International Journal of Academic Engineering Research (IJAER)
ISSN: 2643-9085
Vol. 4 Issue 10, October - 2020, Pages: 44-52

Titanic: Statistical Analytics from Disaster

Joao Negreiros, Samia Loucif, Mohammad Amin Kuhail, Ahmed Seffah

Faculty of Technological Innovation, Zayed University
Abu Dhabi, UAE

***Abstract**: This study aims at utilizing statistical models to analyze the sort of passengers who likely survived the Titanic shipwreck in April 1912. A comparison between three approaches (logistic regression, decision tree, neural networks) was accomplished, including an exploratory analysis of the given dataset (composed of nine variables). Since the performance of those methods was quite similar, path analysis was computed to set up a statistical model for the survival (or not) factors of the RMS Titanic event. According to our calculations, this approach only explains 51.7% of the survival aftermath. Positively, other qualitative and quantitative critical factors were not included here.*

Keywords— Titanic, Data analytics, K-Means, Logistic regression, Decision trees, Neural networks, Path analysis.

1. INTRODUCTION

A myth is not a falsehood. Rather, a myth is a sophisticated social representation; a complex relationship between history, reality, culture, imagination and identity. The story of the Titanic is a modern myth par excellence. To describe it as a myth is not, however, to attempt to reduce it to a primitive level. On the contrary, the sinking of the Titanic is one of the most infamous shipwrecks in history [1]. On April 15, 1912, the unsinkable RMS Titanic sank after colliding with an iceberg, resulting in the death of 1502 out of 2,224 passengers and crew. While there was some element of luck involved in surviving, it seems that some groups of passengers were more likely to survive than others [2]. Twenty-eight different nationalities were present such as British, French, American, Portuguese, Turkish, Mexican, Chinese, Australian, Swiss, Japanese, Syrian, Uruguayan and Russian. In a first overall view, there were 434 women (75% survived), 112 children (50% survived) and 1680 men (32% survived). In a further detail, 97%, 86% and 49% of First, Second and Third class women survived, respectively. On the other hand, 86%, 100% and 31% of First, Second and Third class children survived. At last, 62%, 43% and 25% of First, Second and Third class men survived. Besides all this descriptive statistics, what can inferential statistics state about this event? Is it possible to build a model that predicts the likelihood of any specific passenger based on his/her records data (age, gender, socio-economic class, ticket fare, number of siblings and parents, port of embarkation)?

The use of data analysis tools has been widely considered during the last few years not only because of the increased availability of public databases but also because of a set of friendly and intuitive data processing and visualization tools, such as Microsoft Power BI, Tableau, R or OpenStat [3]. IBM SPSS V22 was the chosen statistical tool for the sequent analysis.

This article is divided into seven main parts. Besides the present introduction, the status-of-the-art is reviewed in the following section while section 3 summarizes the main descriptive statistics and ANOVA results. Clustering with K-Means and a performance comparison analysis for the survival outcome with Logistic Regression, Decision Trees and Neural Networks are presented in the following two sections. A global explanatory model with Path Analysis technique is unveiled in section 6. We conclude this research in the last section.

2. LITERATURE REVIEW

The Titanic has been an inspiring true story in many different fields of science besides movies, plays, music, paintings, artistic works or memoirs. Zhengyou Zhan describes a vision-based, large-area, simultaneous localization and mapping (SLAM) algorithm that respects the low-overlap imagery constraints typical of underwater vehicles while exploiting the inertial sensor information that is routinely available on such platforms [4]. His technique is based upon solving a sparse system of linear equations coupled with the application of constant-time Kalman updates and it produces consistent covariance estimates suitable for robot planning and data association. For that, a real-world result is reported by this author for a vision-based by using data from a recent survey of the wreck of the RMS Titanic.

The recent foundering of the Costa Concordia in January 2012 in the Tyrrhenian Sea demonstrated that accidents can occur even with ships that are considered masterpieces of modern technology and despite more than 100 years of regulatory and technological progress in maritime safety analysis [5]. Some human and organizational factors were present in the Costa Concordia accident as well as in the foundering of the Titanic a century ago and which can be found in many other maritime accidents over the years. They argue that these factors do not work in isolation but in combination and often together with other underlying factors.

Marko et al. demonstrated, based on a comparison and analysis of both previous historical events, how lessons learned and training methods used in the hazardous marine environments of aircraft carrier operations, as well as the near-solo conditions of technical scuba diving, can be better implemented in managing a large ship at sea [6]. Their study showed that the impact of training that minimizes decisions under stress and enable people to make decisions independently in the face of a loss of communications.

With the use of a machine learning tool, JustNN, Barhoom et al. studied which classifications of passengers have a strong relationship with Titanic survival [7]. Their

analysis seeks to identify characteristics of travelers (cabin class, age and point of departure) and the chance of survival for the disaster. By developing an Artificial Neural Network prediction model (one input layer and two hidden layers), they found a prediction accuracy of 99.28%, where gender, passenger class and sleeping cabin have significant effects on Titanic survivors.

Stolz et al. approached the case of the Titanic with a sociological explanation instead of a statistical one [8]. For instance, individuals knew the rule "women and children first" in the abstract. "Being a man" involves honor and with first-class male passengers, in particular when facing the question of whether it was appropriate to board a lifeboat when there were still women on the ship and how they could do so without being seen as a coward. Certainly, one solution open to such a passenger was to reinterpret the issue by pretending to board the lifeboat in order to take care of his wife (or other ladies). In contrast, being a woman leads to a higher probability of survival. However, this only helps those women who have the necessary resources (information, knowledge and access) that allow them to reach the boat deck earlier than others. At the same time, some first-class female passengers chose to stay on board and then died with their husbands although quantitative analysis shows that this phenomenon must have been rare since 96.5% of the first-class female passengers survived [8].

