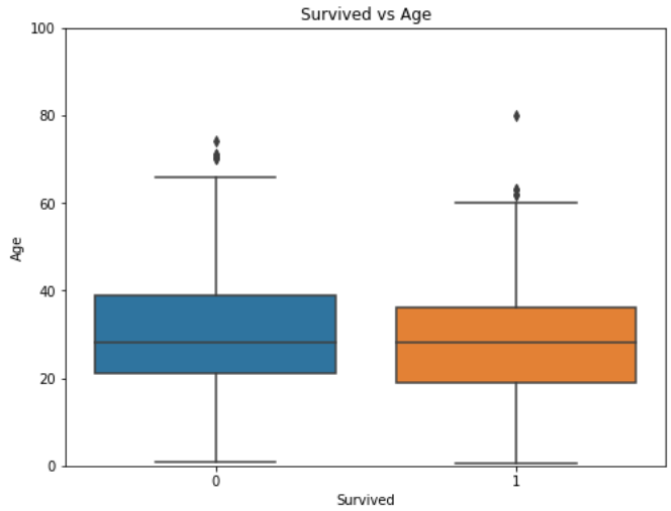


Figure 3. Survived vs. Age

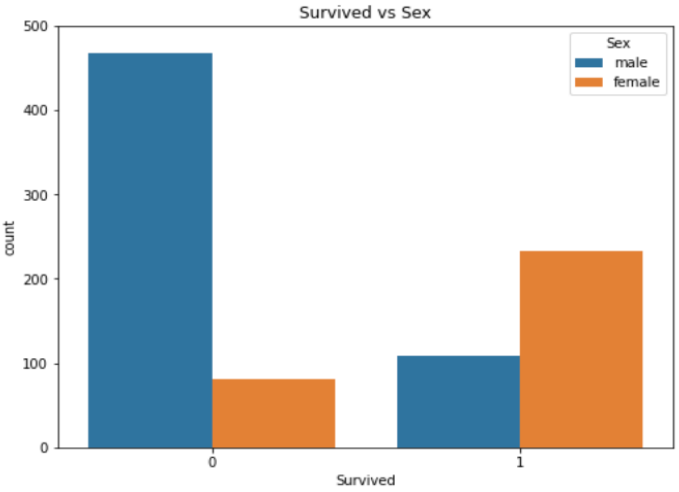


Figure 4. Survived vs. Sex

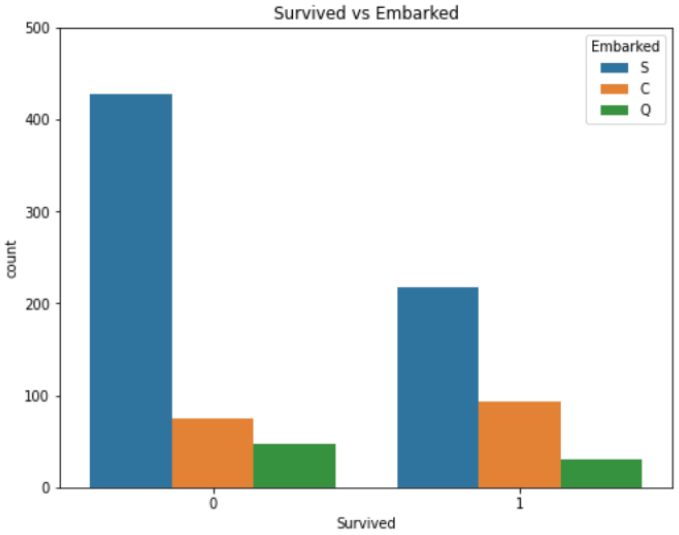


Figure 5. Survived vs. Embarked

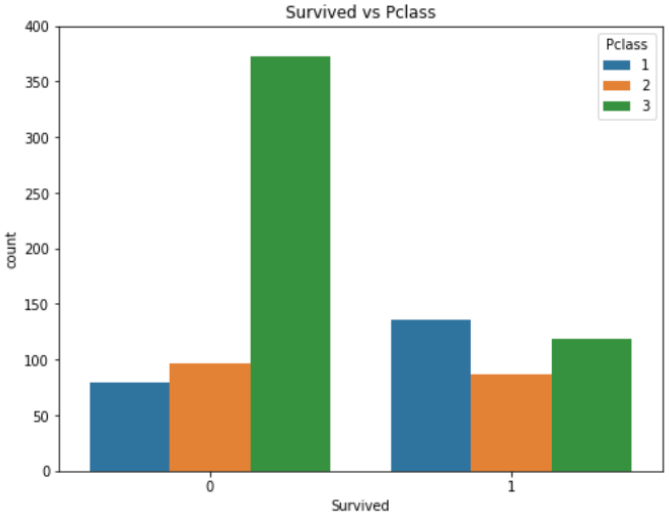


Figure 6. Survived vs.P-class

The exploratory data analysis method is a method that emphasizes data visualization, focusing on the proper distribution of data [9]. The test data is imported here to obtain the survival distribution and various features. The observed rules are analyzed by insights to get inspiration to find a model suitable for the data better. Here, Age, gender, embarked, and P-classcan be seen in Figure 3, Figure 4, Figure 5, and Figure 6, as important survival research features, becoming the most effective data indicators that can be speculated in this study.

2.6. Insights

Insights are the process of finding relations or trends considering multiple features. After the Titanic dataset had been analyzed, two obtained results could be determined from this time. The first is to see what the survival passengers have in common that will help them survive the shipwreck, and the second is to know if the passengers boarded this fateful ship by applying machine learning tools. For Age, a few points are located outside the whiskers of the box plot. However, since it is still within a reasonable age range (around 62 - 80), it will keep the data to see what can be determined from this section. Females were more likely to survive than Males. Upper-class passengers (P Class = 1) were more likely to survive the crash Passengers embarked from S were less likely to survive. (Possibility of a correlation between P-class and embarked)

3. Machine learning models

3.1. Support vector machines

