International Journal of Academic Engineering Research (IJAER)
ISSN: 2643-9085
Vol. 3 Issue 9, September – 2019, Pages: 8-12

Predicting Titanic Survivors using Artificial Neural Network

Alaa M. Barhoom, Ahmed J. Khalil, Bassem S. Abu-Nasser, Musleh M. Musleh, Samy S. Abu-Naser

Department of Information Technology,
Faculty of Engineering and Information Technology,
Al-Azhar University, Gaza, Palestine

Abstract *– Although the Titanic disaster happened just over one hundred years ago, it still appeals researchers to understand why some passengers survived while others did not. With the use of a machine learning tool (JustNN) and the provided dataset we study which factors or classifications of passengers have a strong relationship with survival for passengers that took that trip on 15th of April, 1912. The analysis seeks to identify characteristics of passengers - cabin class, age, and point of departure – and that relationship to the chance of survival for the disaster. Furthermore, we developed a model for classifying passengers. The model was trained and tested and we found the accuracy to be more than 99.28%.*

Keywords—learning; titanic; classification; JustNN

1. INTRODUCTION

The Titanic was a ship disaster that on its maiden voyage sunk in the northern Atlantic on 15th of April, 1912, killing 1502 out of 2224 passengers and its crew [1]. While there exists conclusions regarding the cause of the sinking, the analysis of the data on what impacted the survival of passengers continues to this date [2]. The approach taken is utilize a publically available data set from a web site known as Kaggle[3]. We used the JustNN tool for analysis after data review and normalization.

In Machine Learning, the data is mainly divided into two parts — Training and Testing. Training data is for training our model and testing data is to check how well our model performs [4-10]. The split ratio between the train and test data is usually around 70%–30%. Hence, here we have a total of 891 entries for training and 417 entries for testing[2].

2. ARTIFICIAL NEURAL NETWORK

Artificial Neural Networks (ANN) are the pieces of a computing system designed to simulate the way the human brain analyzes and processes information. They are the foundations of Artificial Intelligence (AI) and solve problems that would prove impossible or difficult by human or statistical standards. ANN have self-learning capabilities that enable them to produce better results as more data become available [11-17].

Artificial Neural Networks (ANN) are paving the way for life-changing applications to be developed for use in all sectors of the economy. AI platforms that are built on ANN are disrupting the traditional way of doing things. From translating web pages into other languages to having a virtual assistant order groceries online to conversing with chatbots to solve problems, AI platforms are simplifying transactions and making services accessible to all at negligible costs[18-28].

ANN are built like the human brain, with neuron nodes interconnected like a web. The human brain has hundreds of billions of cells called neurons. Each neuron is made up of a cell body that is responsible for processing information by carrying information towards (inputs) and away (outputs) from the brain. ANN has hundreds or thousands of artificial neurons called processing units, which are interconnected by nodes. These processing units are made up of input and output units. The input units receive various forms and structures of information based on an internal weighting system, and the neural network attempts to learn about the information presented to produce one output report. Just like humans need rules and guidelines to come up with a result or output, ANNs also use a set of learning rules called backpropagation, an abbreviation for backwards propagation of error, to perfect their output results[29-33].

ANN model initially goes through a training phase where it learns to recognize patterns in data, whether visually, aurally, or textually. During this supervised phase, the network compares its actual output produced with what it was meant to produce, i.e., the desired output. The difference between both outcomes is adjusted using backpropagation. This means that the network works backward going from the output unit to the input units to adjust the weight of its connections between the units until the difference between the actual and desired outcome produces the lowest possible error[34-40].

During the training and supervisory stage, the ANN model is taught what to look for and what its output should be, using Yes/No question types with binary numbers. For example, a bank that wants to detect credit card fraud on time may have four input units fed with these questions: (1) Is the transaction in a different country from the user's resident country? (2) Is the website the card is being used at affiliated with companies or countries on the bank's watch list? (3) Is

the transaction amount larger than $2,000? (4) Is the name on the transaction bill the same as the name of the cardholder? The bank wants the "fraud detected" responses to be Yes Yes Yes No, which in binary format would be 1 1 1 0. If the network's actual output is 1 0 1 0, it adjusts its results until it delivers an output that coincides with 1 1 1 0. After training, the computer system can alert the bank of pending fraudulent transactions, saving the bank lots of money[41-50].

Artificial neural networks have been applied in all areas of operations. Email service providers use ANN to detect and delete spam from a user's inbox; asset managers use it to forecast the direction of a company's stock; Credit rating firms use it to improve their credit scoring methods; e-commerce platforms use it to personalize recommendations to their audience; chatbots are developed with ANN for natural language processing; deep learning algorithms use ANN to predict the likelihood of an event; and the list of ANN incorporation goes on across multiple sectors, industries and countries[51-61].

