Analysis of Titanic Disaster Dataset using Pandas and Machine Learning Libraries

Author: Mohamed Zidan © 2022

Abstract

While the Titanic disaster occurred over one century ago, it still attracts researchers interested in understanding why some passengers survived while others died. With advanced data analysis and machine learning libraries, such as Pandas and Scikit-learn, a publicly available dataset from the Kaggle website, with 891 training examples, 12 features, and 418 testing examples, is analyzed to determine which features of the dataset were more likely to contribute to the survival of the passengers and who is more likely to survive than others. Also, several machine learning-based classification models will be built to predict which passengers survived the Titanic shipwreck and achieve higher accuracy results than previous works. These models include k-nearest neighbors, Naive Bayes, linear support vector machine, Gaussian support vector machine, discriminant analysis, decision trees, random forests, logistic regression, and XGBoost.

1 Introduction

The Royal Mail Ship (RMS) Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours on April 15, 1912 after colliding with an iceberg during its maiden voyage from Southampton, the U.K, to New York City, the U.S. This disaster resulted in killing 1502 out of 2224 passengers and ship personnel (1) (2).

Some authors have studied this disaster using different approaches. For example, the authors in (8) conducted a comparative analysis between three machine learning classifiers, Naive Bayes, support vector machine, and decision trees, with decision trees achieving the highest accuracy score of 79.43% on the test dataset and with a small gap of 2.64% with the worst model, Naive Bayes. Also, the paper in (9) performed cluster analysis and decision tree algorithms to determine the chances of passengers surviving the Titanic disaster. The authors found that the "Gender" feature was the most significant factor impacting survival rate and that the J48 decision tree classifier achieved an 81% accuracy score on the test dataset. Furthermore, the author in (10) used decision trees and random forests classifiers on selected features of the Titanic dataset. The author achieved an accuracy score of 78.46% on the test dataset for their best classifier, the decision trees. Finally, the paper in (11) applied logistic regression, Naive Bayes, decision trees, and random forests, to predict the survival rate of passengers. The applied logistic regression achieved the highest accuracy of 94.26% on the test dataset with a small gap of 2.96% with the worst model, Naive Bayes.

The project's approach will differ from previous works. First, the project will perform different ways of data processing and transformation, including outlier detection and removal and feature engineering, followed by exploratory data analysis to identify important factors impacting the survival rates of passengers. Second, the project will utilize several machine learning classifiers, such as k-nearest neighbors, Naive Bayes, linear support vector machine, gaussian support vector machine, discriminant analysis, decision trees, random forests, logistic regression, and XGBoost. Third, these classifiers will be optimized using two hyperparameter tuning algorithms to achieve accuracy scores greater than 95% on the test dataset. The project will utilize several publicly available libraries for data analysis, data visualization, and machine learning:

• Pandas: Pandas is a powerful and versatile Python data analysis library built on top of two Python libraries: NumPy for mathematical operations and matplotlib for data visualization. Pandas act as a wrapper over these libraries, allowing users to invoke matplotlib's and NumPy's methods. One effective use case of Pandas is for expediting the preprocessing tasks of machine learning projects (4).

Table 1: Features for the Kaggle Titanic Dataset

Features	Description	Keys
PassengerId	An unique index for passenger rows	0 through 890
Survived	A boolean value to show if the passenger survived or not	0 =No, 1= Yes
Pclass	Passengers ticket class cccc	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Passenger's name	
Sex	Passenger's sex. It is either male or female	Male, Female
Age	Passenger's age in years	
Sibsp	Number of siblings/spouses aboard	
Parch	Number of parents/children aboard	
Embarked	Port of Embarkation	
Ticket	Passenger ticket number	
Fare	Passenger fare	
Cabin	Passenger cabin number	
Embarked	Port of Embarkation	C = Cherbourg Q = Queenstown S = Southampton

• Scikit-learn: another powerful Python library for machine learning projects built on top of two Python libraries: NumPy and SciPy for optimization, statistics, and signal processing. In addition, Scikit-learn has a collection of supervised and unsupervised learning algorithms that can be applied in engineering problems, including classification, regression, and clustering problems (5).

This report is structured as follows: Section 2 provides an overview of the dataset and methodology. In Section 3, exploratory data analysis results are presented. Section 4 discusses the implementation steps and results of several machine learning classification models, followed by Section 5, concluding the report and discussing limitations and future work.

