Machine Learning-Based Classification as Heuristic Solution to Spaceship Titanic Problem

Louis Widi Anandaputra
Department of Electronics and Computer Science
Universitas Gadjah Mada
Sleman, Indonesia
louiswidianandaputra@mail.ugm.ac.id

Bagus Alwan Bambang
Department of Electronics and Computer Science
Universitas Gadjah Mada
Sleman, Indonesia
bagusalwanbambang2804@mail.ugm.ac.id

***Abstract*—Heuristic approach may be needed in situation where no formally defined rules are available. The case of Spaceship Titanic where passengers are suddenly transported to another dimension is a prime example. As previously machine learning model implementations showed it can be a solution to determine the outcome of an observation, exploration and evaluation of machine learning-based estimators should be performed in the case of the Spaceship Titanic problem as passengers condition are in crucial terms. This work aim to explore multiple machine learning models and utilize the best model to determine the outcome of passengers whose transportation status is yet to be known. By implementing extensive data exploration and understanding process, many additional features from the original observation results are able to be added to increase the data quality. As new features are added, principal component analysis is used to avoid the curse of dimensionality. The data are then fed to the machine learning models, which a cross validation technique is implemented to acquire the best prediction model. Support vector classifier became the best estimator, reaching a training F1-Score of 0.8 and a testing score of 0.79. The SVC was able to determine the outcome of the unknown transportation status of several passengers.**

Keywords—binary classification, feature engineering, feature selection, K-nearest neighbor, bayesian model

I. Introduction (*Heading 1*)

Newtonian or classical physics have formally defined rules on explaining characteristics and events happening in a mathematical manner. The case of Spaceship Titanic however, lacks the mathematical explanation on cases why passengers are transported to another dimension due to insufficient understanding of the general relativity and quantum mechanics theory [1]. The instant transport of passengers in the Spaceship Titanic into another dimension may also be different from the case of a wormhole [2]. As there are currently no evidence of any existence regarding another reality, a heuristic approach is needed to tackle this case in order to predict whether future passengers of the Spaceship Titanic will be transported to another dimension.

A heuristic approach in determining outcomes of certain observations is needed when there are no formally defined rules. An example of this may be in the case of predicting the damage scale caused by an earthquake. Although there exist formal calculation methods to measure the magnitude of the earthquake, the damage on its surroundings are unable to be determined formally due to the massive factors contributing to it. Machine learning models as heuristics estimators of the damage caused by earthquakes to the buildings around the location of earthquakes are implemented to handle this issue [3]. Heuristic estimators are also able to capture other factors that may not have been known to affect the outcome of certain events. In the financial sectors it has been shown that many rules of investment can be complemented with utilizing machine learning models to predict the outcome as it is able to take external factors into account [4].

Having no rules defined formally on the case of alternate dimensions transportation events, other factors may also affect the transportation of passengers into another dimension. Therefore, a heuristic approach is needed in order to predict the safety of the passengers, specifically by utilizing machine learning models. This work will explore the use of machine learning framework as a heuristic solution to the problem of passengers being transported to another dimension. Prediction on several passengers yet to be determined on their transportation status will be done also as one of the aim of this work.

II. Data Understanding

The dataset, derived from a Kaggle competition, includes various features that describe the passengers on the spaceship Titanic [1]. The primary objective is to predict the binary outcome “Transported”, indicating if a passenger was transported to another dimension. The data is synthetically generated, simulating a fictional interstellar trip. It holds a diverse set of attributes that captures different aspects of the passengers’ profiles and experiences aboard the spaceship.

TABLE I. Dataset Overview

Feature Name	*Description*	*Missing Values*
PassengerId	A unique Id for each passenger. Each Id takes the form gggg_pp where gggg indicates a group the passenger is travelling with and pp is their number within the group.	288
HomePlanet	The planet the passenger departed from.	310
CryoSleep	Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage.	274
Cabin	The cabin number where the passenger is staying. Takes the form deck/num/side, where side	270

Feature Name	***Description***	***Missing Values***
	can be either P for Port or S for Starboard.	
Destination	The planet the passenger will be debarking to.	296
Age	Age of Passenger	263
VIP	Whether passenger paid for VIP service	289
RoomService	Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.	263
FoodCourt		283
ShoppingMall		306
Spa		284
VRDeck		268
Name	The first and last names of the passenger.	294
Transported (Target)	Whether the passenger was transported to another dimension.	4183

a.

