PREDICTION OF TITANIC DATA ANALYSIS USING LOGISTIC REGRESSION COMPARED WITH NAIVE BAYES FOR BETTER ACCURACY

L.Sreenivasulu[1], V. Chandrasekar[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The main aim of the research is to predict the survival of passengers on the titanic data analysis using Logistic Regression (LR) over Naive Bayes (NB) machine learning algorithm. **Materials** and **Methods:** Logistic Regression and Naive Bayes are implemented in this research work. Sample size is calculated using G -power software and determined as 10 per group with pretest G -power 80%, threshold 0.05% and CI 95%. **Result:** Logistic Regression provides a higher of 92.94% compared to Naive Bayes algorithm with 88.95% in predicting titanic data analysis. There is a significant difference between two groups with a significance value of 0.004 ($p<0.05$). **Conclusion:** Logistic Regression algorithm predicts better information about titanic data analysis than Naive Bayes algorithm.

Keywords: Titanic, Novel Logistic Regression Algorithm, Naive Bayes Algorithm, Classification, Machine Learning.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

[2*]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

1. Introduction

The Titanic disaster occurred about 100 years back but still it attracts the researchers to understand and study how some passengers survived and others perished. In this work, the characteristics of the passengers will be identified and the relationship of survival chance from the disaster is found. Feature engineering techniques will be performed (Haque, Shivaprasad, and Guruprasad 2021). The aim of this study is to get as reliable results as possible from the raw and missing data by using machine learning and feature engineering methods. Therefore one of the most popular datasets in data science, Titanic is used. This dataset records various features of passengers on the Titanic, including who survived and who didn't survive (Frey, Savage, and Torgler, n.d.). It is realized that some missing and uncorrelated features decreased the performance of prediction, was taken from (Gupta, Sharma, and Bouza Herreras 2018). In efforts to study the Titanic passengers;kaggle,a popular data science website, assembled information about each passenger back in the days of the Titanic into a dataset,and made it available. The prediction and efficiency of these algorithms depend greatly on data analysis and the model (Shekhar, Arora, and Sharma 2021). The paper presents an implementation which combines the benefits of feature selection and machine learning to accurately select and distinguish characteristics of passengers age,class,cabin,and port (Farag and Hassan 2018). Bayes theorem can be used to make predictions based on prior knowledge and current evidence. (Theus and Urbanek 2008).

In the last three years, Google scholar identified almost 13,300 research articles on titanic data analysis prediction using machine learning. In this paper survival of passengers is figured out using various machine learning techniques namely logistic regression and naive bayes. The main focus of this work is to differentiate machine learning algorithms to analyze the survival rate of travelers based on the accuracy (Shetty, Pallavi, and Ramyashree 2018). The entire international community was deeply shocked and saddened after hearing the news of this sensational disaster which resulted in improved ship safety legislation ("Prediction of Survivors in the Titanic Cruise" 2019)). Her architect, Thomas Andrews died in the disaster. An eye-opening observation that came forth from the sinking of Titanic is the fact that some individuals had a better chance at surviving than the others (Dasgupta et al. 2021). With accumulating evidence, the prediction is changed. In this work, the characteristics of the passengers will be identified and the relationship of survival chances from the disaster is found. In technical terms, the prediction is the posterior probability that investigators are interested in it,and it was taken from (Zhang 2016). Classification is done using the Logistic Regression learning classification algorithm using two classes which are survived and not survived. Python has been used for its implementation; clustering is performed using machine learning algorithms as implemented by (Farag and Hassan 2018). In the Titanic disaster over the years, data of surviving as well as deceased passengers has been collected. The dataset is publicly available on a website called Kaggle.com (Durmuş and Güneri 2020). This dataset has been studied and analyzed using various machine learning algorithms like Logistic Regression and Naive Bayes Various languages and tools are used to implement these algorithms including Weka, Python, R, Java (D. Chatterjee and Chatterjee, n.d.).

