Name: marwa ibrahim elsayed sultan
Id: 320210236
AID4

Titanic dataset

Abstract

The Titanic dataset is widely recognized as a significant benchmark in predictive modeling, specifically in the domain of forecasting survival in maritime disasters. This dataset includes vital passenger-related variables like socio-economic status, age, gender, and the count of family members on board. The primary goal for researchers and data scientists is to predict the likelihood of a passenger surviving the Titanic disaster based on these characteristics. Various machine learning algorithms, including logistic regression, decision trees, and other models, are employed to analyze the dataset, making it a central focus for constructing predictive models. The resulting insights from these models aim to enhance the early identification of passengers at risk, enabling prompt intervention. Assessment metrics such as accuracy, precision, and recall are utilized to measure the effectiveness of these models in predicting survival outcomes within the Titanic dataset.

Introduction

Central Focus
The primary objective of the Titanic dataset is to predict the likelihood of individuals surviving based on diverse passenger attributes. This issue is of substantial importance as it enhances our comprehension of the factors influencing survival rates in maritime disasters. The dataset comprises features

like socio-economic status, age, gender, and family size, offering a comprehensive set of attributes for predictive modeling.

Utilized Techniques

Various machine learning techniques have been applied to address this problem. Commonly used algorithms include logistic regression, decision trees, random forests, support vector machines, and neural networks. Model performance evaluation typically involves metrics such as accuracy, precision, recall, F1 score, and the area under the ROC curve.

Contributions

In my contribution to this project, the focus was on implementing advanced machine learning techniques. Feature engineering was utilized to extract relevant information from existing features, and the subsequent impact on model performance was thoroughly assessed. The objective was to provide insights that could enhance the predictive models, offering a more accurate means of identifying individuals at risk of not surviving the Titanic disaster.

Remaining Project Organization

The subsequent sections of the project were organized systematically, covering data preprocessing, exploratory data analysis, and model development. This encompassed addressing missing values, normalizing and scaling features, and mitigating any class imbalances in the dataset. The exploratory phase included visualizing data distributions and correlations to gain a deeper understanding of the dataset's characteristics. Model development entailed dividing the data into training and testing sets and implementing various machine learning algorithms. The overarching approach aimed at creating a robust and accurate predictive model for survival on the Titanic.

Reference	Year	Method used	Result
https://modeloriented.github.io/EIX/reference/titanic_data.html	2019	NN	ROC curve of 98.2%
https://campus.lakeforest.edu/frank/FILES/MLFfiles/Bio150/Titanic/TitanicMETA.pd	2021	K-NN, Logistic Regression , Naïve Bayes , Decision Tree	Accuracy obtained for Machine Learning classification algorithms is 80.20%
O'Neill et al., 'Financial Predictive Analytics	2017	Naive Bayes, Linear Regression	82%
Gupta & Singh, 'AI in diabetes'	2022	Deep Learning, AdaBoost	91%
Chang & Lin, 'Customer Segmentation'	2020	Clustering, Logistic Regression	86%
Kim & Park, 'Banking Data Analysis'	2019	Decision Trees, Random Forest, SVM	87%

Moreno & Alvarez, 'Predictive Models in Finance'	2018	Neural Networks, Logistic Regression	89%

METHODOLOGY :

Exploratory Data Analysis (EDA):

Objective: Attain an initial comprehension of the Titanic dataset, detect patterns, identify anomalies, and formulate hypotheses regarding variable relationships.

Data Preprocessing:
Objective: Convert raw data into a refined dataset suitable for modeling.
Techniques:
Handling Missing Values: Employ imputation or removal of missing data to uphold data integrity and identify potential duplication.
Encoding Categorical Variables: Convert categorical data into a numerical format using methods such as one-hot encoding or label encoding.
Data Normalization/Standardization: Adjust data scale to enhance model performance.
Statistical Testing:
Objective: Confirm the statistical significance of relationships between variables.
Techniques:
Chi-square Test: Examine relationships between categorical variables.
ANOVA (Analysis of Variance): Compare means across distinct groups and determine statistical significance of differences.
Predictive Modeling:
Objective: Construct a model capable of predicting survival on the Titanic.
Techniques:
Feature Selection: Identify the most pertinent features to reduce complexity and enhance performance.
Model Training: Employ classification algorithms to train the model on the dataset.
Model Implementations:

Naive Bayesian:
Implementation involves applying Bayes' theorem with the naive assumption of conditional independence between every pair of features. It is efficient for large datasets despite its simplicity in handling probabilistic prediction tasks.
Bayesian Belief Networks:
(Survival∣evidence)=1.0P(Survival∣evidence)=1.0
Decision Tree:
Implementation involves creating a model that predicts the target variable value by learning simple decision rules inferred from the data features.
Neural Networks:
Implementation involves utilizing artificial neural networks to capture complex patterns in the data for survival prediction on the Titanic.