A support vector machine (SVM) is the basic model of the system, which is a linear classifier with the most considerable interval defined in the feature space. It is a generalized linear machine that supervises binary classification. It solves the classification problem of two groups of categories by using classification algorithms. After providing a set of SVM model sets marked with training data for each class, they can classify the new text.

From the results of EDA, the SVM method of this project is still a problem to be improved. Before EDA, it was concluded that the survival rate of women and the upper class was higher than other classes. Therefore, eliminate unnecessary data to retrain the SVM model. Finally, 82.82% of the results can be determined by this improvement.

3.2. K-nearest neighbors

The k-nearest neighbor algorithm is a data classification method often used in machine learning. It compares the groups to which the nearest data points belong in the feature space based on their distance and estimates the possibility that the data points become members of one group or another.

The test environment divides the data into vertical and horizontal axes. Through this model construction, the result is 79.79%.

3.3. Logistic regression

Logical regression can be regarded as the process of modeling the probability of discrete results under the condition of given input variables. This study uses the most common binary logistic regression model, which can take a dependent variable with two possible values for analysis and comparison to obtain an answer relatively close to the result. The accuracy of the model data in the test is 79.12%

3.4. Principal component analysis

PCA is a multivariate statistical method for observing correlations among multiple variables. It is suitable for reducing data without losing its properties. This method can describe the composition of variance and covariance for several linear combinations of the main variables and retain essential parts of the original information that may be lost [10].