Our institution is passionate about high quality evidence based research and has excelled in various domains (Vickram et al. 2022; Bharathiraja et al. 2022; Kale et al. 2022; Sumathy et al. 2022; Thanigaivel et al. 2022; Ram et al. 2022; Jothi et al. 2022; Anupong et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Palanisamy et al. 2022). This project involves implementation of data analytics and machine learning. The data analysis will be done on applied algorithms and accuracy will be checked. Based on the performance of the mentioned algorithms; Logistic Regression and Naive Bayes, in this paper,Logistic Regression proved to be the best algorithm by outperforming other implemented algorithms for the Titanic classification problem since it achieved the highest accuracy. Also, the values for Logistic Regression appear to be the highest as compared with Naive Bayes algorithms. On the basis of accuracy the best performing model suggested for the suvival predictions is Novel Logistic Regression Algorithm.

2. Materials and Methods

This study setting was done in the Soft Computing Laboratory, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The number of required samples in research are two in which group 1 is Logistic Regression compared with group 2 of Naive Bayes Algorithm. The samples were taken from the device and iterated 10 times to get desired accuracy with G power 80%, threshold 0.05% and CI 95% A dataset consisting of a collection of titanic data analysis was downloaded from Github repository (venky n.d.; datasciencedojo n.d.)).

Logistic Regression

Logistic regression is the technique which works best when the dependent variable is dichotomous (binary or categorical). The data description and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables is done with the help of

Pseudocode for Logistic Regression

Step1: Import packages.

Step2: Create an input dataset.

Step3: Analyze the size of the taken input data.

Step4: Split the datasets for testing and training the dataset.

Step5: Apply Logistic Regression algorithm.

Step6: Predict the results.

Naive Bayes

Naive Bayes, which is known as an effective inductive learning algorithm, achieves efficient and fast classification in machine learning applications. The algorithm is based on Bayes theorem assuming all features are independent given the value of the class variable. This is conditional independence assumption and true in real world applications. Due to this assumption Naive Bayes performs well on high dimensional and complex datasets.

Pseudocode for Naive Bayes

Step1: Import packages.

Step2: Create an input dataset.

Step3: Analyze the size of the taken input data.

Step4: Split the datasets for testing and training the dataset.

Step5: Apply Naive Bayes algorithm.

Step6: Predict the results.

Recall that the testing setup includes both hardware and software configuration choices. The laptop has
an Intel hp 5th generation CPU with 12GB of RAM, an x86-based processor, a 64-bit operating system, and a hard drive. Currently, the software runs on Windows 10 and is programmed in Python. Once the program is finished, the accuracy value will appear. Procedure: Wi-Fi laptop connected. Chrome to Google Collaboratory search Write the code in Python. Run the code. To save the file, upload it to the disc, and create a folder for it. Log in using the ID from the message. Run the code to output the accuracy and graph.

Statical Analysis

SPSS is a software tool used for statistics analysis. The proposed system utilized 10 iterations logistic regression. It is used to solve binary classification problems, some of the real life examples are spam detection- predicting if an email is spam or not, health-Predicting if a given mass of tissue is benign or malignant, marketing-predicting if a given user will buy an insurance product or not.
for each group with predicted accuracy noted and analyzed. Independent samples t-test was done to obtain significance between two groups. The prediction is to perform exploratory data analytics to mine various information in the dataset available and to know the effect of each field on survival of passengers by applying analytics between every field of the dataset with the "Survival" field (Foster 2004).

3. Result

Table 1 shows the accuracy value of iteration of Logistic Regression and Naive Bayes. Table 2 represents the Group statistics results which depicts Logistic Regression with mean accuracy of 92.94%, and standard deviation is 1.94. Naive Bayes has a mean accuracy of 88.95% and standard deviation is 1.81. Proposed Logistic Regression algorithm provides better performance compared to the Naive Bayes algorithm. Table 3 shows the independent samples T-test value for Logistic Regression and Naive Bayes with Mean difference as 3.99, std Error Difference as 0.84. Significance value is observed as 0.004 ($p<0.05$).

Figure 1 shows the bar graph comparison of mean of accuracy on Logistic Regression and Naive Bayes algorithm. Mean accuracy of Logistic Regression is 92.94% and Naive Bayes 88.95%. The Logistic Regression looks to perform significantly better than Naive Bayes.

4. Discussion

In this study, predicting titanic data analysis using the Logistic Regression algorithm has significantly higher accuracy, approximately 92.94% in comparison to Naive Bayes 88.95%. Logistic Regression appears to produce more consistent results with minimal standard deviation.

