PREDICTING TITANIC SURVIVAL BY USING MACHINE LEARNING MODELS

Asiya Begum [1] Dr. B. Sasi Kumar[2]

M. Tech Student -CSE, Department of Computer Science Engineering, Dr. V.R.K Women's College of Engineering & Technology, Hyderabad, Telangana, India.

Principal & Professor, Department of Computer Science Engineering, Dr. V.R.K Women's College of Engineering & Technology, Hyderabad, Telangana, India

ABSTRACT

Machine learning, a subfield of artificial intelligence, has emerged as a transformative technology with farreaching applications across various industries. It focuses on the development of algorithms that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed. By leveraging vast datasets and sophisticated models, machine learning has shown remarkable success in tasks such as image and speech recognition, natural language processing, recommendation systems, and predictive analytics. The sinking of the RMS Titanic remains one of the most iconic maritime disasters in history, making it a focal point for predictive modeling and data analysis. This study presents a comprehensive comparative analysis of nine distinct machine learning models for predicting passenger survival on the Titanic. The objective is to determine which model yields the highest accuracy and reliability in forecasting passenger survival, shedding light on the strengths and weaknesses of each approach.

Key words: RMS Titanic, prediction, Random forest, K nearest neighbors, Support Vector Machines, Decision Tree, Logistic regression, Linear SVC, Perceptron, Stochastic Gradient Decent, Feature engineering.

INTRODUCTION:

The advancements in machine learning and data science have opened up new avenues for understanding historical events like the Titanic tragedy. By leveraging these technologies and analyzing the available data, we can attempt to unravel the factors that influenced the survival of the passengers and ultimately build predictive models that shed light on the possible outcomes.

The objective of this study is to explore the application of machine learning algorithms to predict the survival of Titanic passengers based on various attributes such as age, sex, ticket class, family size, and cabin location. To achieve this, we will utilize a dataset compiled from passenger records that includes information on those who survived and those who perished during that fateful voyage.

Machine learning [1] is the umbrella term for any method that can be implemented on a computer and applied to a set of data to search for patterns. In essence, this refers to all algorithms used in data science, regardless of whether they are supervised or unsupervised, employed for segmentation, classification, or regression. Machine learning is an essential tool in many fields, including autonomous driving, handwriting identification, language translation, speech recognition, and image categorization. The pixels that make up an image's characteristics are used to create predictions for picture categorization, and machine learning algorithms have the ability to make these discoveries. In autonomous vehicles, information from cameras, range sensors, and GPS is combined with the pitch and volume of sound samples for voice recognition. How many Titanic survivors there will be is predicted using machine learning techniques. A number of features, such as name, title, age, sex, and class, will be used to produce the forecasts. In order to find meaningful and useful patterns in vast amounts of data, predictive analysis uses computational approaches. The likelihood of survival is estimated using machine learning approaches based on different feature combinations. The objective is to do exploratory data analytics on the currently accessible dataset and to examine the effect of each field on passengers' survival by applying analytics between each dataset field and the "Survival" field. Numerous algorithms are examined for accuracy, and the most accurate algorithm is then recommended for predictions. The most notorious catastrophe is known as the sinking of "The Titanic," which happened on April 15, 1912, more than a century ago. The Titanic suffered extensive damage as a result of the collision with the iceberg. On that terrible night, a wide variety of people of all ages and genders were present, but it was unfortunate that there weren't enough lifeboats available for rescue. There were many men among the dead, and the numerous women and kids on board took their place. The men taking the second-class flight were already dead. [2] In order to forecast which people survived when the Titanic sank, machine learning techniques are used. Features such as ticket price, age, sex, and class. By applying analytics between each dataset field and the "Survival" field, it will be possible to undertake exploratory data analytics to mine the available dataset for diverse information and determine the impact of each field on passenger survival. A machine learning method is used to make predictions about newer data sets. The correctness of the data analysis using the implemented algorithms will be examined. The most accurate algorithm is offered for predictions after being compared to other algorithms' accuracy levels.

