ISSN 1870-4069

# Predicting the Survival of Titanic Passengers Using Machine Learning and Smart Data Cleaning

José Alberto Hernández-Aguilar[1], Julio César Ponce Gallegos[2], Víctor Hugo Pacheco Valencia[1], Juan Carlos Bonilla Robles[1]

[1] Universidad Autónoma de Morelos, Mexico

[2] Universidad Autónoma de Aguascalientes, Mexico

jose_hernandez@uaem.mx

**Abstract.** In this paper we discuss how machine learning can be used to predict the survival of Titanic passengers, we present the general pipeline to predict the survival of passengers, and we focus in the first stage on what we call smart data cleaning, this process can reduce the number of variables and increase precision, recall, and f-score metrics. We compare eight machine learning algorithms: Logistic Regression, Decision Tree, KNN, Gaussian Naïve Bayes, Perceptron, LSCV, Random Forest, and Stochastic Gradient Descend (SGD). The best results were obtained with cross-validation and Logistic Regression with a precision 0.8238, Perceptron 0.8142, and SGD 0.8142.

**Keywords:** Machine learning, smart data cleaning, survival of passengers.

## 1 Introduction

An example of a ship disaster is the Titanic, it sunk in the Atlantic sea in April 1912, only (32%) 722 passengers survived of a total 2224 and its crew, titanic sank after hitting an iceberg [1, 2]. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. According [3] the interesting observation is that in the case of Titanic, some people were more likely to survive than others, like women, children were the ones who got the priority to the rescue, like in Hollywood movies script "children and ladies first".

The purpose of this research is to predict passenger survival rate on the Titanic using different machine learning models and intensive data cleaning. In machine learning, data is divided into Training and Testing, the split ratio could be 70-30 or 80-20; in this research, we use the last one, we have 1047 entries for train and 262 for the test. We use Scikit-learn Machine Learning in Python [4, 5] for making our experiments.

**Table 1.** Related recent work.

| Author(s) | Technique | Results |
|---|---|---|
| AM Barhoom, AJ Khalil, BS Abu-Nasser, MM Musleh (2019) [1] | Neuronal networks | 99.28% accuracy |
| B. Balakumar, P. Raviraj, K. Sivaranjani (2019)[7] | Various machine learning algorithms namely Logistic Regression, Naive Bayes, Decision Tree, Random Forest | 94.26 accuracy with Logistic Regression |
| Farag, N., & Hassan, G. (2018, May) [8] | Decision Trees and Naïve Bayes | The Decision Tree algorithm has accurately predicted 90.01% of the survival of passengers, while the Gaussian Naïve Bayes witnessed 92.52% accuracy in prediction |
| Kakde, Y., & Agrawal, S. (2018)[3] | LR, Decision Tree, Random Forest, and SVM | LR. 0.8372 best result of Accuracy |
| Tabbakh, A., Rout, J. K., & Rout, M. (2020) [9] | logistic regression, k-nearest neighbors, SVM, naive Bayes, decision tree, and random forest | NA |

# 2 Related Work

## 2.1 Machine Learning

According [6] insufficient quality of data was the second biggest obstacle to employing AI, narrowly behind a shortage of internal talent. In Table 1 is shown related work of recent research results of predictions about Titanic survivor's prediction.

As shown in the above table, there is a lot of research discussing the prediction of survivors of Titanic, this is a due database of Titanic's passengers is used in several courses of machine learning around the world. The best results were obtained by using neuronal networks reported in [1], and with Logistic Regression in [7].

## 2.2 Smart Data Cleaning

Quality of data has been analyzed from early days of computing, with the emergent techniques for big data and machine learning techniques this is mandatory to carry out, in [10] is discussed different faces of data quality in the context of this new scenario, in this research tools available and trends are discussed to going beyond just data cleaning.