2 DATASET AND METHODOLOGY

2.1 Dataset description

The dataset consists of 891 training examples with 12 features and 471 testing examples, as provided by Kaggle (1) (3). Table 1 above summarizes the 12 features of the dataset.

Also, Table 2 and Table 3 below represent a sample of the train dataset.

2.2 Dataset questions

The following questions attempt to determine what factors made some passengers more likely to survive than others:

• Question 1: Did gender determine chances of survival for passengers?

• Question 2: Which age range were the more likely to survive? did age, regardless of gender, determine chances of survival for passengers?

• Question 3: Was social-economic status a factor in survival rate?

• Question 4: How many passengers embarked from each port? Was port of embarkation an indicator of survival rate for passengers?

• Question 5: Was family size of passengers a factor in survival rate?

2.3 Dataset pre-processing

The dataset was processed by encoding categorical data, imputing missing feature values, engineering new features from some of the existing features, and eliminating unnecessary features, thus facilitating data analysis and machine learning classification tasks. Table 4 below identifies new modifications to the dataset.

Also, Table 5 below presents a sample of the final processed dataset for analysis.

3 EXPLARTORY DATA ANALYSIS RESULTS

For Question 1, it was observed that there was a strong correlation between the "Sex" and "Survived" features, as shown in Figure 1 below.

It was also found that females tend to have higher survival rates than males (68.2% for females compared to 31.8% for males). Among the females, the percentage of survived females was 74.2%.

Table 2: Kaggle Titanic disaster sample train dataset

PassengerId	Survived	Pcalss	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
3	1	3	Heikkinen, Miss. Laina	female	26

Table 3: Kaggle Titanic disaster sample train dataset (Continued)

SlibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	A/5 21171	7.25		S
1	0	PC 17599	71.2833	C85	C
0	0	STON/O2.3101282	7.925		S

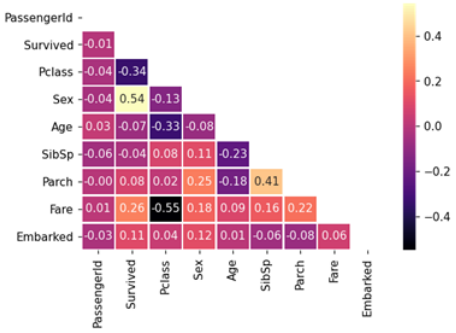

Figure 1: Strong correlation between the "Sex" and "Survived" features

Regarding Question 2, Figure 2 below shows a violin plot displaying the distribution of the dataset and its probability density for both sexes who survived across a wide age range.

Figure 2: Survival rate versus Age and Sex

It can be noticed that the survival rate is good for children, high for women in the age range 20-40, and less for men in the same range of 20-40. Also, for both sexes, the survival rate decreases as the age passes the 40-age mark. Therefore, it can be concluded that the younger passengers were the more likely to survive.

For Question 3, It was observed that the survival rate depends on the ticket fare, which is closely related to the Passenger Pclass. Those passengers with higher Pclass paid higher ticket fares, and therefore had higher chances of survival than those with lower Pclass, as demonstrated in Figure 3 and Figure 4 below. Please note that Pclass 1 is considered the highest class.

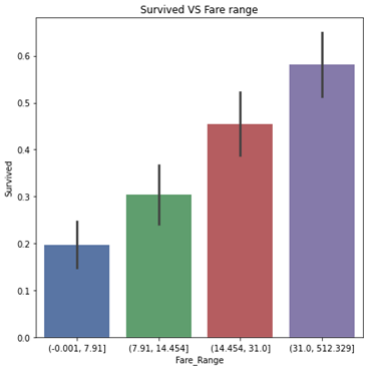

Figure 3: Survival rate versus fare range

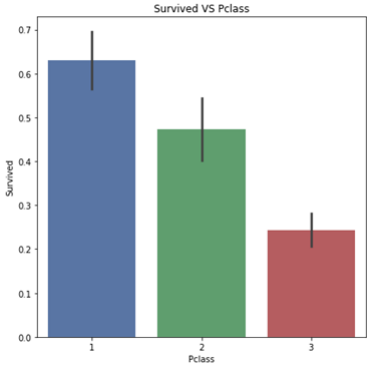

Figure 4: Survival rate versus Pclass

Therefore, it can be concluded that social-economic status was a factor in the survival rate of passengers.