By using R to build logistic regression, decision tree (supervised learning algorithm), random forest (supervised classification algorithm) and support vector machine (supervised machine learning) approaches, Kakde et al. assessed the accuracy (Table 1) of these four approaches (0.837, 0.826, 0.817 and 0.831, respectively) with the Titanic dataset (861 samples) reaching the following conclusions [9]: (A) The survival rate of females is 79% while the survival rate of males is only 16%, passengers who were travelling in first-class is more likely to survive; (B) Survival rate increases when family size lies from 0 to 3 (but when family size becomes greater than 3, survival rate decrease); (C) The higher the fare, the higher of the survival rate; (D) Logistic regression proved to be the best predictive model. Yet, this same predictors ranking inference has already been announced by Balakumar et al. one year before [10]. Ultimately, Farag and Hassan stated that their decision tree algorithm has accurately predicted 90.01% of the survival of passengers while the Gaussian Naïve Bayes witnessed 92.52% accuracy in prediction [11].

Table 1: *Sensitivity defines the percentage of actual positive which are correctly identified and also measures the proportion of negatives which are properly categorized [12].*

Confusion Matrix		Target Positive	Target Negative		
Model	Positive	a	b	Positive Predictive Value	a/(a+b)
	Negative	c	d	Negative Predictive Value	d/(c+d)
		Sensitivity a/(a+c)	Specificity d/(b+d)	**Accuracy= (a+d)/(a+b+c+d)**	

Meanwhile, Kshirsagar *et al.* achieved good prediction accuracy with logistic regression of about 95% [13]. The values shown in the below confusion matrix are their probability of survival of individuals. For instance, the cell on first column and is of age and the 7th row is sex_male represents the probability of surviving when depending on age and gender as if he is male (8.1%).

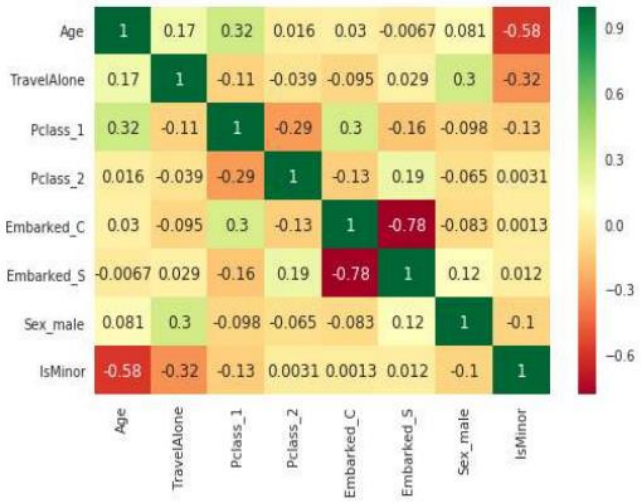


Fig. 1. *Confusion matrix [13].*

Sherlock et al. used extensively K-means clustering with Weka [14]. According to these researchers, some appealing highlights should be stressed: (A) We see that sex of the passenger shows significant clustering around survival chances; (B) Cabin class had significant clustering with the lower tiered cabins showing significant weight towards non-survival. This is shown in figure 2 with a fairly dominant clustering for those in 3rd class that did not survive (and somewhat clear clustering for those in 1st class surviving); (C) Adults (age 20 – 49) were amongst those that perished; (D) The analysis identified that point of embarking was also an indicator of survival rate, although not as strong one (3rd class passengers mostly embark at Southampton).

International Journal of Academic Engineering Research (IJAER)
ISSN: 2643-9085
Vol. 4 Issue 10, October - 2020, Pages: 44-52

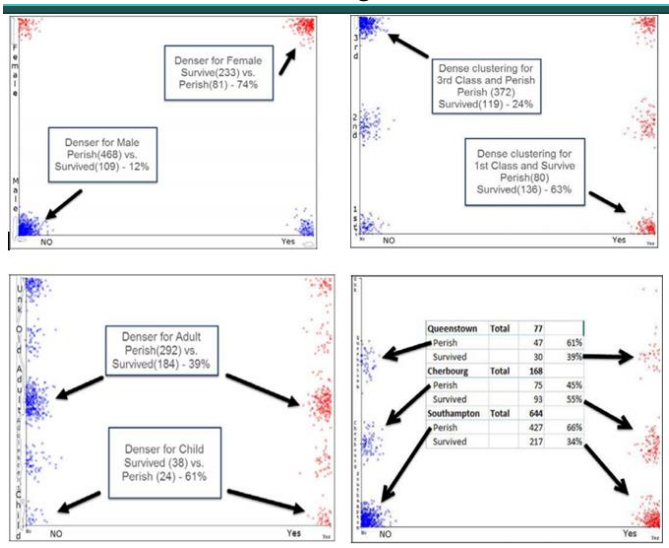


Fig. 2. *Simple K Means: Survived vs. Sex Classification (top left), Survived vs. Class (bottom left), Survived vs. Age Group (top right) and Survived vs Embarked (bottom right).*

In a pedagogical contribution, Asarta et al. extend the traditional three-class tariff employed in the French passenger railway system with the three passenger classes of the ill-fated RMS Titanic [15]. In order for students to reach a deeper understanding of finance concepts, the well-known motion picture Titanic has been used into the teaching of economics to demonstrate modern travel examples of price discrimination.