As described by Table I, the dataset has several missing values due to incomplete observations. The *Transported* column has the most missing values as it shows there are 4183 passengers yet figured out whether they will be transported to another dimension. Those information would then be predicted to know whether the passengers are to be transported to another dimension.

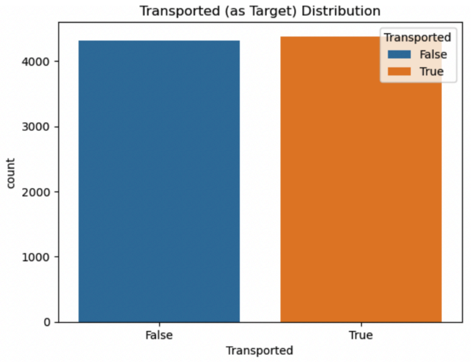


Fig. 1. Class distribution of target.

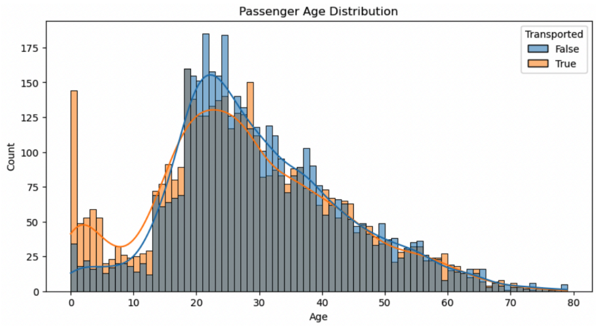


Fig. 2. Age distribution of passengers.

The target variable Transported is balanced. As shown by Fig 1, 50.4% of passengers are being transported and 49.6% not transported. This would eliminate the need for sampling techniques to be used on the dataset, ensuring no information is loss through undersampling or overfitting through oversampling steps. Another information came up on the age distribution. As shown by Fig 2, younger passengers below 18 are more likely to be transported, passengers within the age range of 20 until 45 are less likely to be transported, and passengers above the age of 45 are equally likely to be transported. This would indicate that age groups are able to be extracted.

In terms of the expenditure features ranging from *RoomService* to *VRDeck*, most of the passengers did not make a spending on the luxury amenities. As exampled by Fig 3, majority of the passengers did not make a spending on room service or food court. This would indicate there is another information can be extracted, namely whether the passengers make an expenditure on the spaceship.

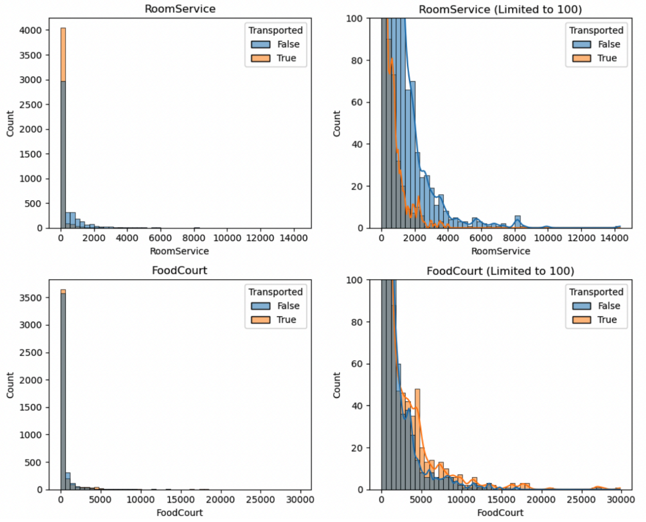


Fig. 3. Sample of expenditures on different luxury amenities.