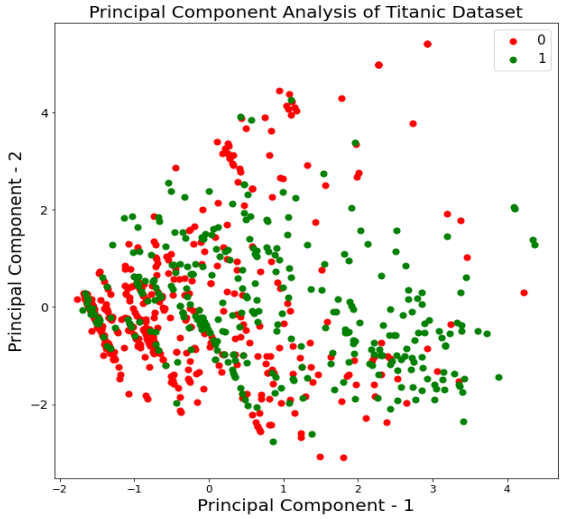


Figure 7. PCA Dataset

The PCA method converts feature with high correlation and equivalent information into a single part. The correlation matrix gives a high correlation of 54%,as shown in Figure 7, between Fare and P-class, also closely related depending on the question. Check the correlation between Embarked and P-class from PCA, as shown in Figure 8. The results of this problem can be replaced by other methods, proving that in this learning process, PCA is unsuitable for this Experimental modeling.

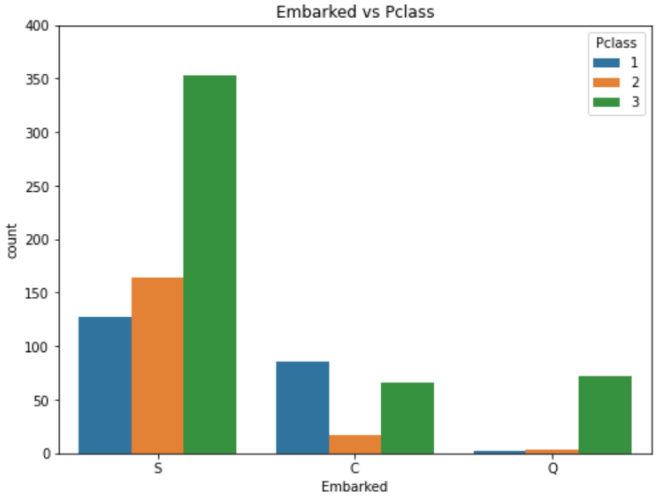


Figure 8. Checks the correlation between Embarked & P-class from PCA.

3.5. Featured engineering

The featured engineering can be a core site to predict the result by changing data to create new variables. To select features that are implemented during training and, therefore, during prediction, domain awareness can be used in feature engineering to acknowledge data set aspects that help create machine learning models. Modeling helps explain data collection, and incorrect feature selection may lead to inaccurate and shocking prediction models. Reliability and predictability depend on the precise nature of the workpiece. It filters out unused and unimportant functions. The following characteristics of the exploratory study were completed using a person's Age, gender, title, class number, P-class, boarding place, ticket price, and several families. The survival column was selected as the reaction column. These characteristics were chosen because of their ideal impact on survival. If incorrect parts are selected, inaccurate predictions may occur. Therefore, functional engineering is the backbone of creating effective forecasting models.

As seen below, some data was lost during the test, and some columns have missing values. Display missing function shows the count of missing values in every column in the training and test set.

The "Age," "Cabin," and "Boarding" columns of the training set are missing values, while the "Age," "Cabin," and "Fare" columns of the test set are not marked with relevant matters [11]. This process requires feature engineering to show some unnecessary columns to improve this problem. Then it would help to view each column's value names and counts to make a more accurate chart. Nonredundant rows and columns and their rows and missing values should be deleted for data disassembly to improve data accuracy.