On the other hand, Perez-Alvaro and Manders explored the process on how a common object (the violin that have been played to calm the passengers while the cruise ship was sinking) has gained prestige both as cultural heritage and allure as a treasure by recognition of various values by different stakeholders [16]: an historical value by the museum, an emotional value by the media and an economic value by the auction market. Based on their words, on 19th October 2013, the auction house Henry Aldridge & Son sold a violin rescued from the Titanic for more than $1.7 million.

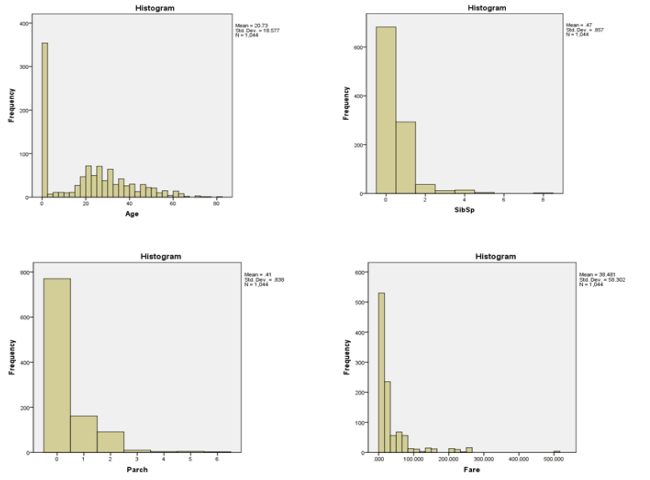
3. DESCRIPTIVE STATISTICS

The given data (kaggle.com source) for this study consists of 1044 records with 10 fields: PassengerID, Survived (0 stands for perished and 1 means lived), Pclass (1st, 2nd and 3rd ticket class, a proxy for socio-economic status), Gender (0 for male and 1 for female), Age (in years), Sibsp (number of siblings and spouses aboard the Titanic), Parch (number of parents and children aboard the Titanic), Fare (ticket price in pounds) and Embarked (C=Cherbourg, Q=Queenstown, S=Southampton). Three derived categorical variables were created for analysis purposes: Agegroup (Seven subsets: 0-10, 10-20 … and over 70 years old), Faregroup (Ten subsets: 0-10, 10-20 … and over 100 units of money) and BabyAdult (0 denotes individuals under 5 years old and 1 otherwise).

The first analysis step concerns the assessment of univariate descriptive indexes such as central tendency measures (mean, mode, median), distribution graphs (histogram), spread measures (variance, standard deviation, range and inter-quartile range), skewness and kurtosis (overall type of distribution). After all, familiarity with the dataset is an asset in any statistical study.

Figure 3 summarizes the main descriptive parameters of the current dataset. The independent T-test confirms different means between survivors and perished people when facing Fare, Parch and Pclass for a 95% level of confidence (the variable Sibsp holds a non-significant p-value of 0.073). Globally, 40% of the 1044 clients survived to this catastrophic event, the average age was 21 years old (young people looking for the American Dream), the number of siblings and parents traveling with each passenger was quite low and the mean ticket fare was 39 pounds.

Since these variables do not follow the Gaussian distribution, the non-parametric Spearman's Rho correlation was computed instead of the conventional Pearson coefficient. The most significant correlations are as follows: (A) Pclass holds a negative natural significant correlation with Fare (-0.774); (B) Sibsp presents a positive correlation with Parch (0.418) leading to the large family association concept; (C) Parch has a positive correlation with Gender (0.265), that is, a predisposition of females travelling with their parents; (D) Gender embraces a positive correlation with Fare (0.236), i.e., females spend more on their travelling ticket rather than men; (E) Age has a negative correlation with Pclass (-0.129) leading to the assumption that older passengers travelled in the upper and middle deck.



	N	Range	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Survived	1044	1	0	1	.40	.490	.411	-1.83
Age	1044	80	0	80	20.73	18.57	.442	-.699
SibSp	1044	8	0	8	.47	.857	3.252	16.64
Parch	1044	6	0	6	.41	.838	2.777	10.37
Fare	1044	512.32	.000	512.329	38.48	58.30	3.817	19.87

Fig. 3. *Positive skewness may be found in these four histograms concerning age, number of siblings (Sibsp), number of parents, brothers and sisters (Parch) and fare cost. In fact, the Kolmogorov-Smirnov and Shapiro-Wilk tests confirm the non-normality assumption of all these distributions (p-value<0.05).*

Developed by Ronald Fisher, one-way analysis of variance (ANOVA) is a technique used to compare means of two or more datasets by using the F non-symmetrical distribution, where the null hypothesis clearly states that samples of two or more groups are drawn from populations with the same mean. This one-way F-test equals the variance ratio calculated among the means between the different regions (explained variation) over the variance within the samples of each region (not explained variation). If both sub-groups means are drawn from populations with the same mean, their variance should be less than the variance of the samples (Table 2). A higher ratio, therefore, implies that samples were drawn from populations with different averages [17]. In the end, the Tukey HSD index allows to test which sub-groups are statistically different from each other in case of a statistical difference stressed by ANOVA.

Table 2: *Sibsp, Parch and Fare hold a statistical difference of the mean regarding the created seven age groups.*

	F	Sig.
Survived	.967	.446
SibSp	3.558	.002
Parch	3.116	.005
Fare	5.681	.000

The Tukey HSD stresses the following group ages towards their statistical average differences among them: (A) Teenagers/twenties/thirties versus children for Sibsp (the number of siblings for 10-40 years old passengers are close to zero); (B) Twenties versus children for Parch (the number of 20-30 years old group travelled without their own families on a search of a better and future life); (C) Fifties/sixties versus teenagers/twenties for Fare (50-70 years old people travelled in first-class cabins).