3. METHODOLOGY

3.1 Data Sets

The data we used for our study was provided on the Kaggle website [3]. We were given 891 passenger samples for our training set and their associated labels of whether or not the passenger survived. For each passenger, we were given his/her passenger class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, cabin embarked, and port of embarkation (as shown in Table 1). For the test data, we had 418 samples in the same format. The dataset is not complete, meaning that for several samples, one or many of fields were not available and marked empty (especially in the latter fields – age, fare, cabin, and port). However, all sample points contained at least information about gender and passenger class.

Table 1: Description of each attribute in our dataset

Feature	type	Description
PassengerId	int	Id
Survived	int	Survival (0=No; 1=Yes)
Pclass	int	Passenger Class
Name	object	Name
Sex	object	Sex
Age	float	Age
SibSp	int	Number of Siblings/Spouses Aboard
Parch	int	Number of Parents/Children Aboard
Ticket	object	Ticket Number
Fare	float	Passenger Fare
Cabin	object	Cabin number
Embarked	object	(C=Cherbourg; Q=Queenstown; S=Southampton)

3.2 Feature Engineering

Feature engineering is measuring the impact of each feature on the output. But the more important thing is that it is not just about using existing features, it is about creating new ones that can make a significant improvement in our output. Andrew Ng said, "*Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning is basically feature engineering.*" We will go through each feature we are using so that we can understand how to use existing features and how to create new ones[1,2].

3.3 Passenger Class

It is obvious that the class of passenger is directly proportional to survival rate. If the importance of a person is more than others, they'll get out of the disaster first. And our data tells the same story. 63% of people survived from Class 1. Therefore, this feature is definitely impactful. Data in Pclass column is complete hence no need to manipulate it.

3.4 Sex

Sex is again important and directly proportional to survival rate. Female and children were saved first during this tragedy. We can see that 74% of all females were saved and only 18% of all males were saved. Again, this will impact our outcome.

3.5 Family Size

Next two columns are SibSp and Parch, which are not directly related to whether a person has survived or not. That is where the idea of creating a new feature came in. For each row/passenger, we will determine his/her family size by adding SibSp + Parch + 1(himself/herself). Family size differs from a minimum of 1 to a maximum of 11, where the family size of 4 having the highest survival rate of 72%.

It seems to have a good effect on our prediction but let's go further and categorizes people to check whether they are alone in this ship or not. And after looking through it too, it seems to have a considerable impact on our output.

3.6 Embarked

From which place a passenger embarked has something to do with survival (not always). So, let's take a look. In this column, there are plenty of Not Available (NAs). To deal with it, we are going to replace NAs with 'S' because it is the most occurred value.

3.7 Fare

There are missing data in this column as well. We cannot deal with every feature in the same way. To fix the issue here, we are going to take the median value of the entire column. When you cut with qcut, the bins will be chosen so that you

International Journal of Academic Engineering Research (IJAER)
ISSN: 2643-9085
Vol. 3 Issue 9, September – 2019, Pages: 8-12

have the same number of records in each bin (equal parts). Looking through the output, it is considerable.

3.8 Age

Age has some missing values. We will fill it with random numbers between (average age minus average standard deviation) and (average age plus average standard deviation). After that, we will group it in the set of 5. It has a good impact as well.

3.9 Name

This one is a little tricky. From the name, we have to retrieve the title associated with that name, i.e. Mr or Captain. First, we get the title from the name and store them in a new list called *title*. After that, we cleaned the list by narrowing down to common titles.

3.10Mapping

After cleaning our features, they are now ready to use. However; there is one more step before we feed our data to JustNN tool. The thing about ML algorithms is that they only take numerical values and not strings. So, we have to map our data to numerical values and convert the columns to the integer data type.

3.11ANN Model Architecture

The resulted architecture ANN model for Titanic survivors is shown in Figure 1. It consists of one impute layer, two hidden layers, and one output layer.