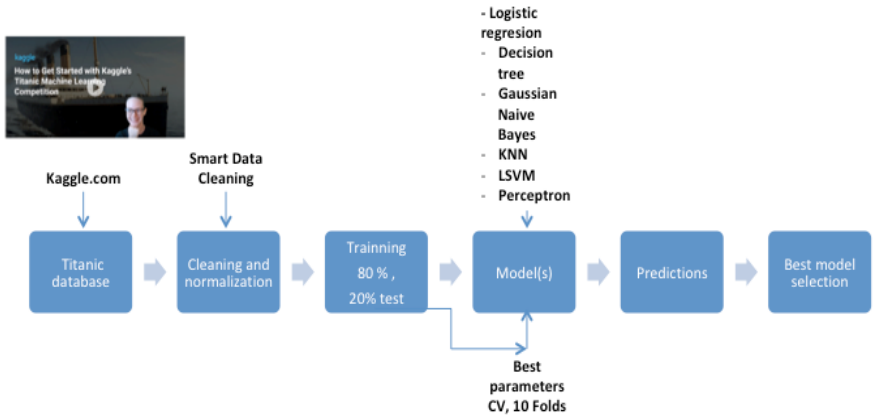



**Fig. 1.** Proposed Methodology (own source).

| Data base | Kaggle | Used |
|---|---|---|
| Train | 891 | 1047 |
| Test | 418 | 262 |
| Total | 1309 | 1309 |
| Proportion for test | 31.9% | **20%** |

**Fig. 2.** Test and Train proportion Used in this research (own source).

## 3 Methodology (Pipeline)

The proposed methodology includes six stages: database selection, cleaning and normalization, Training and Test, model selection, prediction, and best model selection (see Fig. 1).

Next, we will discuss briefly each step.

### 3.1 Database

The database used was obtained from the Kaggle competitions [2] web page, data is publically available. We used 80% of data for training and 20% for testing stages (see Figure 2).

### 3.2 Data Cleaning

This stage is one of our main contributions, we called it Smart data cleaning, this idea was inspired in data engineering applied by large IT companies like IBM and Oracle, the algorithm is shown below:

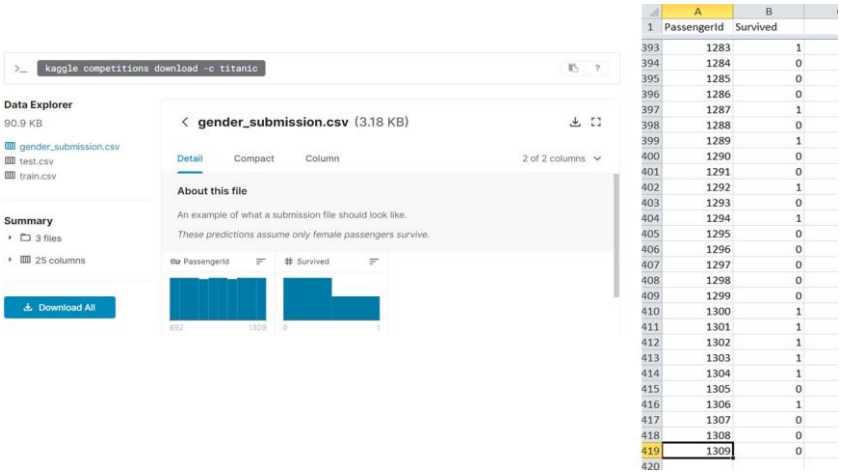



**Fig. 3.** Detail of gender submission and classification of survivors according to Passenger ID.

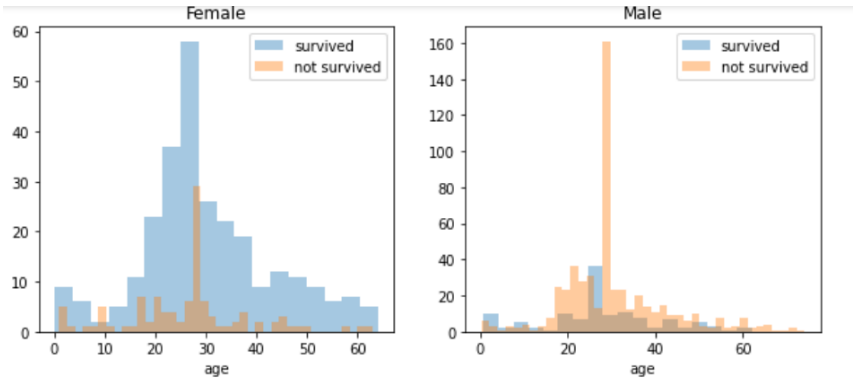



**Fig. 4.** Detail of gender versus age contrasted survived and not survived.

**Smart data cleaning**

1. Deep knowledge of data
    a. Visualization of data (Human Analyst)
    b. Crosstab
2. Auto-fill of missing data (by imputation through software-agent)
3. Elimination of not useful columns (Only the columns that contribute to the result are used)

**Deep Knowledge of Data.** To get deep knowledge of data is necessary to identify what is inside the data. Where does the information come from? How was the information generated? In this stage participates the human analyst/expert to improve the pipeline. In next figure is shown in detail the gender of passengers.