1.1 Introduction to python A general-purpose, high-level programming language, Python has gained popularity recently. It enables programmers to write code in fewer lines, something that is not achievable in other languages. Python programming is notable for its support for several programming paradigms. Python has a huge collection of comprehensive standard libraries that are expandable. Python's key characteristics include its simplicity and ease of learning, freeware and open source status, high-level programming language, platform independence, portability, dynamically typed, both procedure- and object-oriented design, interpreted, extendable, embedded nature, and sizeable library. 1.2 Introduction to Data Science Data Science is a multidisciplinary field that employs scientific methods, practises, tools, and systems to glean knowledge from both structured and unstructured data. Big data, data mining, and data analytics are all connected to data science. It is aware of the phenomenon behind the data. It uses methods and theories that are derived from a variety of disciplines in the context of mathematics, statistics, computer science, and information science. 1.3 Introduction to Machine Learning Automatically identifying meaningful patterns in data is a process known as machine learning. In the recent years, it has transformed into a common tool for almost any task needing information extraction from large data sets. The technology that permeates our lives nowadays includes machine learning. Search engines figure out how to provide us the best results while putting profitable adverts, anti-spam software figures out how to filter our email communications, and fraud-spotting software safeguards credit card transactions. Face recognition is possible with digital cameras, while voice recognition is possible with personal assistant apps on smartphones. 1.4 Python for Data Science The most important data science libraries to be familiar with are as follows: • Numpy • Matplotlib • Scipy Numpy: Numpy will greatly improve our ability to manage multi-dimensional arrays. Although doing so directly might be challenging, Numpy is the foundation upon which many other libraries (indeed, virtually all of them) are built. Simply put, using Pandas, Matplotlib, Scipy, or Scikit-Learn is challenging without Numpy. Matplotlib: The visualisation of data is crucial. Data visualisation enables us to more effectively comprehend the data, locate information that would not be seen in the raw form, and present our discoveries to others. Matplotlib is the top-rated and most well-known Python data visualisation library. Although it is not user-friendly, it often offers a variety of capabilities, such as bar charts, scatterplots, pie charts, and histograms, which are helpful for projecting multidimensional data. Scipy: Numerous concepts that are very significant but also complicated and time-consuming are covered in mathematics. But Python has a whole scipy library that takes care of this problem for us. We will learn how to use this library in this programme, along with a few functions and illustrations of how they work

EXISTING SYSTEM:

The existing system gives us a result with decision tree having the highest score with 85.6% correct prediction. By using KNN the accuracy is 83% and with SVM the accuracy is 79%. The dataset for Titanic survival prediction is often imbalanced, with a higher number of non-survivors compared to survivors. This imbalance can bias the model's performance and lead to inaccurate predictions, especially for the minority class (survivors). The dataset used in existing systems might contain missing values, outliers, or other data quality issues that can adversely impact the performance of the prediction model. Preprocessing and handling such issues is critical to improving prediction accuracy. Certain machine learning models, especially complex ones, lack interpretability. This can be problematic when trying to understand the factors that contribute to a passenger's predicted survival. For certain applications (e.g., medical or legal contexts), interpretability is crucial.

PROPOSED SYSTEM:

The proposed system culminates in the identification of the most effective machine learning model for predicting Titanic survival, emphasizing the significance of model selection and tuning. Additionally, this system underscores the value of feature engineering to enhance predictive accuracy and provides an essential framework for historical data analysis.

Our proposed system aims to contribute to the field of data analysis and machine learning by emphasizing the importance of model selection and the applicability of these techniques to historical events. By offering a practical approach to predicting passenger survival on the Titanic, it serves as a valuable reference for researchers and data scientists interested in historical data analysis and predictive modeling, while also highlighting the potential for broader applications in other domains.