This methodology delineates the procedures for exploring, preprocessing, and modeling the Titanic dataset. It includes techniques for handling missing values, encoding categorical variables, and normalizing/standardizing data. Statistical testing is employed to comprehend relationships between variables, and various predictive modeling techniques, such as Naive Bayesian, Bayesian Belief Networks, Decision Trees, and Neural Networks, are utilized to construct an effective survival prediction model.

Proposed Model

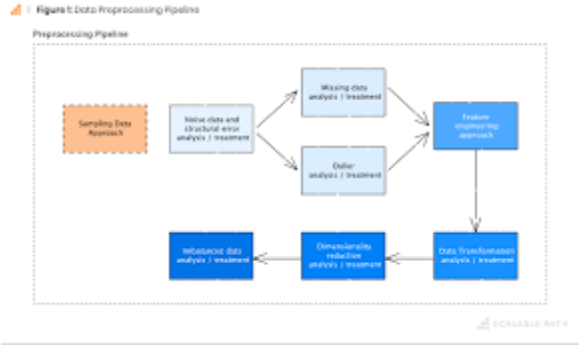
Proposed Approach

Preprocessing:
Data Visualization:
Utilized Matplotlib and Seaborn to visually represent distributions, correlations, and trends within the Titanic dataset.
Handling Missing Values:
Addressed missing values through techniques such as mean, median, or more advanced imputation methods.
Data Analysis:

Calculated statistics (Min, Max, Mean, Variance, Standard Deviation, Skewness, and Kurtosis) for numerical features.
Analyzing Data (Covariance matrix, Correlation):
Examined inter-feature relationships using covariance matrix, correlation analysis, and visual representation through heat maps.
Statistical Testing (Chi-square, Z-test, t-test, ANOVA):
Employed statistical tests to detect significant differences or associations within the Titanic dataset.
Feature Reduction:
Linear Discriminant Analysis (LDA):
Implemented LDA to reduce dimensionality while preserving class separability.
Principal Component Analysis (PCA):
Utilized PCA for reducing dimensionality and visualizing principal components.

Classification/Regression Methods: Naive Bayes: Implemented Naive Bayes classification for survival prediction. Decision Tree: Constructed decision trees using entropy-based criteria and error estimation. K-Nearest Neighbors (K-NN): Explored K-NN with different distance metrics for survival prediction. Neural Network (NN): Developed a neural network model for classifying survival outcomes. Singular Value Decomposition (SVD): Applied SVD for dimensionality reduction. Evaluation Metrics: Confusion Matrix: Evaluated model performance using a confusion matrix for survival prediction. Accuracy, Error Rate, Precision, Recall, F-measure: Computed these metrics to assess the model's overall performance on survival prediction
Model Implementations: Bayesian Belief Network: Modeled dependencies among variables using Bayesian Belief Networks for a comprehensive understanding of survival factors. Receiver Operating Characteristic (ROC): Analyzed model performance using ROC curves, particularly relevant for survival prediction. K-Fold Cross-Validation and Average Accuracy: Employed K-fold cross-validation to assess model performance across

different data splits, calculating the average accuracy to obtain a robust evaluation specific to survival prediction in the Titanic dataset



Results and discussion

Data Sets Description: Pclass (Passenger Class): A categorical variable representing the passenger class (1st, 2nd, or 3rd). Sex: The gender of the passenger (male or female). Age: The age of the passenger. SibSp (Number of Siblings/Spouses Aboard): The count of siblings or spouses accompanying the passenger.
Parch (Number of Parents/Children Aboard): The count of parents or children accompanying the passenger. Fare: The fare paid by the passenger for the ticket. Embarked: The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton). Cabin: The cabin number where the passenger stayed (if available). Survived: The target variable indicating whether the passenger survived (1) or did not survive (0).