In a multivariate exploratory context, Principal Component Analysis (PCA) is a technique for reducing the dimensionality of datasets variables, increasing interpretability but at the same time minimizing information loss. Basically, it is often used to make data easy to explore and visualize. Varimax Rotation Method with Kaiser Normalization was applied and after 4 iterations, two components (eigenvalues greater than 1, according to the Scree plot) were achieved for a 54% cumulative variation. The KMO measure of sampling adequacy was 0.544 (values greater than 0.5 indicate the sampling is barely adequate to ensure the use of this factor analysis technique). The Chi-Square Barlett's test of sphericity was 781.2 (statistical significance of 99%, that is, rejection of the null hypothesis that clearly states that the correlation matrix is an identity one), meaning that the two factors than forms the six variables are satisfactory (i.e., there is some scope of reducing the number of dimensions in the present dataset). This outcome also reveals that there is no high correlation or coefficient among the items.

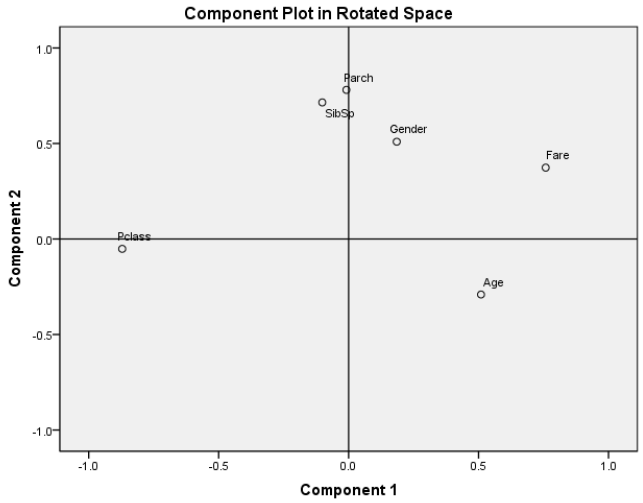


Fig. 4. *As expected, upper-deck (Pclass=1) holds a negative relationship with Age and Fare (PC1) while the family concept (Sibsp and Parch) of PC2 are closely related with Gender variable (lonely males versus family females pattern).*

3.1 CLUSTERING WITH K-MEANS

Clustering is an exploratory data analysis technique used to identify sub-groups in the data such that data points in the same subgroup (cluster) are quite similar while data points in different clusters are exceptionally different. As an alternative of hierarchical clustering, K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping clusters where each data point belongs to only one sub-group. In other words, the less variation we have within clusters, the more homogeneous the data points are within the same cluster.

After the standardization of the values of Gender (male and female), Pclass (1st, 2nd and 3rd cabin class), Agegroup (7 classification sets), Faregroup (10 classification sets), Sibsp (number of siblings) and Parch (number of parents), K-means clustering was accomplished within SPSS. After several trial and error attempts, four clusters were stressed and characterized in Figure 5. The ANOVA confirmed a significant statistical differences between those four clusters based on these six variables for a 99% level of confidence (p-value=.000, high F values) after 10 iterations. As well, the Tukey HSD statistical test, regarding all internal multiple comparisons between variables and clusters, revealed a 100% close statistical level significance with the exception of cluster 1, 3 and 4 for Gender, cluster 2 and 4 for AgeGroup and cluster 1 and 2 for FareGroup.

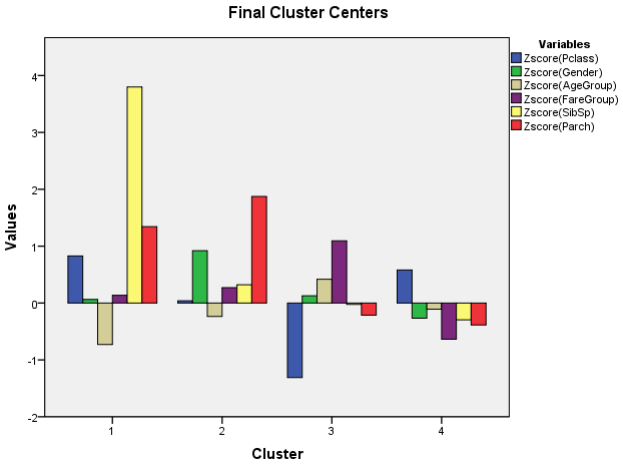


Fig. 5. *Cluster 1 represents younger travelers with large number of siblings and parents in the lower deck (37 individuals), Cluster 2 denotes mainly females that are travelling with their parents (128) in the middle deck, Cluster 3 embodies first-class and older propensity passengers (289) while Cluster 4 is characterized by third-class (male bias) customers with small or no families at all (590).*

4. Prediction: A Performance Comparison Analysis

4.1 Logistic Regression

Income, hours of TV watched or the grade point average of students are ratio variables. However, there are lots of things that we may want to explain that involve dichotomies. There are things in life that the outcome process results in a binary value: Do or do not, did or did not, have or have not, believe or believe not, for or against. If a researcher is interested in explaining how many times someone has been married, the outcome could be 0, 1, 2, 3 or 4, for instance. Another plausible possibility is to determine whether someone has been married or not. The same logic pattern follows if you want to explain why people get involved in car accidents. This measured dependent variable (DV) could be zero accidents, one accident and so on [17]. Again, it also makes sense to ask if in the past five years, a driver has been involved in any serious car accidents. In this case, the outcome could only be yes (1) or no (0).