The similar findings of the paper (Nair et al. 2017) had an accuracy of 91% with Naive Bayes which was used to predict the titanic data analysis. The proposed work of (T. Chatterjee 2018) reported Naive Bayes has 78% accuracy which is used to predict the accuracy of titanic data analysis. The work proposed by (Nair et al. 2017) shows the Logistic Regression has a better accuracy of 93%. Naive Bayes is a parameter to measure accuracy of titanic data analysis which is used in both traditional and modern methods. In the same way (Whitley 2015) had accuracy of 80.2%

with Logistic Regression and (Singh, Nagpal, and Sehgal 2020) had accuracy of 76.79% with Naive Bayes. So Logistic Regression performs better with a combination of other machine learning algorithms.

The limitation of this research is that it cannot give appropriate results for smaller data. In this model it is not able to consider all given feature variable parameters for training. The future scope of proposed work will be prediction of titanic data analysis based on classification using class labels for lesser time complexity.

5. Conclusion

In this research, Titanic Data Analysis is performed to analyze survival of people by using a dataset for Novel Logistic Regression and Naive Bayes. The accuracy value of the Logistic Regression is 92.94% whereas the accuracy value of Naive Bayes is 88.95%.The prediction of Titanic Data Analysis the survival of male and female accuracy using Logistic Regression appears to be better than Naive Bayes.

Declaration

Conflict of Interests

No conflict of interests in this manuscript.

Authors Contribution

Author LS was involved in data collection, data analysis, manuscript writing. Author VC was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, are providing the necessary Infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Inoaura Technologies, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

6. References

Anupong, Wongchai, Lin Yi-Chia, Mukta Jagdish, Ravi Kumar, P. D. Selvam, R. Saravanakumar, and Dharmesh Dhabliya. 2022. "Hybrid Distributed Energy Sources Providing Climate Security to the Agriculture Environment and Enhancing the Yield." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102142.

Bharathiraja, B., J. Jayamuthunagai, R. Sreejith, J. Iyyappan, and R. Praveenkumar. 2022. "Techno Economic Analysis of Malic Acid Production Using Crude Glycerol Derived from Waste Cooking Oil." *Bioresource Technology* 351 (May): 126956.

Chatterjee, Devlina, and Anindya Chatterjee. n.d. "Binary Logistic Regression Using Survival Analysis." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1672759.

Chatterjee, Tryambak. 2018. "Prediction of Survivors in Titanic Dataset: A Comparative Study Using Machine Learning Algorithms." *International Journal of Emerging Research in Management and Technology*. https://doi.org/10.23956/ijermt.v6i6.236.

Dasgupta, Anasuya, Ved Prakash Mishra, Sanjiv Jha, Bhopendra Singh, and Vinod Kumar Shukla. 2021. "Predicting the Likelihood of Survival of Titanic's Passengers by Machine Learning." *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. https://doi.org/10.1109/iccike51210.2021.9410757.

datasciencedojo. n.d. "Datasets/titanic.csv at Master · Datasciencedojo/datasets." Accessed October 9, 2021. https://github.com/datasciencedojo/datasets.

Durmuş, Burcu, and Öznur İşçi Güneri. 2020. "Analysis and Detection of Titanic Survivors Using Generalized Linear Models and Decision Tree Algorithm." *International Journal of Applied Mathematics Electronics and Computers*. https://doi.org/10.18100/ijamec.785297.

Farag, Nadine, and Ghada Hassan. 2018. "Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster." *Proceedings of the 7th International Conference on Software and Information Engineering - ICSIE '18*. https://doi.org/10.1145/3220267.3220282.

Foster, John Wilson. 2004. "The Titanic Disaster: Stead, Ships and the Supernatural." *The Titanic in Myth and Memory*. https://doi.org/10.5040/9780755604845.ch-003.

Frey, Bruno S., David A. Savage, and Benno Torgler. n.d. "Surviving the Titanic Disaster: Economic, Natural and Social Determinants." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1347962.

Gupta, Kshitiz, Prayas Sharma, and Carlos N. Bouza Herreras. 2018. "Surviving the Titanic Tragedy: A Sociological Study Using Machine Learning Models." *Suma de*

Negocios. https://doi.org/10.14349/sumneg/2018.v9.n20.a2.