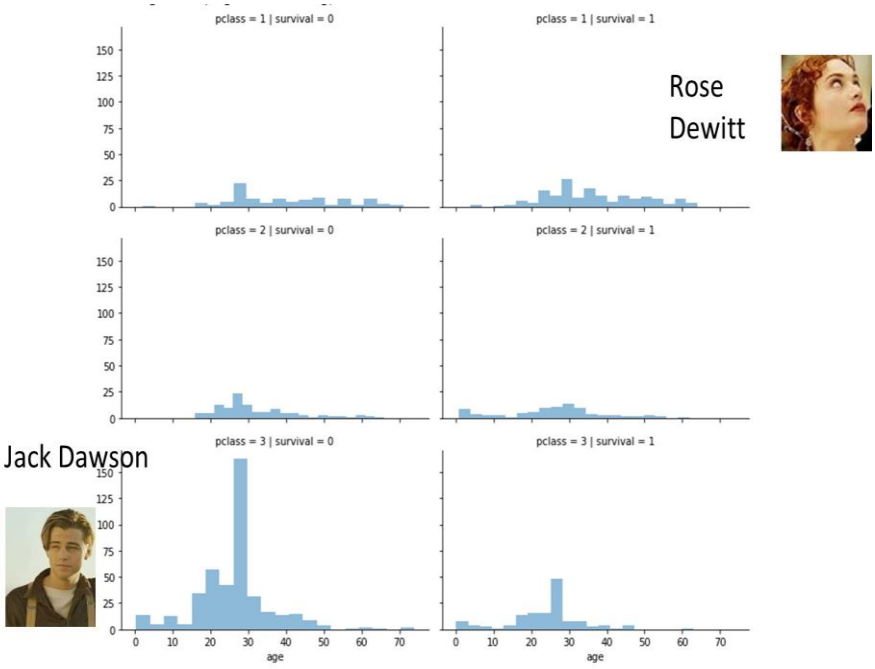



**Fig. 5.** Social class (pclass) versus age, compared by not survivors (left side) and survivors (right side).

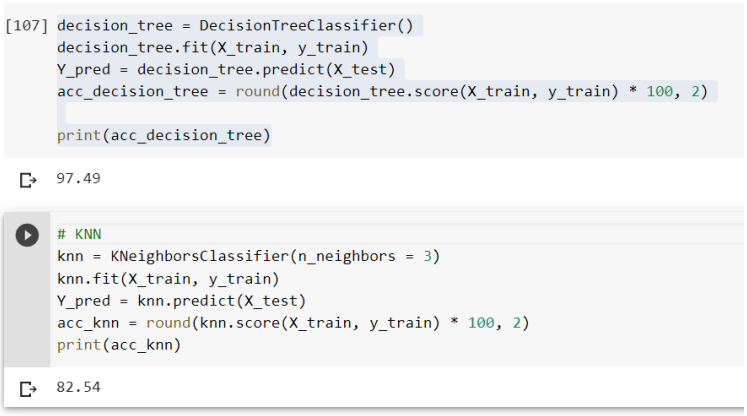
| | Total | % |
|---|---|---|
| **Cabin** | 687 | 77.1 |
| **Age** | 177 | 19.9 |
| **Embarked** | 2 | 0.2 |
| **Fare** | 0 | 0.0 |
| **Ticket** | 0 | 0.0 |

**Fig. 6.** Crosstab of fields by missing values.

**Data Visualization.** To get deep knowledge of data is recommended to visualize data into graphs. The above figure shows what happens during the Titanic's evacuation, the instructions were to abandon the ship "Ladies and children first", therefore there were more females (in all ranges of age) and children who survived. In figure 5, shows how rich first-class women and children survived.

**Crosstab.** To analyze the quality of data a crosstab of fields will reveal which data must be completed, imputed, or deleted. Figure 6 shows some fields are complete (i.e. Fare Ticket), but some others must be imputed (i.e. Age) or deleted (i.e. Cabin).