Logistic regression is the appropriate maximum likelihood estimator (MLE) regression analysis to conduct when the dependent variable (DV) is dichotomous or binary. Used with predictive purposes, logit regression describes data and tries to explain the relationship between one dependent dualistic variable and one or more metric (interval or ratio scale) independent variables (IV). The output is actually the percentage chance that a particular case gets a yes (1) on the DV. To determine this, the probability computation equals $p=ez/((1+ez))$, where e=2.71828 while z denotes the MLE logit regression equation. If z=0 then the yes probability totals 50%. If z<0 then p<0.5, otherwise p>0.5 [17]. Regarding assumptions, there should be no high inter-correlations (multicollinearity) among the predictors.

The Cox & Snell R Square and Nagelkerke R Square indices equal 0.573 and 0.775, respectively, for the given model (Table 3): survived=-2.583-0.26×Pclass(1)+3.166× Pclass(2)+ 1.305×Pclass(3)+5.544×Sex-.036×Age-0.4×Parch. Some novelties may be found in this equation: (A) The women probability of surviving was 5.544 times then men; (B) Since Pclass(1) is not statistical significant, we may clearly state that passengers of class 2 hold a greater likelihood of survival when compared with class 3 by 3.166/1.305=2.42 times; (C) Youngest and lonely passengers hold greater chance of survival. For reference and analogous to the non-pseudo R-squared, the Cox & Snell R Square and the Nagelkerke provide an indication of the amount of variation in the dependent variable explained by the model (from a minimum value of 0 to a maximum of 1).

Table 3: *The total percentage of correct prediction was 92% in a total of 1038 samples (600 and 360 for true positive and true negative, respectively, against 57 and 21 false positive and false negative, correspondingly).*

	B	Sig.	Exp(B)
Pclass		.000	
Pclass(1)	-.260	.540	.771
Pclass(2)	3.166	.000	23.716
Pclass(3)	1.305	.001	3.688
Sex(1)	5.544	.000	255.689
Age	-.036	.001	.965
Parch	-.400	.003	.671
Constant	-2.583	.000	.076

Due to the non-significance of first-class passengers and in a more restrict way (class 1 travelers would not be include in the initial dataset), this logit model would change to survived=-.566-.748×Pclass+5.234×Sex-0.018×Age-0.406× Parch (Cox & Snell R Square and the Nagelkerke R Square equal 0.539 and 0.749, respectively, with an accuracy of 93%). If the new coefficients of Sex, Age and Parch remain quite similar to the previous model, the probability of survival that depends on the middle and lower deck would changes from 2.42 to 0.748. Yet, this latter model should be disregarded since it does not match the true reality of RMS Titanic (by not including first-class passengers).

4.2 Decision Trees

As an exploratory and confirmatory analysis, Decision Trees creates a tree-based classification model by classifying individual cases into groups in order to predict values of a dependent (target) variable based on values of independent (predictor) variables. It may be used for segmentation (patterns) purposes and prediction such as the likelihood that someone will default (or not) his/her down payments when facing a personal loan. By default, SPSS uses CHAID (Chi-

squared Automatic Interaction Detection) algorithm where, at each step, the independent (predictor) chosen variable is the one that has the strongest interaction with the dependent variable.

The root of Figure 6 clearly states that 60.1% passed away and 91.3% of those were men (596). On the contrary, 92.1% of the original 39.9% of the survivors (417) were female (360). In a more depth and detail description, for instance, we may articulate that regarding the middle and lower class (Pclass variable equals 2 and 3), only 5.3% males survived (26) of the total 8.7% for this gender. Yet and for the first-class males, this likelihood increases to 19.1% (31). It is curious to see children and elderlies (females of the lower deck) in the right leaf of this tree with a survival probability of 91.7% (66 out of 72). The teenagers, twenties, thirties and forties females sub-groups of the same third-class status presented a highly 71.4% chance of survival.

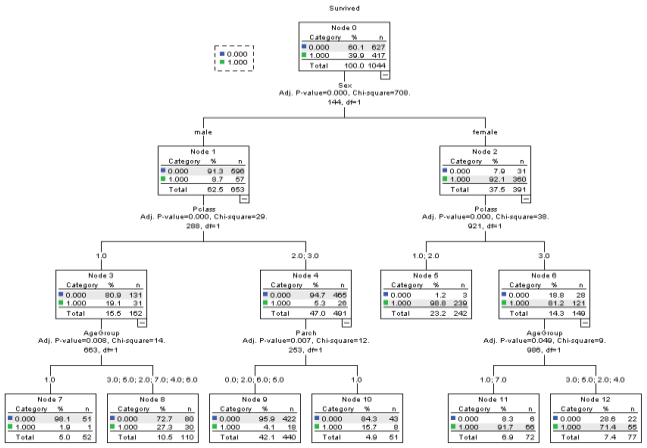


Fig. 6. *Sex, Pclass, AgeGroup and Parch are the most key variables of the dataset regarding the building of this four layer tree.*

Analogous to the previous approach, cross-validation was applied here (table 4), that is, tree models are generated, excluding the data from each subsample in turn. The first tree is based on all of the cases except those in the first sample set, the second tree is based on all of the cases except those in the second sample fold and so on. For each tree, misclassification risk is estimated by applying the tree to the subsample excluded in generating it [18].