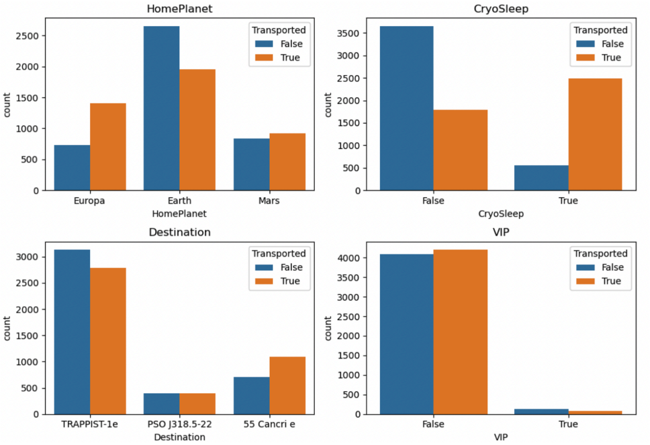


Fig. 4. Distribution of categorical features.

Four categorical features available in this dataset are *HomePlanet, CryoSleep, Destination,* and *VIP.* Among the categorical features, *VIP* does not appear to be useful, as it is illustrated by Fig 4, the distribution is highly imbalanced, suggesting that the *VIP* might need to be dropped. In contrast, the *CryoSleep* feature appears show distinction in which passengers will be transported to another dimension. Passengers who are in *CryoSleep* are more likely to be transported. *HomePlanet* distribution also showed passengers who are travelling from Earth are less likely to be transported, similar to those who are travelling to TRAPPIST 1-e as their destination.

TABLE II. SAMPLE OF QUALITATIVE FEATURES

Passenger ID	***Cabin***	***Name***
0001_01	B/0/P	Maham Ofracculy
0002_01	F/0/S	Juanna Vines
0003_01	A/0/S	Altark Susent
0003_04	A/0/S	Solam Susent
0004_01	F/1/S	Willy Santantines

b.

As Table II described, several information can be extracted from the qualitative features. Passenger groups can be extracted from *passenger ID*, this information can show how many passengers are travelling in a group. The *cabin* feature can be extracted further to the deck, number, and side. Lastly, the *name* feature can help aggregate the families, showing how many passengers are in the same families.

III. METHODOLOGY

Several steps will be involved in order to predict the 4183 passengers' transportation status. Illustrated by Fig 5, the first step that has been discussed is the data understanding process of the given dataset of Spaceship Titanic [1]. Insights from the data understanding step are then utilized to extract and engineer new features in the feature extraction step. The acquired data are then cleaned before being inputted into the train and test split process. Feature selection with principal component analysis (PCA) are then used to select the best features. The process would then continue to split the data into training, validation, and testing subset for model training and evaluation. The testing data would then be predicted usign the best model acquired from the training process to

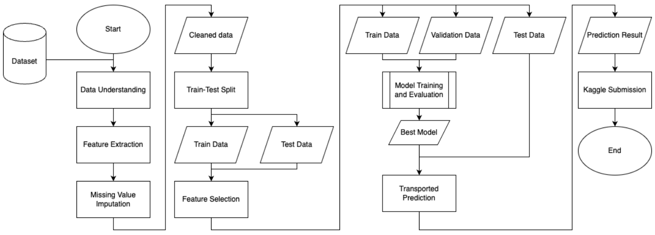


Fig. 5. Diagram of the prediction methodology.

A. Data Understanding

The acquired data are analyzed and explored. The analysis process is aimed to extract information and insights from the original data. The extracted information would then be the basis of the feature extraction process where additional information would be represented in a vector space format.

B. Feature Extraction

From the several information acquired through the data understanding process, additional features are able to be added to the data. The added features would be correlated to several aspects of the original data, namely passenger age, expenses on luxury amenities, seating location (*Cabin*), and the utilization of the *Name* feature.