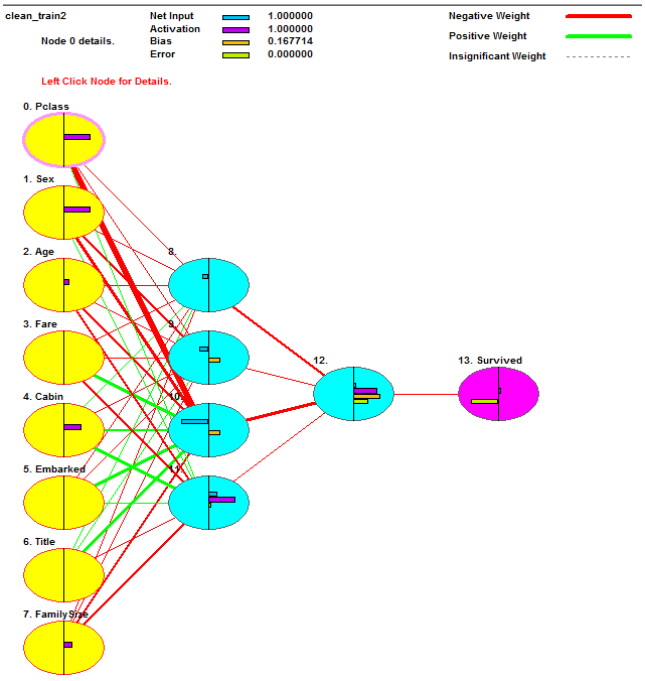

Figure1: ANN Model Architecture

3.12ANN Model Validation

Our ANN model was able to predict Titanic survivors with 99.28% accuracy, with about 0.005 errors as seen in Figure 2. Furthermore, The Model showed that the most effective factors in Titanic survivors are Sex, Pclass, and Cabin. More details are shown in Figure 3.

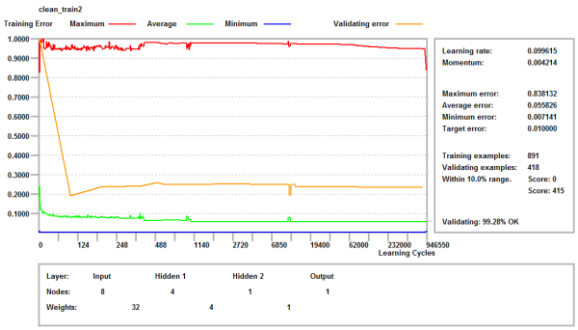

Figure 2: ANN Model training and validation

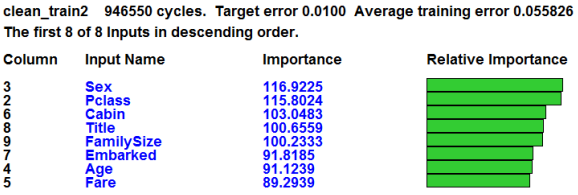

Figure 3: Effective of input valuables to Titanic Survivors

4. CONCLUSION

In this paper, Titanic survivors has been investigated using Artificial Neural Network model to predict Titanic survivors and analysis using JustNN Tool was used to determine the effect of input variables based on the data in the literature. An ANN model gives a very good prediction (99.28%) in comparison with the original data sets of [3]. Sex, Pclass, and Cabin have significant effects on Titanic survivors for the present problem.

REFERENCES

1. Abu Naser, S. S. (2012). "Predicting learners performance using artificial neural networks in linear programming intelligent tutoring system." International Journal of Artificial Intelligence & Applications 3(2): 65.