Table 4: *The total percentage of correct prediction was 91.6%, a very close value to logit regression.*

Classification (Dependent Variable: Survived; Growing Method: CHAID**)**

Observed	Predicted		
	0	1	Percent Correct
0	596	31	95.1%
1	57	360	86.3%
Overall Percentage	62.5%	37.5%	91.6%

5. NEURAL NETWORKS

Neural Networks (NN) are composed of layers of computational units called neurons, with connections in possible layers. These networks transform data until they classify it as an output [19]. Each neuron multiplies an initial value by some weight, sums results with other values coming into the same neuron, adjusts the resulting number by the neuron's bias and then normalizes the output with an activation function [19]. A key feature is its iterative learning process in which input records are presented to the network one at a time while the weights associated to these input values are adjusted each time. During this learning phase, the network trains by adjusting the weights to predict the correct class label of input samples. This push-back method is called backpropagation and it is the key on how a neural network learns a particular task [20].

NN may handle at once numerical and categorical variables either as dependent or as independent variables in a non-linearity setting. Yet, the researcher will never know anything about the distribution of the outcome variables, meaning that it is not possible to propose any type of test regarding how good the model works. As well, its results highly depend on which kind of cases the investigator gives to the NN learning curve to generate the internal weights for that specific problem.

Popular types of neural networks include identifying faces, self-driving cars and language-driven image generation. For real world business issues, NN are capable to segment customers according to basic characteristics including demographics, economic status, location, purchase patterns and attitude towards a product [21]. In retail and sales, forecasting of sales in stores can be of great advantage too. In medicine, recognizing diseases from various scans is another possibility [22, 23]. From the statistical point of view, NN is an alternative to regression analysis.

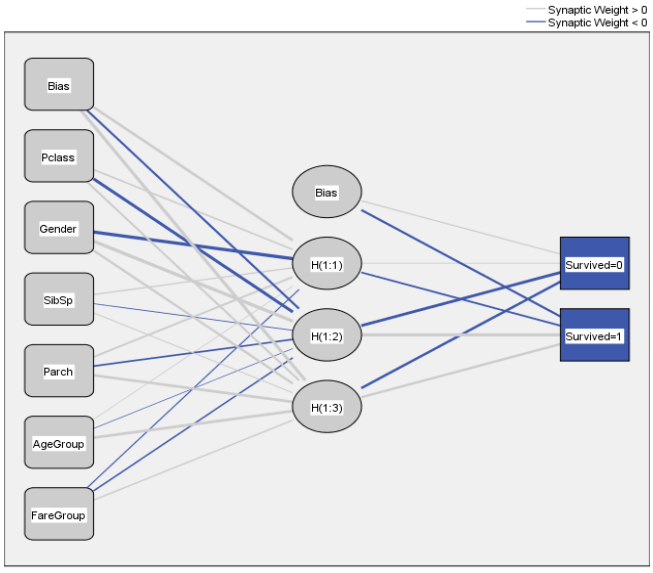


Fig. 7. *Synaptic weights for the different connectors at the learning phase (the darker the line, the stronger the relationship).*

Figure 7 and table 5 presents NN SPSS output for the actual dataset. The outcome, as expected, are only two (survived or not survived) while the input phase is comprised of seven variables. Similar to other statistical approaches, the bias connector represents the error component of the model and, for this particular case, the difficulty to estimate the survivors are more stressed and sensitive than the perish ones. The independent variable normalized importance (their sum up equals one) ranks as follows: Gender (0.443), Sibsp (0.147), Pclass (0.130), Parch (0.121), AgeGroup (0.093) and FareGroup (0.067).

Table 5: *Classification of SPSS NN (827 of the initial dataset was used for learning purposes and the remaining 217 samples for testing) for the dependent variable Survived (0=Perish; 1=Survived). Regarding further quality prognosis measures, 7.9% were considered incorrect predictions during the training phase and 7.8% for the testing one.*

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	479	17	96.6%
	1	48	283	85.5%
	Overall Percent	63.7%	36.3%	92.1%
Testing	0	126	5	96.2%
	1	12	74	86.0%
	Overall Percent	63.6%	36.4%	92.2%

6. PATH ANALYSIS

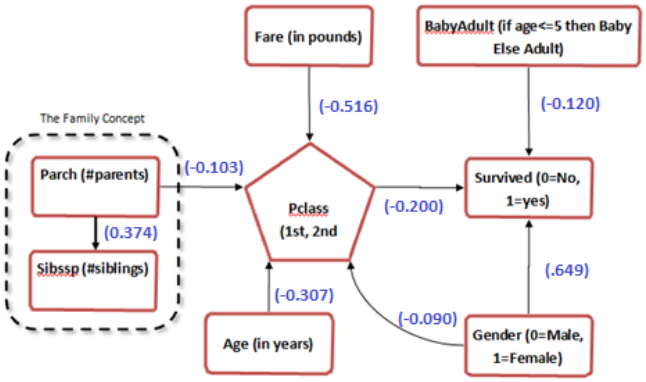
Since the performance of neural networks, logit regression and tree decisions are quite similar, path analysis will be assess here in order to find a possible common statistical model regarding the survival behavior of the RMS Titanic event. Regression can be used to establish the possibility of cause and effect relationships among a set of variables in a one-way cause flow technique [24]. Path analysis and structural equation modeling (SEM) are the two main practices nowadays that fall under this category.