As it was illustrated by Fig 2, there are several age groups able to be represented as another feature. This work would represent the age into four different groups. As illustrated by Fig 6, underage passengers, younger adults, older adults, and elderly passengers that are included in the 45 years old and above are represented individually.

Passengers total expenditures are able to be extracted by summing the spending on the *RommService, VRDeck, FoodCourt, ShoppingMall,* and *Spa* feature values. Indicator of whether the passengers make a spending on the spaceship are also extracted by looking at the total expenditures. As illustrated, if the total expenditure value is 0, the indicator would have the value 0, suggesting that the passenger did not make any spending.

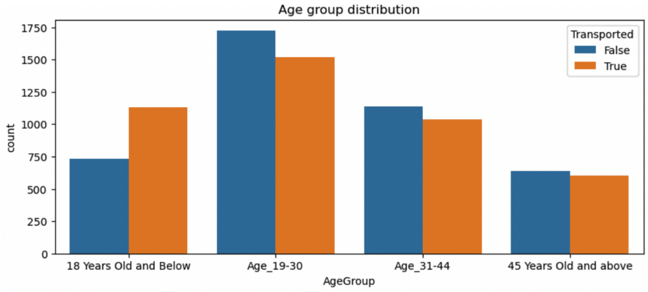


Fig. 6. Age group distribution.

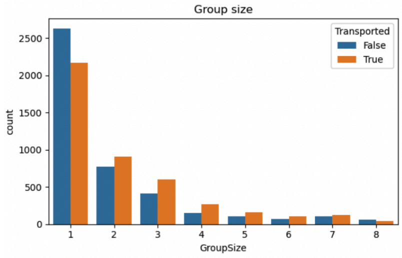


Fig. 7. Passenger group size distribution.

The *PassengerId* feature would also allow the extraction of distinct groups through the first four digits where the distinct groups would then allow the extraction of the *GroupSize* the passenger is travelling in. In further manner, information whether the passenger is travelling solo is also able to be extracted by determining whether the *GroupSize* is one as described both by Fig 7 and Fig 8. Similarly, this method is also applicable to the passenger name, where the *Firstname* and *Surname* are able to be extracted by separating the first and second word of the name. The *Surname* would then give the information about how many family members the passenger is travelling with in *FamilySize* as Fig 9 explains the distribution of *FamilySize.*

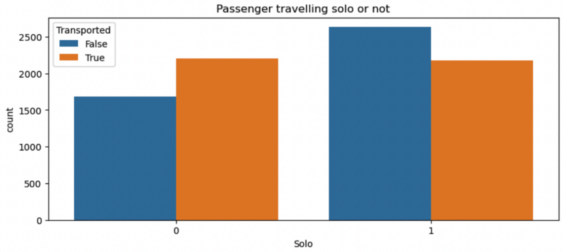


Fig. 8. Passenger group size distribution.

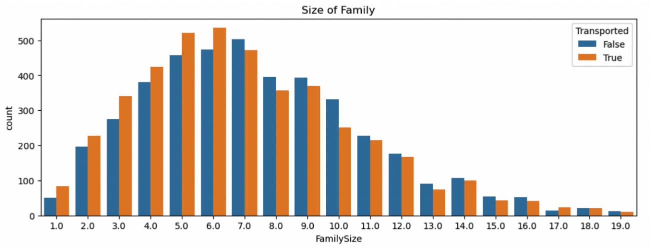


Fig. 9. Passengers family size distribution

Last feature to be extracted is the cabin-related features. As described by Table I, the original *Cabin* feature consist of the deck, number, and the side. These three features are able to be extracted using string separation method with regular expression. As Table II also describes, the cabin deck, number, and side can be separated using '/' as separator.

C. Missing Value Imputation

Described by Table I, all features except for the target *Transported* have around 200 entries missing. Several strategies are implemented to impute the missing values:

- Mode is used to fill the missing values on *HomePlanet, CabinDeck,* and *CabinSide* directly. The missing missing *Surname* entries are filled by the mode of the *Surname* in the same group the passenger is travelling in.
- Median is used to fill the missing values on features relating to expenditures and age.
- *CryoSleep* missing value imputation utilize the information on whether the passenger make any spendings.