Table 4: Modifications to the dataset

Feature	Modification	Comment
PassengerId	Ignored	Unnecessary
Survived	Unchanged	
Pclass	Unchanged	
Name	Ignored	Unnecessary
Sex	Categorically encoded	0.0 for male 1.0 for female
Age	Missing values imputed	Imputation using the mean
SibSp	Unchanged	
Parch	Unchanged	
Ticket	Ignored	Unnecessary
Fare	Unchanged	
Cabin	Ignored	Unnecessary
Embarked	Categorically encoded	0.0 for S 1.0 for C 2.0 for Q
$\text{Fare}_{R}ange$	New column	Added for survival data analysis
$\text{Family}_{S}ize$	New column	Added for survival data analysis
Alone	New column	Added for survival data analysis

Table 5: Final processed sample dataset for analysis

PassengerId	Survived	Pcalss	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	0.0	22.0
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	1.0	38.0
3	1	3	Heikkinen, Miss. Laina	1.0	26.0

Concerning Question 4, the percentage of passengers who boarded from Southampton was the highest of 72.5%, followed by 18.8% for Cherbourg and 8.6% for port Queenstown. Figure 5 below illustrates the result.

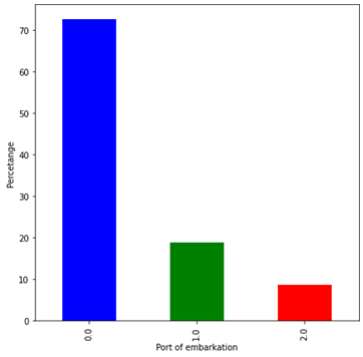

Figure 5: Percentage of passengers who boarded each port

Also, passengers who boarded in Cherbourg had the highest survival rate of 55%. Passengers who boarded in Southhampton were slightly less likely to survive than those who boarded in Queenstown with survival rates of 34% and 39%, respectively, as shown in Figure 6.

Furthermore, it was also found that males and

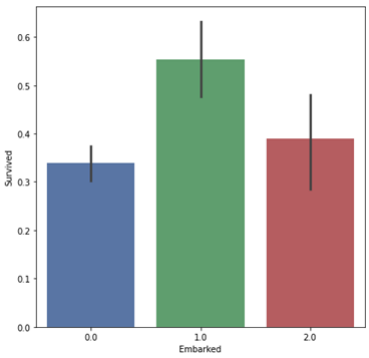

Figure 6: Survival rate versus Embarked

females were more likely to survive from the Cherbourg port, as shown in Figure 7 below.

Finally, for Question 5, it can be concluded that the survival rate is far less if a passenger is alone. Also, if the family size is greater than 3, the chances of survival decrease considerably. Figure 8 and Figure 9 below show the results.

4 MACHINE LEARNING MODELS

4.1 Further data processing

After the initial data processing for the exploratory data analysis, it was found that additional data pro-

Table 6: Final processed sample dataset for analysis (Continued)

SlibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	A/5 21171	7.25		0.0
1	0	PC 17599	71.2833	C85	1.0
0	0	STON/O2. 3101282	7.925		0.0

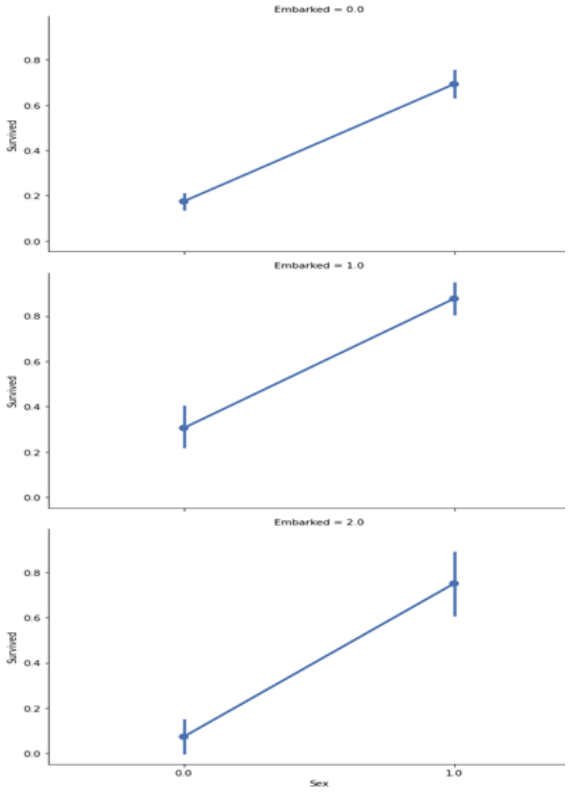

Figure 7: Survival rate with two variables ('Embarked and 'Sex')