As shown in figure 6, there are fields with missing values, but we can improve some of them using auto-filling, as described in the next pseudo-code.



```
[107] decision_tree = DecisionTreeClassifier()
      decision_tree.fit(X_train, y_train)
      Y_pred = decision_tree.predict(X_test)
      acc_decision_tree = round(decision_tree.score(X_train, y_train) * 100, 2)

      print(acc_decision_tree)

      97.49

      # KNN
      knn = KNeighborsClassifier(n_neighbors = 3)
      knn.fit(X_train, y_train)
      Y_pred = knn.predict(X_test)
      acc_knn = round(knn.score(X_train, y_train) * 100, 2)
      print(acc_knn)

      82.54
```

**Fig. 7.** Details of Decision tree and KNN models.

**Auto filling of missing data**

1. If columns are not promising (i.e. Too empty or do not contribute)

   If (ttnc_df[i] >= 70%)

   Delete column i

   If (ttnc_df[i] is not correlated)

   Delete column i

2. If data is incomplete but could be useful

   Auto-fill column by imputation (i.e. Median value, or random values)

3. Convert text fields to categories (numerical data)

### 3.3 Training and Test

We use 80% of data for training and 20% for testing.

### 3.4 Model

We train and test the next models: Logistic Regression, Decision Tree, KNN, Gaussian Naïve Bayes, Perceptron, LSCV, Random Forest, and Stochastic Gradient Descend (SGD). In further research, we will describe each of these models.

### 3.5 Predictions and Model Selection

We obtain predictions for each model, first without cross-validation, and later with 10 folds cross-validation. For our experiments, we used Google Collab, Python 3, and set up the environment GPU enabled. The results are shown in the next section.

**Table 2.** Results of prediction.

| | **LR** | **Deci-sion Tree** | **KNN** | **Gaussian Naive Bayes** | **Perceptron** | **LSVC** | **Random Forest** | **Stochastic Gradient Descent (SGD)** |
|---|---|---|---|---|---|---|---|---|
| **Proposed pipeline** | | | | | | | | |
| Precision | 0.7859 | **0.9749** | 0.8254 | 0.7691 | 0.6938 | 0.7656 | **0.9749** | 0.7392 |
| **Donges (2018)** | | | | | | | | |
| Precision | 0.8114 | 0.9282 | 0.8732 | 0.7710 | 0.8070 | 0.8081 | 0.9282 | 0.7699 |
| **Opt. w/ CV** | **0.8238** | 0.7619 | NA | NA | **0.8142** | 0.8333 | 0.7714-**0.8142** | **0.8142** |

## 4 Results and Discussion

### 4.1 Predictions

Table 2 shows the results of predictions, the first row shows the precision of the proposed pipeline, the second row shows precision obtained in [11], the third row shows the results of our proposal using ten folds cross-validation (CV).

The main results are shown in black, as can be seen, the best results are obtained with the proposed pipeline and Logistic Regression with ten folds cross-validation. Perceptron and SGD get second-best results 0.8142; in these three cases, precision was better than results obtained by [8].

ROC AUC CURVES of [8] and of our most representative results are shown in the next figures. A classifier that is 100% correct, would have a ROC AUC Score of 1 and a completely random classifier would have a score of 0.5.

As can be seen in Table 2, and figures 9 and 10; our results were better than those reported by [11]. Our best result of Accuracy (0.8238) was obtained with Logistic Regression which is promising according to results obtained in [7] and [3].

## 5 Conclusions and Future Work

The smart data-cleaning algorithm employed as a stage of machine learning demonstrates its functionality to improve the results of prediction for survivors of titanic.

Our future work will be to apply the proposed methodology in different databases, including synthetic and real-world databases, for instance in bank databases to analyze information from users to predict payment of credits. In the case of applications in education, we plan to predict performance in online assessments based on student habits, previous grades, preferences, and usage of LMS during this period of the pandemic.