All derived features such as *FamilySize* and *AgeGroup* are then updated according to the imputed values.

D. Train and Test Subset Splitting

The original data have several rows of *Transported* as target value missing, specifically 4183 entries. The 4183 rows of data are regarded as the test subset where a submission to the Kaggle platform will be made and the rest is the train subset.

E. Feature Selection

The cleaned data are then pre-processed using several methods. Numerical data with skewed distribution are first normalized with logarithmic transformation [5]. Standardization to ensure the data are on the same measurement are then performed to the numerical data. In terms of categorical data such as *HomePlanet, AgeGroup, CabinDeck,* and *Destination* are encoded with one-hot encoding, a process of creating new binary features according to the amount of unique values in each of the categorical features. The pre-processing step would produce a total of 33 features.

Among the 33 features, *PassengerId, Age,* and *Group* features are dropped to remove redundancy. The pre-processed data are then being inputted to the PCA process. The PCA, which is a dimensionality reduction method to transform the dataset into a lower dimension [6] showed that the optimal value of number of principal components is 15, where 98% of variance is explained.

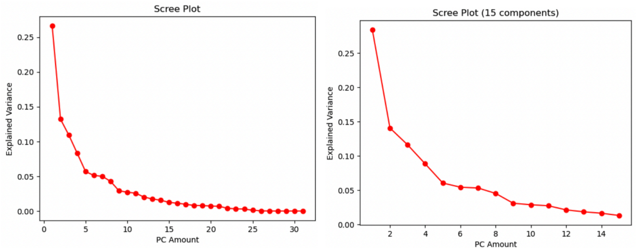


Fig. 10. Optimal principal component amount

F. Model Training and Evaluation

The model training and evaluation process will involve several steps. As illustrated by Fig 11, the training data acquired from the train and test set split is then partitioned into training and validation subset. The subsets having proportions of 25% for validation and 75% for training, the amount of training and testing data entires are 6519 and 2174 respectively. The training subset is then used for the model training process with hyperparameter tuning using grid search cross validation (grid-search CV) [7].

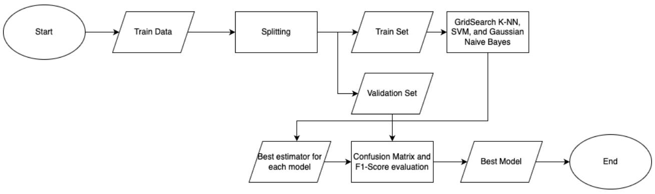

Start
Train Data
Splitting
Train Set
GridSearch K-NN, SVM, and Gaussian Naive Bayes
Validation Set
Best estimator for each model
Confusion Matrix and F1-Score evaluation
Best Model
End

Fig. 11. Model training, evaluation, and selection process.

The compared machine learning models in this work are classifiers, specifically K-Nearest Neighbor (K-NN) [8], Support Vector Machine Classifier (SVC) [9], and a Gaussian Naïve Bayes classifier (GaussianNB) [10]. The parameters evaluated in the grid-search CV for each models are described in Table III. The K-NN would evaluate between manhattan and euclidean distance as well as several odd number neighbors. The SVC would evaluate the type of kernel and C. Lastly, the GaussianNB would evaluate the variable smoothing value.

TABLE III. PARAMETERS TO BE SELECTED

Model	Parameters
K-NN	{Metric:[Minkowski] P : 1,2] Number of neighbors: [3, 5, 7, 9, 11, 13, 15]}
SVC	{kernel : [rbf], gamma : [1e-3, 1e-4], C : [1, 10, 100, 1000]}, {kernel : [linear], C: [1, 10, 100, 1000]}
GaussianNB	variable smoothing: [1.0, …, 2.848035868435799e-05, …, 1e-09]

The best estimator for each of the K-NN, SVC, and GaussianNB models are then acquired through the grid-search CV. Each of the models are then evaluated using the confusion matrix regarding the predicted results and the real value [11] by using the validation set. Another evaluation metric of F1-score is then used to determine the performance of each model. The F1 score would be an average to the precision, a measure describing how many predicted transported passengers are truly transported, and the recall, a measure describing how many transported passengers are truly predicted [12]. The evaluation process would then provide the best predicting model with the best performing hyperparameters.