4. Result

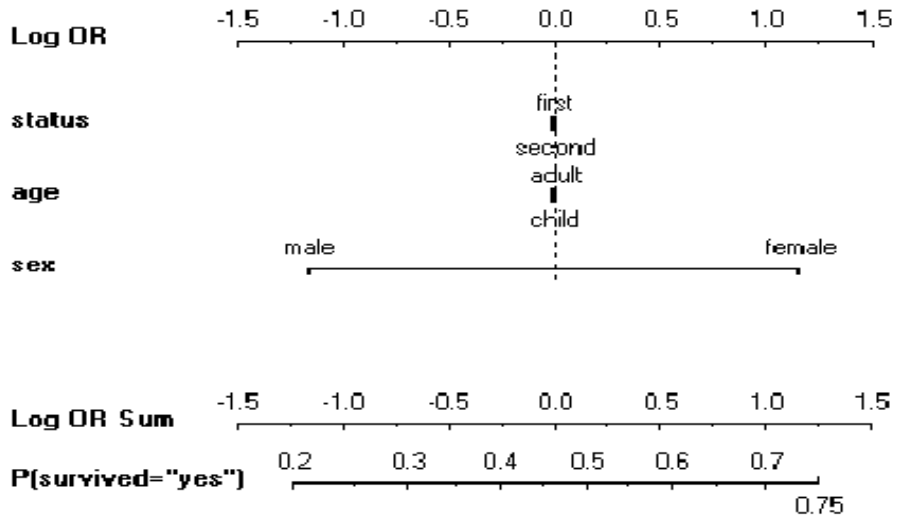


Figure 9. Classification of survivals

The relevant data can be obtained through a series of model building and analysis in insights, as shown in Figure 9. After classifying different methods and algorithms, finding the best implementation score to recognize objects and separate them into categories is one of the processes. As a result, this study gained valuable experience building prediction systems from the SVM model and achieved the best score on Kaggle, 82.82% of correct predictions.

5. Conclusion

Extensive hyperparameters should be adjusted on multiple machine learning models to develop overall results. It can be further enhanced through comprehensive learning. This paper starts with data exploration and then examines the valid data through analysis to see which features are essential to the model. When the data preprocessing sections arrive, classifications are used to calculate the exact

values of each model and convert them into digital visual elements. Then, by creating some functions, different machine learning models are trained. Finally, the confusion matrix, f-score, accuracy of computer models, and recall rates were investigated and evaluated. Data cleanup is the first and most critical step when trying to analyze data. Exploratory data analysis allows people to identify datasets and associated link characteristics. EDA is implemented to find relationships between data set elements. This is achieved by using different graphics techniques. Using EDA to draw some conclusions and discover facts is an integral part of this study. Optimizing the efficiency and accuracy of the model is also a critical process in the experimental process when learning about data recycling using valid models.

References

[1] The National Archives, J. C. 2012. Titanic, 100 years on. https://blog.nationalarchives.gov.uk/titanic-100-years-on/

[2] Kaggle. 2020, Aug 1. Titanic - machine learning from disaster. https://www.kaggle.com/c/titanic/data

[3] Cronan, J. 2012. *National Archives*.https://blog.nationalarchives.gov.uk/titanic-100-years-on/

[4] Lord, W., & Philbrick, N. 2017. I Believe She's Gone, Hardy. In A night to remember (pp. 47–50).

[5] Mishra, V. P., Singh, B., Shukla, V. K., & Dasgupta, A. 2021. Predicting the likelihoodof survival of Titanic's passengers by Machine Learning. https://www.researchgate.net/publication/351155499_Predicting_the_Likelihood_of_Survival_of_Titanic%27s_Passengers_by_Machine_Learning

[6] Brownlee, J. 2020. Evaluating Machine Learning Algorithms.https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

[7] ABDILLAH, F. R. 2021. Titanic Prediction - ANN. https://www.kaggle.com/code/farizhaykal/titanic-prediction-ann

[8] Cook, A. 2019. Titanic - Machine Learning from Disaster. https://www.kaggle.com/code/alexisbcook/titanic-tutorial

[9] Swain, A. 2017. EDA To Prediction.https://www.kaggle.com/code/ash316/eda-to-prediction-dietanic

[10] Houssem Ben Braieka, Foutse Khomha. 2018. *On Testing Machine Learning Programs.* Montreal: SWAT lab. https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab

[11] Niklas Donges, 2018, Predicting the Survival of Titanic Passengers. https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8