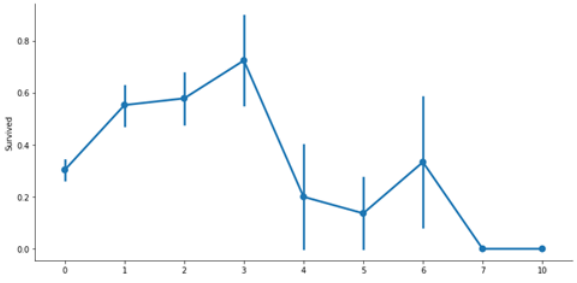

Figure 8: Survival rate vs Family size

Table 7: Final processed sample dataset for analysis (Continued)

Fare$_{R}ange$	Family$_{S}ize$	Alone
1	0	0
1	0	1
0	1	0

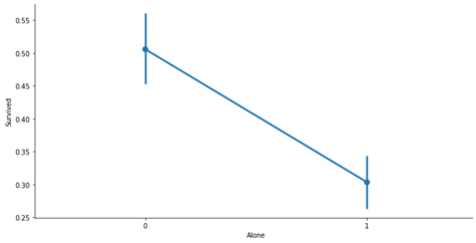

Figure 9: Survival rate vs Alone

cessing was needed to improve the accuracy of the machine learning classification models. Therefore, two additional features, namely 'Has Cabin' and 'Title', were engineered using two advanced Pandas methods, Apply and Lambda. Also, irrelevant features, such as 'PassengerId', 'Name', 'Ticket', 'Cabin', and 'Fare Range', were eliminated. Furthermore, an outlier detection and removal algorithm was applied to the training dataset for enhanced machine learning predictive modelling performance. Finally, each feature was standardized by subtracting the mean and then scaling to unit variance.

4.2 Machine learning algorithms

The project will apply and optimize several machine learning classifiers, including k-nearest neighbors, Naive Bayes, linear support vector machine, Gaussian support vector machine, discriminant analysis, decision trees, random forests, logistic regression, and XGBoost (12).

4.3 Confusion matrix

A confusion matrix contains information about actual and predicted classifications performed by a classifier. The performance of classifiers is evaluated using the data in the matrix (7). Table 8 below shows the confusion matrix for a binary classifier.

Table 8: Confusion matrix for a binary classifier

		Predicted	Predicted
		Negative	Positive
Actual	Negative	True Negative (TN)	False Negative (FN)
Actual	Positive	False Positive (FP)	True Positive (TP

The entries in the confusion matrix have the following meanings in the context of the Titanic dataset:

• True Negative: True Negative is the number of correct predictions that an instance is negative. The classifier here predicted 'Not Survived' and the passenger did not survive.

• False Negative: False Negative is the number of incorrect predictions that an instance is positive. The classifier here predicted 'Survived', but the passenger did not survive.

• False Positive: False Positive is the number of incorrect predictions that an instance is negative. The classifier here predicted 'Not Survived', but the passenger survived.

• True Positive: True Positive is the number of correct predictions that an instance is positive. The classifier here predicted 'Survived' and the passenger survived.

4.4 Model evaluation metrics

There are three standard evaluation metrics for comparing the performance of machine learning classification models: model training time, model accuracy, and model memory utilization (6). In this project, the model accuracy is the evaluation metric considered for evaluating the nine different machine learning classifiers.

The model accuracy is the proportion of the total number of correct predictions and is calculated using the equation below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (1)$$

4.5 Model validation technique

80% of the training examples were used in training each of the nine machine learning classification models. In contrast, the remaining 20% were employed for initial testing of each model to find the best model with the highest accuracy. Then, the best-trained model was tested against the testing examples, and a corresponding accuracy result was calculated locally and online on the Kaggle website. It is worth mentioning that GridSearchCV and RandomizedSearchCV were used for tuning the hyperparameters of most of the models during the training process. The only difference between these two hyperparameter searching algorithms is that GridSearchCV finds the combinations and performs model training, while in RandomizedSearchCV, the model selects the combinations randomly.

4.6 System specifications

All machine learning classification models were tested on a Jupyter Notebook running Python 3.8.8 on macOS 10.14.3 version with 16 GB RAM.