G. Prediction

Through the training and evaluation process, the best model would predict the outcomes of the test data set. The 4183 passengers yet to be determined of the transportation status would be predicted based on the same features the model is trained on.

IV. Results and Discussions

As Table III described the pool of hyperparameter selection, the grid-search CV provided the best model parameters. Table IV describes that for the K-NN classifiers, the best distance measure is the euclidean distance with the optimal number of neighbors to provide good classification results is 15. As for the SVC, the best kernel turned out to be the radial basis function (rbf) kernel with a gamma value to be 0.001 and a C value of 1000. Lastly, the GaussianNB turned out to be having the best performance with variable smoothing value of around 0.2848.

TABLE IV. Best Parameters for Each Models

Model	Best Parameters
K-NN	{Metric: Minkowski P : 2 Number of neighbors: 15}
SVC	{kernel : rbf, gamma : 0.001, C : 1000}
GaussianNB	variable smoothing: 0.28480358684435802

TABLE V. Confusion Matrix for Each Classifier

Model			Predicted Values	
			Not Transported	Transported
K-NN	**True Values**	Not Transported	849	233
		Transported	250	842
SVC		Not Transported	**797**	**285**
		Transported	**175**	**917**
GaussianNB		Not Transported	826	256
		Transported	275	818

When the best models from the grid-search CV is tested upon the validation data, the result showed that the SVC provided the best F1-Score as described by Table VI with a value of 0.80. However, its results showed that the model leaned more towards predicting the *transported* class as the confusion matrix in Table V showed a lower score of precision for the *transported* class compared to the *not transported.*

TABLE VI. F1 Metric on Each Model

Model	F1-Score
K-NN	0.77
SVC	0.80
GaussianNB	0.76

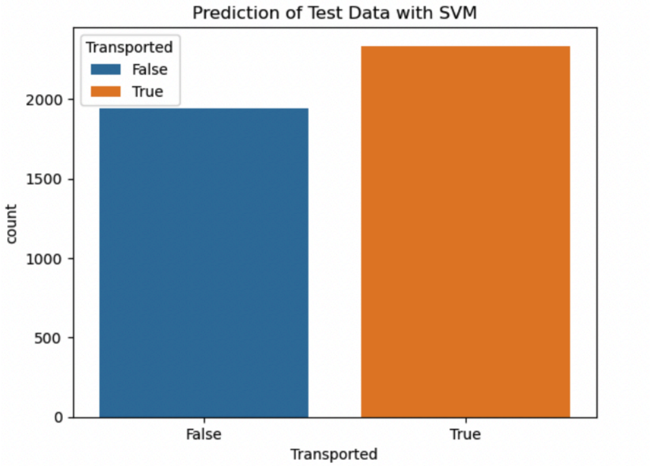


Fig. 12. Test data prediction distribution with support vector classifier

Although having much higher F1-Score than the other classifier, it has a bias towards predicting a certain class. This may due to the fact that support vector machines are spatial-aware models and it applies to this case of using SVC. Certain features may cause the SVC to lean more towards a certain class as the rbf kernel is not linear. In contrast to the issue presented by SVC, K-NN that utilizes distance without spatial information, or GaussianNB that utilizes probabilities, did not exhibit the issue as their precision and recall values are more stable. However, since the disparity between values in the SVM, it can be selected as the best classifier to predict the unknown *Target* columns. in the test dataset.

The outcome of the test dataset was able to be determined using the SVC. However, illustrated by Fig 12, the prediction showed more transported class are being predicted. When submitted to the Kaggle competition, the result showed little difference in terms of score, as the value is 0.79. Meaning, the SVC model is able to capture the information and patterns in order to predict whether the passengers are to be transported.