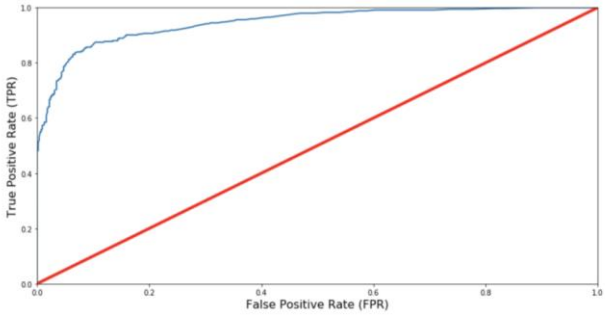



**Fig. 8.** Random Forest Classifier, precision 0.8019, Roc AUC Score=0.9450 [11].

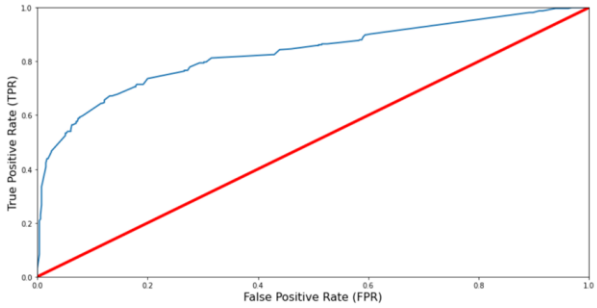
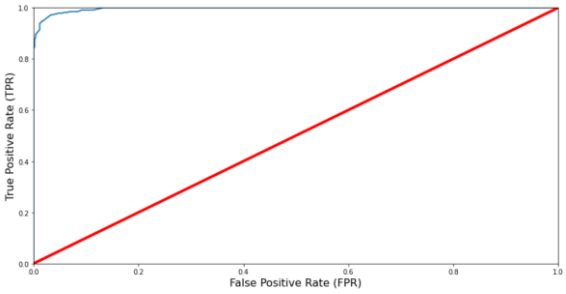



**Fig. 9.** Logistic Regression, precision 0.8238, ROC AUC Score = 0.8286.




.

**Fig. 10.** Random Forest, precision 0.8142, ROC-AUC-score = 0.9963.

## References

1. Barhoom, A.M., Khalil, A.J., Abu-Nasser, B.S., Musleh, M.M., Abu-Naser, S.S.: Predicting Titanic Survivors using Artificial Neural Network. Int. J. Acad. Eng. Res., 3, pp. 8–12 (2019)
2. Kaggle: Kaggle (2019) https://www.kaggle.com/c/titanic.
3. Shefali Agrawal, I.: Kakde Asst Professor AITR, Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques.

Artic. Int. J. Comput. Appl., 179, pp. 975–8887 (2018) doi: 10.5120/ijca2018917094.

4. Scikit-learn: Scikit-learn: machine learning in Python — scikit-learn 0.23.1 documentation (2020) https://scikit-learn.org/stable/.
5. Raschka, S., Mirjalili, V.: Python Machine Learning: Machine Learning and Deep Learning with Python (2020) https://books.google.com.mx/books?hl=es&lr=&id=sKXIDwAAQBAJ&oi=fnd&pg=PP1&dq=Raschka,+S.+and+Mirjalili,+V.+(2019).+Python+machine+learning&ots=V8LmsSTDEn&sig=Zxkoa3nncSs4gJvLlg8IUlO9DAA&redir_esc=y#v=onepage&q=Raschka%2C S. and Mirjalili%2CV.
6. MIT, T.R.I.: Insights survey: Talent shortage is the top AI challenge – MIT Technology Review Insights (2020) https://insights.techreview.com/live-ai-poll-key-stats/.
7. Balakumar, B., Raviraj, P., Sivaranjani, K.: Prediction of survivors in Titanic dataset: A comparative study using machine learning algorithms (2020) https://pdfs.semanticscholar.org/545a/9e5da57058cf08e32eae6b5816839505ac3c.pdf.
8. Farag, N., Hassan, G.: Predicting the Survivors of the Titanic-Kaggle, Machine Learning From Disaster (2018) doi: 10.1145/3220267.3220282.
9. Tabbakh, A., Rout, J.K., Rout, M.: Analysis and Prediction of the Survival of Titanic Passengers Using Machine Learning. In: Lecture Notes in Networks and Systems, pp. 297–304, Springer (2021) doi: 10.1007/978-981-15-4218-3_29.
10. Gudivada, V.N., Ding, J., Apon, A.: Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations Big Data Management View project Cognitive Computing (2017)
11. Donges, N.: Predicting the Survival of Titanic Passengers (2020) https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8.