Haque, Md Arfinul, G. Shivaprasad, and G. Guruprasad. 2021. "Passenger Data Analysis of Titanic Using Machine Learning Approach in the Context of Chances of Surviving the Disaster." *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899x/1065/1/012042.

Jothi, K. Jeeva, K. Jeeva Jothi, S. Balachandran, K. Mohanraj, N. Prakash, A. Subhasri, P. Santhana Gopala Krishnan, and K. Palanivelu. 2022. "Fabrications of Hybrid Polyurethane-Pd Doped ZrO2 Smart Carriers for Self-Healing High Corrosion Protective Coatings." *Environmental Research*. https://doi.org/10.1016/j.envres.2022.113095.

Kale, Vaibhav Namdev, J. Rajesh, T. Maiyalagan, Chang Woo Lee, and R. M. Gnanamuthu. 2022. "Fabrication of Ni–Mg–Ag Alloy Electrodeposited Material on the Aluminium Surface Using Anodizing Technique and Their Enhanced Corrosion Resistance for Engineering Application." *Materials Chemistry and Physics*. https://doi.org/10.1016/j.matchemphys.2022.125900.

Nair, Dr Prabha Shreeraj, Prabha Shreeraj Nair, Tulsiramji Gayakwade Patil College of Engineering and Technology, and Nagpu. 2017. "Analyzing Titanic Disaster Using Machine Learning Algorithms." *International Journal of Trend in Scientific Research and Development*. https://doi.org/10.31142/ijtsrd7003.

Palanisamy, Rajkumar, Diwakar Karuppiah, Subadevi Rengapillai, Mozaffar Abdollahifar, Gnanamuthu Ramasamy, Fu-Ming Wang, Wei-Ren Liu, Kumar Ponnuchamy, Joongpyo Shim, and Sivakumar Marimuthu. 2022. "A Reign of Bio-Mass Derived Carbon with the Synergy of Energy Storage and Biomedical Applications." *Journal of Energy Storage*. https://doi.org/10.1016/j.est.2022.104422.

"Prediction of Survivors in the Titanic Cruise." 2019. *International Journal of Recent Technology and Engineering*. https://doi.org/10.35940/ijrte.c4408.098319.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102102.

Shekhar, Shashank, Deepak Arora, and Puneet Sharma. 2021. "Classifying Titanic Passenger Data and Prediction of Survival from Disaster." *Advances in Information Communication Technology and Computing*. https://doi.org/10.1007/978-981-15-5421-6_18.

Shetty, Jyothi, S. Pallavi, and Ramyashree. 2018. "Predicting the Survival Rate of Titanic Disaster Using Machine Learning Approaches." *2018 4th International Conference for Convergence in Technology (I2CT)*. https://doi.org/10.1109/i2ct42659.2018.9058280.

Singh, Karman, Renuka Nagpal, and Rajni Sehgal. 2020. "Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset." *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. https://doi.org/10.1109/confluence47617.2020.9057955.

Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Kumar Shukla, and Stephen Jeswinde Nuagah. 2022. "Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System." *Computational Intelligence and Neuroscience* 2022 (February): 5906797.

Thanigaivel, Sundaram, Sundaram Vickram, Nibedita Dey, Govindarajan Gulothungan, Ramasamy Subbaiya, Muthusamy Govarthanan, Natchimuthu Karmegam, and Woong Kim. 2022. "The Urge of Algal Biomass-Based Fuels for Environmental Sustainability against a Steady Tide of Biofuel Conflict Analysis: Is Third-Generation Algal Biorefinery a Boon?" *Fuel*. https://doi.org/10.1016/j.fuel.2022.123494.

Theus, Martin, and Simon Urbanek. 2008. "The Titanic Disaster Revisited." *Interactive Graphics for Data Analysis*. https://doi.org/10.1201/b17187-15.

venky. n.d. "Data-Analysis-Project-Titanic/titanic_data.csv at Master · venky14/Data-Analysis-Project-Titanic." Accessed October 8, 2021. https://github.com/venky14/Data-Analysis-Project-Titanic.