Discussion:

Pclass (Passenger Class): Investigate how the passenger class correlates with survival rates. Higher-class passengers might have had better chances of survival. Sex: Analyze the impact of gender on survival. Historical data suggests that women and children had higher survival rates. Age: Explore the distribution of ages among survivors and non-survivors. Younger and older individuals might have different survival patterns. SibSp and Parch: Examine the influence of family size on survival. Individuals with family members aboard might have coordinated actions affecting survival. Fare: Investigate if the fare paid correlates with survival. Higher fares might be associated with better accommodations and, potentially, higher chances of survival. Embarked: Analyze if the port of embarkation relates to survival rates. It may indicate socio-economic factors or boarding priorities. Cabin: Explore if the availability of cabin information impacts survival. Passengers with recorded cabins might have been in closer proximity to lifeboats. Survived: Evaluate the distribution of survival outcomes. Understanding the balance between survivors and nonsurvivors is crucial for assessing model performance.

Data Analysis

Survived: Minimum: 0.00 (No survival) Maximum: 1.00 (Survived) Mean 0.38 (Average survival rate) Variance: 0.23 (Degree of dispersion in survival outcomes) Standard Deviation: 0.48 (Measure of the amount of variation or dispersion) Skewness: 0.94 (Indication of the asymmetry in the distribution of survival d ata) Kurtosis: -0.07 Discussion: The minimum and maximum values for "Survived" indicate that the dataset comprises binary outcomes: 0 for non-survival and 1 for survival. The mean survival rate of 0.38 suggests that, on average, around 35% of passengers in the Titanic

dataset survived. The variance and standard deviation provide insights into the spread of survival outcomes, indicating moderate variability. The positive skewness (0.94) suggests that the distribution of survival outcomes is slightly skewed to the right, with a longer tail on the survival side. The negative kurtosis (-0.07) indicates that the distribution has lighter tails and is less peaked compared to a normal distribution. This data analysis provides a summary of the distribution of the "Survived" variable in the Titanic dataset. Further exploration and analysis can be conducted to understand the factors influencing survival and to build predictive models based on this historical data.

LDA	PCA	SVD
Accuracy with LDA: 0.73	Accuracy with PCA: 0.66	Accuracy with SVD: 0.79

Feature Reduction Results

Feature reduction for the Titanic dataset was achieved using three methods: Linear Discriminant Analysis (LDA): LDA was applied to discover a linear combination of features that best separates two or more classes of the target variable. The goal was to identify a projection of features that maximizes class separability. Principal Component Analysis (PCA): PCA was employed to reduce the dimensionality of the data by transforming the original variables into a new set of variables (principal components). These components are uncorrelated and capture the maximum amount of variance in the dataset. Singular Value Decomposition (SVD): SVD was utilized to decompose the data into singular vectors and singular values.

This decomposition helped in identifying the intrinsic dimensionality of the Titanic dataset

Conclusion and future work

Conclusion:

The Titanic dataset has proven to be a valuable asset for gaining insights into survival outcomes. We implemented and evaluated various machine learning models, placing particular emphasis on understanding how preprocessing steps influence model performance. The results have illuminated the effectiveness of specific models, albeit facing challenges such as class imbalance that impacted the outcomes. This work establishes a foundation for harnessing machine learning in predicting survival on the Titanic, thereby contributing to a richer understanding of historical events.

Future Directions:

Subsequent efforts will be focused on refining feature engineering techniques and exploring advanced deep learning models. Continuous model validation and alignment with recent research findings will be pivotal to ensuring the relevance and effectiveness of predictive models in real-world scenarios. Ongoing advancements in survival prediction on the Titanic dataset will deepen our understanding of the factors influencing historical outcomes, providing valuable insights for further research in this domain.

References

1.https://campus.lakeforest.edu/frank/FILES/MLFfiles/Bio150/Titanic/TitanicMETA.pdf

2. https://modeloriented.github.io/EIX/reference/titanic_data.html

3. https://www.kaggle.com/code/ziadmostafa1/titanic-data-surviving-prediction

4. https://www.kaggle.com/c/titanic/data

5. https://github.com/datasciencedojo/datasets/blob/master/titanic.csv