SYSTEM ARCHITECHTURE:

It is highly likely that the data we collected contains errors, missing numbers, and corrupted values because it is still in its raw form. Before making any conclusions from the data, feature engineering and data wrangling, often known as data preparation, are required. In order to make large, complicated data sets easy to access and analyze, data wrangling involves organizing and cleaning them up. To increase the predictive power of learning algorithms, a method called feature engineering seeks to generate more pertinent features from the raw features of the data.

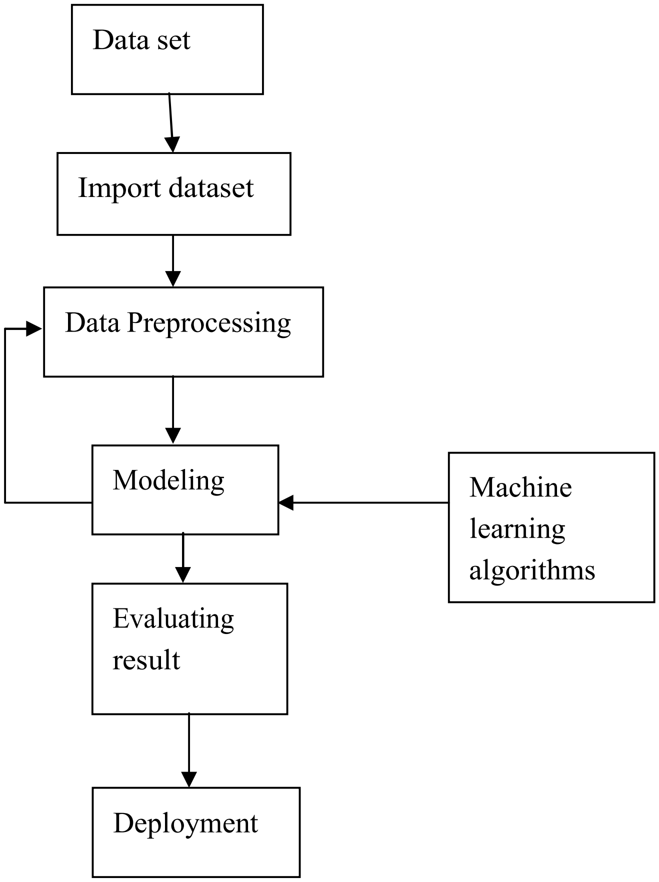


Fig: Operational flow chart

IMPLEMENTATION:

Data Collection and Preprocessing: Data is gathered from historical records, including passenger demographics, ticket class, family relationships, and other relevant attributes.

Data preprocessing techniques, such as missing value imputation, outlier handling, and feature scaling, are applied to ensure data quality.

Feature Engineering:

Feature selection and extraction techniques are employed to enhance the predictive power of the models.

Categorical variables are encoded, and new features may be created to capture important information.

Model Selection:

Nine distinct machine learning models, including Decision Trees, Random Forest, Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Linear SVC, Perceptron, Naive Bayes, and SGD, are integrated into the system.

Training and Cross-Validation:

The dataset is split into training and validation sets to train and evaluate each model.

Cross-validation techniques are applied to assess the model's robustness and generalization ability.

Model Evaluation:

Various performance metrics, such as accuracy, precision, recall, and F1-score, are used to evaluate the models' effectiveness. The models are ranked based on their predictive performance.

Hyperparameter Tuning:

Hyperparameter optimization is performed for each model to fine-tune their configurations and improve predictive accuracy.

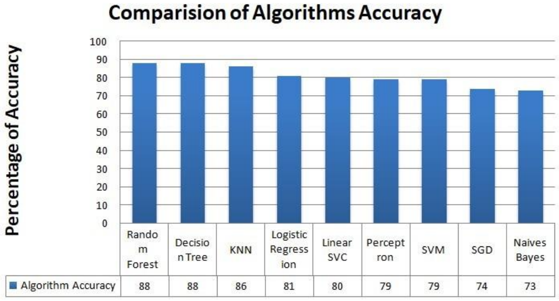
Model Integration:

The best-performing model is selected based on the comparative analysis and integrated into the system for deployment.