2. Abu-Naser, S., et al. (1995). "& Beattie, GA (2000)." Expert system methodologies and applications-a decade review from: 9-26.
3. Afana, M., et al. (2018). "Artificial Neural Network for Forecasting Car Mileage per Gallon in the City." International Journal of Advanced Science and Technology 124: 51-59.
4. Ahmed, A., et al. (2019). "Knowledge-Based Systems Survey." International Journal of Academic Engineering Research (IJAER) 3(7): 1-22.
5. Akkila, A. N., et al. (2019). "Survey of Intelligent Tutoring Systems up to the end of 2017." International
6. Al-Ani, I. A. R., et al. (2007). "Water pollution and its effects on human health in rural areas of Faisalabad." Journal of Environmental Science and Technology 5(5): 1-17.
7. Alghoul, A., et al. (2018). "Email Classification Using Artificial Neural Network." International Journal of Academic Engineering Research (IJAER) 2(11): 8-14.
8. Alkronz, E. S., et al. (2019). "Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation Neural Network." International Journal of Academic and Applied Research (IJAAR) 3(2): 1-8.
9. Almasri, A., et al. (2019). "Intelligent Tutoring Systems Survey for the Period 2000-2018." International Journal of Academic Engineering Research (IJAER) 3(5): 21-37.
10. Al-Massri, R., et al. (2018). "Classification Prediction of SBRCTs Cancers Using Artificial Neural Network." International Journal of Academic Engineering Research (IJAER) 2(11): 1-7.
11. Al-Mubayyed, O. M., et al. (2019). "Predicting Overall Car Performance Using Artificial Neural Network." International Journal of Academic and Applied Research (IJAAR) 3(1): 1-5.
12. Al-Shawwa, M. and S. S. Abu-Naser (2019). "Predicting Birth Weight Using Artificial Neural Network." International Journal of Academic Health and Medical Research (IJAHMR) 3(1): 9-14.
13. Al-Shawwa, M. and S. S. Abu-Naser (2019). "Predicting Effect of Oxygen Consumption of Thylakoid Membranes (Chloroplasts) from Spinach after Inhibition Using Artificial Neural Network." International Journal of Academic Engineering Research (IJAER) 3(2): 15-20.
14. Al-Shawwa, M., et al. (2018). "Predicting Temperature and Humidity in the Surrounding Environment Using Artificial Neural Network." International Journal of Academic Pedagogical Research (IJAPR) 2(9): 1-6.
15. Anderson, J., et al. (2005). "Adaptation of Problem Presentation and Feedback in an Intelligent Mathematics Tutor." Information Technology Journal 5(5): 167-207.
16. Ashqar, B. A. M. and S. S. Abu-Naser (2019). "Identifying Images of Invasive Hydrangea Using Pre-Trained Deep Convolutional Neural Networks." International Journal of Academic Engineering Research (IJAER) 3(3): 28-36.
17. Ashqar, B. A. M. and S. S. Abu-Naser (2019). "Image-Based Tomato Leaves Diseases Detection Using Deep Learning." International Journal of Academic Engineering Research (IJAER) 2(12): 10-16.
18. Ashqar, B. A., et al. (2019). "Plant Seedlings Classification Using Deep Learning." International Journal of Academic Information Systems Research (IJAISR) 3(1): 7-14.
19. Atallah, R. R. (2014). "Professor Samy S." Abu Naser, Data Mining Techniques in Higher Education an Empirical Study for the University of Palestine, IJMER 4(4): 48-52.
20. Atallah, R. R. and S. S. Abu Naser (2014). "Data mining techniques in higher education an empirical study for the university of Palestine." International Journal Of Modern Engineering Research (IJMER) 4(4): 48-52.
21. Baker, J., et al. "& Heller, R.(1996)." Information Visualization. Information Technology Journal 7(2).
22. Baker, J., et al. (1996). "Information Visualization." Information Technology Journal 7(2): pp: 403-404.
23. Chen, R.-S., et al. (2008). "Evaluating structural equation models with unobservable variables and measurement error." Information Technology Journal 10(2): 1055-1060.
24. Dalffa, M. A., et al. (2019). "Tic-Tac-Toe Learning Using Artificial Neural Networks." International Journal of Engineering and Information Systems (IJEAIS) 3(2): 9-19.
25. El_Jerjawi, N. S. and S. S. Abu-Naser (2018). "Diabetes Prediction Using Artificial Neural Network." International Journal of Advanced Science and Technology 121: 55-64.
26. El-Khatib, M. J., et al. (2019). "Glass Classification Using Artificial Neural Network." International Journal of Academic Pedagogical Research (IJAPR) 3(2): 25-31.
27. Elzamly, A., et al. (2015). "Classification of Software Risks with Discriminant Analysis Techniques in Software planning Development Process." International Journal of Advanced Science and Technology 81: 35-48.
28. Elzamly, A., et al. (2015). "Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods." Int. J. Adv. Inf. Sci. Technol 38(38): 108-115.
29. Elzamly, A., et al. (2016). "A New Conceptual Framework Modelling for Cloud Computing Risk Management in Banking Organizations." International Journal of Grid and Distributed Computing 9(9): 137-154.