Path analysis allows the testing of a global model including both direct and indirect effects on some outcomes by exploring the inter-correlation among variables. It begins with the investigator developing a path diagram with arrows connecting variables and depicting the cause flow by using the traditional OLS regression standardized parameters. After all, these beta weights reflect the direct explanation effect of an independent variable (IV) on a dependent one (DV). However, if the path coefficient is not statistically significant, one should consider dropping it from the model unless there is a strong theoretical support for its inclusion [17]. Under this semantic, the DVs that are being explained by the IVs are named endogenous variables (have incoming arrows) while the external causes are referred to as exogenous (no arrows going to them). Yet, keep in mind that this technique does not confirm causation in the model. It merely illuminates which of two or more competing models is most consistent with correlation patterns found in the data [25].

Like any typical model, path analysis (an extension of multiple regression) holds four major assumptions: (1) The relationship between the IVs and the DV is linear; (2) The DV errors are independent among themselves whose total mean tends to be zero; (3) No collinearity among the independent variables; (4) The adequate sample size, according to [26], should be 10 times as many cases as parameters.

In a first a priori theory-driven model, the present authors believe that the survival (or not) depends heavily on the Pclass in the first place and on Gender and "being a baby/child" variables on the second position (this happens due to the "women and children first" rule). As expect, the upper, middle or lower deck is subject to Fare, mainly. Certainly, the family concept (honored here by the variables Sibsp and Parch) are not significant for the dependent Survived outcome in a direct way. From the previous sections, we have learnt that the propensity of females is to travel in higher classes rather than men. This same trend may be found for older people and families whose parents are travelling too (higher budget resources). Figure 8 translates this belief where the standardize betas coefficients are presented between brackets, decoding the importance and sign (relationship type) of each variable in the overall model. Perceive that within path analysis, feedback loops between variables are not allowed.

Fig. 8. *The below family concept holds a significant global statistical sub-model for the ANOVA test (p-value=.000), its Durbin-Watson (DW) residuals equals 1.833 (a slightly positive first order correlation on the residuals), no collinearity (VIF=1), the adjusted R-square is 13.9% with an individual significance T-test (p-value=.000) for the constant and the independent Sibssp variable.*



The Pclass regression results of four direct inputs whose key statistical indicators are presented next: Significant Anova (p-value=.000), individual and significant T-test (p-value=.000) for all independent variables, Adjusted R-square of 43%, DW=2.119 (random first-order autocorrelation of the residuals) and VIF=1.13 (no collinearity).

Model Summary

a. Predictors: (Constant), Gender, BabyAdult, Pclass

b. Dependent Variable: Survived

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.720[a]	.518	.517	.341	1.942

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	129.918	3	43.306	373.073	.000[b]
	Residual	120.722	1040	.116		
	Total	250.640	1043			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.659	.059		11.247	.000		
	Pclass	-.116	.013	-.200	-9.103	.000	.964	1.037
	BabyAdult	-.260	.047	-.120	-5.506	.000	.979	1.021
	Gender	.658	.022	.649	29.726	.000	.972	1.029

a. Dependent Variable: Survived

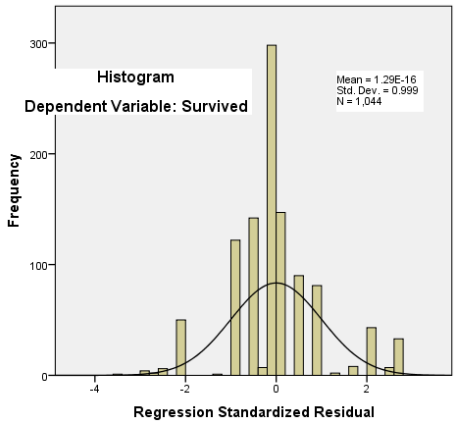


Fig. 9. *Notice that a new explanatory variable was setup, BabyAdult (analogous to Gender, 0 if the age of the passenger is below 5 years old and 1 otherwise), in order to interpret the children first rule.*

If Figure 8 reproduces the direct effects of all variables, indirect effects can be calculate in order to compute the total effect of each variable on the system. Hence, the total indirect effect of Fare, Parch, Age and Gender on Survived equals 20.32%, which corresponds to the sum of the following three parcels (path multiplication rule): (A) Fare -> Pclass -> Survived: -0.516x-0.2=0.1032; (B) Parch -> Pclass -> Survived: -0.103x-0.2=0.0206; Age -> Pclass -> Survived: -0.307x-0.2=0.0614; Gender -> Pclass -> Survived: -0.09x-0.2=0.018. The total direct affect is -0.2-0.12+0.649 =32.9%. This means that indirect effect on Survived equally totals 38.18%, 0.2032/(0.2032+0.329), whereas the remaining corresponds to direct influence.

7. CONCLUSIONS

The mathematical numbers make it all too clear that a rule of "first-class first" far outweighed the principle of "women and children first". The statistical performance of Logistic Regression, Neural Networks and Tree Decisions were quite similar in terms of the confusion matrix. In the end, the present path analysis model only explains 51.7% of the RMS Titanic survival. Definitely, a lot qualitative and quantitative data is missing in this study.

Lifeboat is one of them, for instance. For the first six launched, it only contained passengers from first-class plus crew members to do the work and notorious for being launched at less than half capacity [27]. Testimony has also made it clear that Captain Smith knew before the first lifeboat was launched that the Titanic would sink, but no one among the crew took measures to ensure that all boats were adequately filled [27].

In a personal touch, the DNA factor of each passenger also helped for the survival (or not) outcome. For illustration, the musicians, who traveled in second-class, had the chance to save themselves but all perished by their own will. At last, panic and confusion had been an even greater consideration for the loss of life by many and particularly highlighted in Cameron's blockbuster movie of 1997.