4.7 Results and discussion

4.7.1 Accuracy scores using the training examples

Based on the model validation technique discussed earlier, it was found that the random forest and decision tree algorithms are ranked in the first place since they all have the highest accuracy score of 85.87%. The Gaussian support vector machine algorithm comes in second place with an accuracy score of 84.74%. The third place is shared equally by the KNN, discriminant analysis, logistic regression, and XGBoost algorithms with an accuracy score of 84.18%. Figure 10 below summarizes the results.

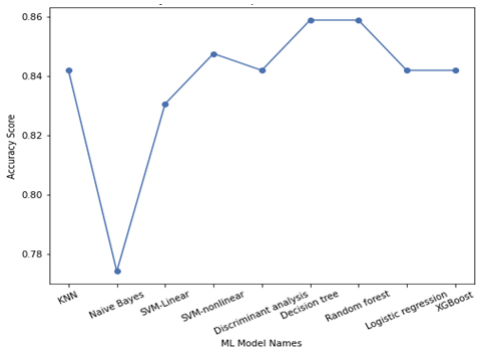

Figure 10: Accuracy scores of multiple machine learning models on the train dataset

The confusion matrices for the random forest and decision tree algorithms based on the train

datasets are shown in Table 9 and Table 10, respectively.

Table 9: Confusion matrix for the random forest classifier

		Predicted	Predicted
		Negative	Positive
Actual	Negative	95	11
Actual	Positive	14	57

Table 10: Confusion matrix for the decision tree classifier

		Predicted	Predicted
		Negative	Positive
Actual	Negative	94	12
Actual	Positive	13	58

All nine algorithms will be tested against the test dataset in the following sub-section since they all yielded very approximate accuracy scores.

4.7.2 Local accuracy scores using the testing examples

The accuracy score results are presented in Figure 12 below, showing the discriminant analysis model achieving the highest accuracy score of 96.88%, followed by the random forest model with 95.45% logistic regression and linear support vector with 94.73%, and finally, the decision trees and Naive Bayes models with 93.54% each.

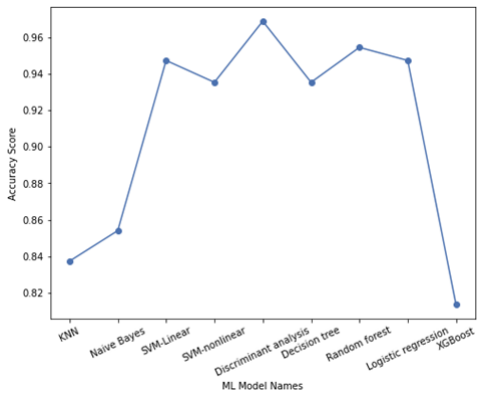

Figure 11: Local accuracy scores of multiple machine learning models on the test dataset

The confusion matrices for the discriminant analysis and random forest and algorithms based on the test datasets are shown in Table 12 and Table 13, respectively

Table 11: Confusion matrix for the discriminant analysis classifier

		Predicted	Predicted
		Negative	Positive
Actual	Negative	257	9
Actual	Positive	4	148

Table 12: Confusion matrix for the random forest classifier

		Predicted	Predicted
		Negative	Positive
Actual	Negative	252	14
Actual	Positive	5	147

4.7.3 Kaggle online accuracy score using the testing examples

The online Kaggle accuracy score results are presented in Figure 11 below, showing the random forest model achieving the highest accuracy score of 78.708% and placing the team in the 1653rd place among other competitors, as shown in Figure 13 below.

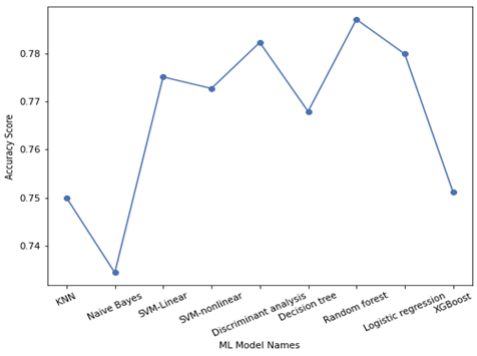

Figure 12: Online Kaggle accuracy scores of multiple machine learning models on the test dataset

It is worth mentioning that the online accuracy score obtained is reasonably good, given that it is the first machine learning project the team is working on. Furthermore, the team's ranking is also reasonably good. Many competitors submitted fake CSV files (i.e. same truth tables provided by Kaggle), allowing them to achieve an accuracy score of 100%, placing other legitimate competitors in low rankings. Therefore, it can be concluded that the online Kaggle scoreboard scores are not very reliable as some competitors used dishonest techniques to improve their rankings.