30. Elzamly, A., et al. (2017). "Predicting Critical Cloud Computing Security Issues using Artificial Neural Network (ANNs) Algorithms in Banking Organizations." International Journal of Information Technology and Electrical Engineering 6(2): 40-45.
31. Elzamly, A., et al. (2019). "Critical Cloud Computing Risks for Banking Organizations: Issues and Challenges." Religación. Revista de Ciencias Sociales y Humanidades 4(18).
32. Heriz, H. H., et al. (2018). "English Alphabet Prediction Using Artificial Neural Networks." International Journal of Academic Pedagogical Research (IJAPR) 2(11): 8-14.
33. Hissi, H. E.-., et al. (2008). "Medical Informatics: Computer Applications in Health Care and Biomedicine." Journal of Artificial Intelligence 3(4): 78-85.
34. Jamala, M. N. and S. S. Abu-Naser (2018). "Predicting MPG for Automobile Using Artificial Neural Network Analysis." International Journal of Academic Information Systems Research (IJAISR) 2(10): 5-21.
35. Kashf, D. W. A., et al. (2018). "Predicting DNA Lung Cancer using Artificial Neural Network." International Journal of Academic Pedagogical Research (IJAPR) 2(10): 6-13.
36. Kashkash, K., et al. (2005). "Expert system methodologies and applications-a decade review from 1995 to 2004." Journal of Artificial Intelligence 1(2): 9-26.
37. Li, L., et al. (2011). "Hybrid Quantum-inspired genetic algorithm for extracting association rule in data mining." Information Technology Journal 12(4): 1437-1441.
38. Marouf, A. and S. S. Abu-Naser (2018). "Predicting Antibiotic Susceptibility Using Artificial Neural Network." International Journal of Academic Pedagogical Research (IJAPR) 2(10): 1-5.
39. Masri, N., et al. (2019). "Survey of Rule-Based Systems." International Journal of Academic Information Systems Research (IJAISR) 3(7): 1-23.
40. Metwally, N. F., et al. (2018). "Diagnosis of Hepatitis Virus Using Artificial Neural Network." International Journal of Academic Pedagogical Research (IJAPR) 2(11): 1-7.
41. Nasser, I. M. and S. S. Abu-Naser (2019). "Artificial Neural Network for Predicting Animals Category." International Journal of Academic and Applied Research (IJAAR) 3(2): 18-24.
42. Nasser, I. M. and S. S. Abu-Naser (2019). "Lung Cancer Detection Using Artificial Neural Network." International Journal of Engineering and Information Systems (IJEAIS) 3(3): 17-23.
43. Nasser, I. M. and S. S. Abu-Naser (2019). "Predicting Books' Overall Rating Using Artificial Neural Network." International Journal of Academic Engineering Research (IJAER) 3(8): 11-17.
44. Nasser, I. M. and S. S. Abu-Naser (2019). "Predicting Tumor Category Using Artificial Neural Networks." International Journal of Academic Health and Medical Research (IJAHMR) 3(2): 1-7.
45. Nasser, I. M., et al. (2019). "A Proposed Artificial Neural Network for Predicting Movies Rates Category." International Journal of Academic Engineering Research (IJAER) 3(2): 21-25.
46. Nasser, I. M., et al. (2019). "Artificial Neural Network for Diagnose Autism Spectrum Disorder." International Journal of Academic Information Systems Research (IJAISR) 3(2): 27-32.
47. Ng, S., et al. (2010). "Ad hoc networks based on rough set distance learning method." Information Technology Journal 10(9): 239-251.
48. Owaied, H. H., et al. (2009). "Using rules to support case-based reasoning for harmonizing melodies." Journal of Applied Sciences 11(14): pp: 31-41.
49. Sadek, R. M., et al. (2019). "Parkinson's Disease Prediction Using Artificial Neural Network." International Journal of Academic Health and Medical Research (IJAHMR) 3(1): 1-8.
50. Salah, M., et al. (2018). "Predicting Medical Expenses Using Artificial Neural Network." International Journal of Engineering and Information Systems (IJEAIS) 2(20): 11-17.
51. Sulisel, O., et al. (2005). "Growth and Maturity of Intelligent Tutoring Systems." Information Technology Journal 7(7): 9-37.
52. Zaqout, I., et al. (2015). "Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology." International Journal of Hybrid Information Technology 8(2): 221-228.
53. Zaqout, I., & Al-Hanjori, M. (2005). An improved technique for face recognition applications. Information and Learning Science, 119 (9/10), 529-544.
54. Zaqout, I. S. (2012). Printed Arabic Characters Classification Using A Statistical Approach. International Journal Of Computers & Technology, 3 (1), 1-5.
55. Zaqout, I. (2019). Diagnosis of skin lesions based on dermoscopic images using image processing techniques. Pattern Recognition-Selected Methods and Applications.
56. Zaqout, I., Zainuddin, R., & Baba, S. (2004). Human face detection in color images. Advances in Complex Systems, 7 (03n04), 369-383.

57. Zaqout, I. S. (2005). An integrated approach for detecting human faces in color images. Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Malaya.
58. Zaqout, I. (2011). A Statistical Approach For Latin Handwritten Digit Recognition. IJACSA Editorial.
59. Zaqout, I. S. (2017). An efficient block-based algorithm for hair removal in dermoscopic images. Компьютерная оптика, 41 (4).
60. Zaqout, I., Zainuddin, R., & Baba, S. (2005). Pixel-based skin color detection technique, Machine Graphics and Vision. 14 (1), 61.
61. https://www.kaggle.com/c/titanic
62. https://www.kaggle.com/gokultalele/titanic-machine-learning-from-disaster
63. https://www.kaggle.com/biswajee/titanic-dataset