V. Conclusions

This work managed to explore the use of machine learning classifiers as heuristic solutions to the problem of passengers being transported mysteriously to another dimension. The SVC became the best classifier among other tested models and is able to capture patterns and information of the observation data. The similarity in score of the train and test data provided by the prediction results of SVC showed that a heuristic approach can aid the process of determining an outcome of an observation without the need of a formally defined rule.

Future works should try to improve the heuristic solution of machine learning model implementation by adding more robust pre-processing steps. Improved missing value imputation strategy that combines the exploratory data analysis insights should be done to improve the quality of the data.

References

[1] Kaggle, "Spaceship Titanic Competition Overview," Retrieved June 14, 2024, from https://www.kaggle.com/competitions/spaceship-titanic/overview.

[2] D.-C. Dai and D. Stojkovic, 'Observing a wormhole', Phys. Rev. D, vol. 100, p. 083513, Oct. 2019.

[3] K. Chaurasia, S. Kanse, A. Yewale, V. K. Singh, B. Sharma, and B. R. Dattu, 'Predicting Damage to Buildings Caused by Earthquakes Using Machine Learning Techniques', in 2019 IEEE 9th International Conference on Advanced Computing (IACC), 2019, pp. 81–86.

[4] B. T. Kelly and D. Xiu, "Financial Machine Learning," SSRN, Jul. 1, 2023.

[5] C. Feng et al., 'Log-transformation and its implications for data analysis', Shanghai Arch Psychiatry, vol. 26, no. 2, pp. 105–109, Apr. 2014.

[6] I. Jolliffe, 'Principal Component Analysis', in International Encyclopedia of Statistical Science, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1094–1096.

[7] S. M. LaValle, M. S. Branicky, and S. R. Lindemann, 'On the relationship between classical grid search and probabilistic roadmaps', The International Journal of Robotics Research, vol. 23, no. 7–8, pp. 673–692, 2004.

[8] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, 'k-Nearest Neighbor Classification', in Data Mining in Agriculture, New York, NY: Springer New York, 2009, pp. 83–106.

[9] N. Cristianini and E. Ricci, 'Support Vector Machines', in Encyclopedia of Algorithms, M.-Y. Kao, Ed. Boston, MA: Springer US, 2008, pp. 928–932.

[10] G. I. Webb, 'Naive Bayes', in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 713–714.

[11] K. M. Ting, 'Confusion Matrix', in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 209–209.

[12] C. Goutte and E. Gaussier, 'A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation', in Advances in Information Retrieval, 2005, pp. 345–359.

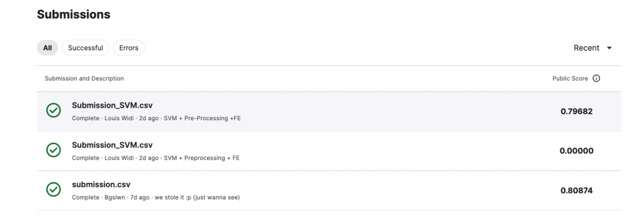
Appendix

The complete source code for the implementation described in this paper is available in a GitHub repository. The repository contains the Python scripts, Jupyter Notebooks, and any other relevant files used in the development and evaluation of the proposed approach. The repository can be accessed at the following URL: https://github.com/louiswids/Spaceship-Titanic

This repository contains:

- Data preprocessing scripts: Scripts used to clean and preprocess the dataset.
- Feature engineering scripts: Scripts used to create new features from the raw data.
- Model training and evaluation scripts: Scripts for training and evaluating the machine learning models.

Kaggle Competition participation



Submissions

Submission and Description	Public Score
Submission_SVM.csv Complete · Louis Widi · 2d ago · SVM + Pre-Processing +FE	0.79682
Submission_SVM.csv Complete · Louis Widi · 2d ago · SVM + Preprocessing + FE	0.00000
submission.csv Complete · Bgslwn · 7d ago · we stole it :p (just wanna see)	0.80874