Also, there is a significant difference between local and online accuracy scores. It is possible that

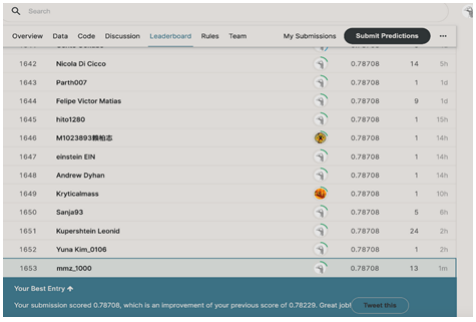

Figure 13: Team's 1653rd place ranking on the Kaggle competition

Kaggle used different ways to calculate accuracy scores.

5 CONCLUSION AND FUTURE WORK

Many factors are impacting the chances of survival for passengers. For example, being a female or a child increases the chances of survival. Also, if a passenger has a first-class ticket, they are more likely to survive than a passenger with second or third-class tickets. Furthermore, males and females are more likely to survive if they embark from the Cherbourg port. Moreover, if a passenger is alone, their survival rate is far less. However, if a passenger travels with a family of size greater than 3, the chances of survival decrease considerably. Finally, the younger passengers are the ones more likely to survive.

The project used a discriminant analysis classifier to obtain an accuracy result of 96.88% on the test dataset, thus scoring higher than all the other algorithms mentioned in previous works.

Regarding limitations, the project only considered the accuracy metric for the evaluation of the classifiers. The project would consider other important evaluation metrics for future work, such as model training time and memory utilization. Also, additional feature engineering, a one-hot encoder for categorical data encoding, and cross-validation could help improve the online Kaggle accuracy score. Furthermore, data analysis could also be improved by answering more detailed questions, including but not limited to the following questions: did women with children have a better survival rate than women without children?

References

[1] Kaggle.com. n.d. Titanic - Machine Learning from Disaster. [online] Available at: <https://www.kaggle.com/c/titanic> [Accessed 29 November 2021].

[2] En.wikipedia.org. n.d. Titanic- Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Titanic> [Accessed 29 November 2021].

[3] Kaggle.com. 2021. Titanic - Machine Learning from Disaster. [online] Available at: <https://www.kaggle.com/c/titanic/data> [Accessed 29 November 2021].

[4] Kaggle.com. 2021. Titanic - Machine Learning from Disaster. [online] Available at: <https://www.kaggle.com/c/titanic/data> [Accessed 29 November 2021].

[5] Mode. n.d. Pandas. [online] Available at: <https://mode.com/python-tutorial/libraries/pandas/> [Accessed 29 November 2021].

[6] Brownlee, J., 2016. Metrics To Evaluate Machine Learning Algorithms in Python. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/> [Accessed 29 November 2021].

[7]]uregina. n.d. Confusion Matrix. [online] Available at: <http://www2.cs.uregina.ca/ dbd/cs831/notes/confusion matrix/confusion matrix.html> [Accessed 29 November 2021].

[8] Lam, E. and Tang, C., 2012. CS229 Titanic Machine Learning From Disaster. [online] Docplayer. Available at: <https://docplayer.net/30593543-Cs229-titanic-machine-learning-from-disaster.html> [Accessed 29 November 2021].

[9] Sherlock, J., Muniswamaiah, M., Clarke, L. and Cicoria, S., 2018. Classification of Titanic Passenger Data and Chances of Surviving the Disaster. arXiv, [online] Available at: <https://arxiv.org/abs/1810.09851> [Accessed 30 November 2021].

[10] Stephens, T., 2014. Titanic: Getting Started With R - Part 3: Decision Trees. [online] Trevor Stephens. Available at: <https://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/> [Accessed 29 November 2021].

[11] A. Singh, S. Saraswat and N. Faujdar, "Analyzing Titanic disaster using machine learning algorithms," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 406-411, doi: 10.1109/CCAA.2017.8229835.

[12] Saini, B., 2021. The Most Common Machine Learning Classification Algorithms for Data Science and Their Code. [online] Medium. Available at: <https://medium.com/swlh/the-most-common-machine-learning-classification-algorithms-for-data-science-and-their-code-9a99c3d32b27> [Accessed 29 November 2021].