Vickram, Sundaram, Karunakaran Rohini, Krishnan Anbarasu, Nibedita Dey, Palanivelu Jeyanthi, Sundaram Thanigaivel, Praveen Kumar Issac, and Jesu Arockiaraj. 2022. "Semenogelin, a Coagulum Macromolecule Monitoring Factor Involved in the First Step of Fertilization: A Prospective Review." *International Journal of Biological*

Macromolecules 209 (Pt A): 951–62.
Whitley, Michael Aaron. 2015. *Using Statistical Learning to Predict Survival of Passengers on the RMS Titanic*.
Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123814.
Zhang, Zhongheng. 2016. "Naïve Bayes Classification in R." *Annals of Translational Medicine* 4 (12): 241.
Narayanasamy, S., Sundaram, V., Sundaram, T., & Vo, D. V. N. (2022). Biosorptive ascendency of plant based biosorbents in removing hexavalent chromium from aqueous solutions–Insights into isotherm and kinetic studies. Environmental Research, 210, 112902.

Tables and Figures

Table 1. Accuracy Values for Logistic Regression and Naive Bayes

S.NO	Logistic Regression	Naive Bayes
1	95.00	92.50
2	93.00	90.00
3	94.00	91.00
4	93.50	88.60
5	92.00	89.30
6	91.70	87.90
7	90.40	88.90
8	95.80	86.90
9	94.20	87.00
10	89.80	87.40

Table 2. Group Statistics Results-Logistic Regression has an mean accuracy (92.94%), std.deviation (1.94), whereas for Naive Bayes has mean accuracy (88.95%), std.deviation (1.81).

Group Statistics					
	Groups	**N**	**Mean**	**Std deviation**	**Std. Error Mean**
Accuracy	Logistic Regression	10	92.94	1.94	0.61
	Naive Bayes	10	88.95	1.81	0.57

Table 3. Independent Samples T-test - Logistic Regression seems to be significantly better than Naive Bayes (p=0.004)

Accuracy	Independent Samples Test								
	Levene's Test for Equality of Variances		T-test for Equality of Means						
	F	Sig	t	df	Sig(2-tailed)	Mean Difference	Std.Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	0.136	0.004	4.737	18	0.000	3.99	0.84228	2.22044	5.75956
Equal variances not assumed			4.737	17.910	0.000	3.99	0.84228	2.21980	5.76020

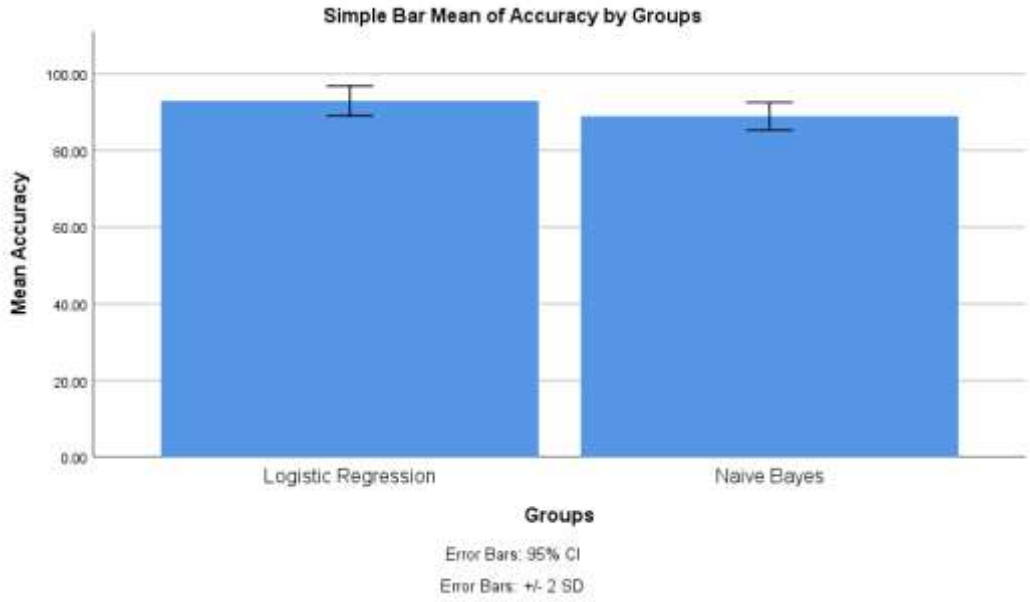

Fig. 1. Bar Graph Comparison on mean accuracy of Logistic Regression (92.94%) and Naive Bayes (88.95%). X axis: Logistic Regression, Naive Bayes, Y axis: Mean accuracy +/- 2 SD