Real-time Prediction: The deployed model is used to predict the survival likelihood of passengers in realtime.

GRAPH ANALYSIS:

The findings of this research provide a deeper understanding of the predictive power of machine learning models when applied to the Titanic dataset from a graph perspective. This approach can be instrumental in identifying influential passengers, understanding the dynamics of survival within different passenger groups, and improving the overall accuracy of survival predictions. Moreover, it highlights the potential for utilizing graph analysis in other historical and network-related prediction tasks. Traditional performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the performance of each model. However, we also introduce new metrics that consider the network characteristics of the graph, such as centrality measures and community detection.



CONCLUSION:

This paper presents a comparative study on machine learning techniques to analyze Titanic dataset to learn what features effect the classification results and which techniques are robust.

The findings of this study reveal insights into the efficacy of different machine learning models in predicting Titanic survival. The best-performing model, is Random forest and Decision tree with 88% accuracy. Additionally, this study underscores the importance of feature engineering, hyperparameter tuning, and model selection in enhancing predictive accuracy for this specific historical event.

REFERENCES:

[1]https://www.kaggle.com/competitions/titanic/data.

[2] Karman Singh, Renuka Nagpal and Rajini Sehgal. “Exploratory Data Analysis and Machine learning on Titanic Disaster Dataset”.

[3] Jyothi Shetty, Pallavi S and Ramyashree. “Predicting the survival rate of Titanic Disaster using Machine learning Approaches”.

[4] Analyzing Titanic disaster using machine learningalgorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.

[5] Jain, Nikita, and Vishal Srivastava. "Data mining techniques: a survey paper." IJRET: International Journal of Research in Engineering and Technology 2.11 (2013): 2319-1163.

[6] Zhao, Zheng, and Huan Liu. "Spectral feature selection for supervised and unsupervised learning." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

[7] Farag, Nadine, and Ghada Hassan. Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster. ICSIE '18 Proceedings of the 7th International Conference on Software and Information Engineering, May 2018, dl.acm.org/citation.cfm?id=3220282.

[8] Disaster, CS229 Titanic-Machine Learning From. "Eric Lam Stanford

University."

[9] Singh, Aakriti, Shipra Saraswat, and Neetu Faujdar. "Analyzing Titanic disaster using machine learning algorithms." 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017.

[10] Han, Jun; Morag, Claudio (1995). "The influence of the sigmoid function parameters on the speed of backpropagation learning". In Mira, Jose; Sandoval, Francisco (eds.). From Natural to Artificial Neural Computation. Lecture Notes in Computer

Science. 930. pp. 195-201. doi:10.1007/3-540- 59497-3_175. ISBN 978-3-540-59497-0.

[11]J C. Bezdek *Introduction of statistical model* 1973. 6.Vapnik V.N. The Nature of Statistical Learning Theory[M]. New YorkSpringer-Verlag.1995 7.V. Vapnik, “Statistical learning theory,” Wiley, New York, 1998. 8.Atakurt, Y., 1999, Logistic Regression Analysis and an Implementation in Its Use in Medicine, Ankara University Faculty of Medicine Journal, C.52, Issue 4, P.195, Ankara 9.M Jamel Selim S Z *The construction of decision tree* vol. 61 pp. 177-188 1994.

10.https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

[12] Unwin A, Hofmann H (1999). \GUI and Command-line{ Conict or Synergy?" In K Berk,M Pourahmadi (eds.), Computing Science and Statistics. 12.Galit Shmueli and Otto R. Koppius MIS Quarterly, Predictive Analytics in Information System Research, , Vol. 35, No. 3(September 2011), pp. 553-572. 13.Michalski R S,et al.Machine Learning:Challenges of the eighties.Machine Learning,1986,99-10