8. REFERENCES

[1] Howells R. (1999). The myth of the Titanic. Palgrave Macmillan, London. https://doi.org/10.1057/9780230510845_3.

[2] kaggle, 2020. https://www.kaggle.com/c/titanic.

[3] Negreiros, J. & Lobo, A. (2019). Microsoft Power BI Desktop for data analytics. International Conference in Management, Economics and Marketing, Vienna, Austria, ISBN 978-80-88203-11-7, pp. 132-149.

[4] Zhang, Z. 1998. Determining the epipolar geometry and its uncertainty: a review. International Journal Computer Vision 27(2), pp. 161-198.

[5] Schröder-Hinrichs, J., Hollnagel, E. & Baldauf, M. (2012). From Titanic to Costa Concordia—a century of lessons not learned. WMU journal of maritime affairs, 11(2), 151-167.

[6] Marko, M., Gilman, L., Vasulingam, S., Miliskievic, M. & Spell, C. (2020). Leadership lessons from the Titanic and Concordia disasters. Journal of Management History, Vol. 26 No. 2, pp. 216-230. https://doi.org/10.1108/JMH-09-2018-0050.

[7] Barhoom, A., Khalil, J., Bassem, A., Musleh, M. & Naser, S. (2019). Predicting Titanic Survivors using Artificial Neural Network. International Journal of Academic Engineering Research (IJAER) 3 (9):8-12.

[8] Stolz, J., Lindemann, A. & Antonietti, J. (2019). Sociological explanation and mixed methods: the example of the Titanic. Qual Quant 53, 1623–1643. https://doi.org/10.1007/s11135-018-00830-0.

[9] Kakde, Y. & Agrawal, S. (2018). Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. International Journal of Computer Applications (0975 – 8887), Volume 179 – No.44.

[10] Balakumar, B., Raviraj, P. & Sivaranjani, K. (2017). Prediction of survivors in Titanic dataset: A comparative study using Machine Learning algorithms (retrieved on July 2020 from https://pdfs.semanticscholar.org/545a/9e5da57058cf08e32eae6b5816839505ac3c.pdf).

[11] Farag, N. & Hassan, G. (2018). Predicting the survivors of the Titanic Kaggle, machine learning from disaster. ICSIE '18: Proceedings of the 7th International Conference on Software and Information Engineering, Pages 32–37, https://doi.org/10.1145/3220267. 3220282.

[12] Kakde, Y. & Agrawal, S. (2018). Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. International Journal of Computer Applications (0975 – 8887), Volume 179 – No.44.

[13] Kshirsagar, V. & Phalke, N. (2019). Titanic Survival Analysis using Logistic Regression. International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 08, p-ISSN: 2395-0072.

[14] Sherlock, J., Muniswamaiah, M., Clarke, L. & Cicoria, S. (2018). Classification of Titanic passenger data and chances of surviving the disaster. Retrieved from https://arxiv.org/ftp/arxiv/papers/1810/1810.09851.pdf.

[15] Asarta, C., Mixon Jr., F. & Upadhyaya, K. (2018). Multiple product qualities in monopoly: Sailing the RMS Titanic into the economics classroom. The Journal of Economic Education, 49:2, 173-179, DOI: 10.1080/00220485.2018.1438862.

[16] Perez-Alvaro, E. & Manders, M. (2016). Playing the values: Sound and vision of the violin of the Titanic. Journal of Cultural Heritage, ISSN: 1296-2074, Vol: 21, Page: 869-875.

[17] Negreiros, J. (2017). Spatial Analysis Techniques Using MyGeoffice®. IGI Global Press, USA, 336p, ISBN10: 1522532706.

[18] IBM (2020). Validation (retrieved from https://www.ibm.com/support/knowledgecenter/ SSLVMB_23.0.0/spss/tree/idh_idd_tree_validation.html

[19] Kwon, Y., Won, J., Kim, B. & Paik, M. (2020). Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Computational Statistics & Data Analysis, 142, 106816.

[20] Shah, J. (2017). Neural Networks for Beginners: Popular Types and Applications: An introduction to neural networks learning (retrieved at https://blog.statsbot.co/neural-networks-for-beginners-d99f2235efca).

[21] Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D. & Yang, J. (2020). Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics, Volume 36, Issue 1, 1 January 2020, Pages 41–48, https://doi.org/10.1093/bioinformatics/ btz477.

[22] Garofalo, A., Rusci, M., Conti, F., Rossi, D. & Benini, L. (2020). PULP-NN: accelerating quantized neural networks on parallel ultra-low-power RISC-V processors. Philosophical Transactions of the Royal Society A, 378(2164), 20190155.

[23] Garg, V. K., Jegelka, S. & Jaakkola, T. (2020). Generalization and representational limits of graph neural networks. arXiv preprint arXiv:2002.06157.

[24] Shen, H., Zheng, S., Xiong, H., Tang, W., Dou, J., & Silverman, H. (2020). Stock market mispricing and firm innovation based on path analysis. Economic Modelling.

[25] Jomnonkwao, S., Uttra, S., & Ratanavaraha, V. (2020). Forecasting road traffic deaths in Thailand: Applications of time-series, curve estimation, multiple linear regression, and path analysis models. Sustainability, 12(1), 395.

[26] Kline, R. (1998). Principles and practice of structural equation modeling. New York, NY: The Guilford Press.

[27] Ithaca College (2017). Demographics of the TITANIC Passengers: Deaths, Survivals, Nationality and Lifeboat Occupancy [retrieved from htt p://www.icyousee.org/titanic.html]