MODERN EDUCATION SOCIETY'S WADIA COLLEGE OF ENGINEERING,

PUNE, 411014.

DEPARTMENT OF COMPUTER ENGINEERING

.A PRELIMENERY REPORT ON

"Titanic People Survival Probability"

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Computer Engineering

By

Sahil Naik	S20111006
Sohan Chidrawar	S20111013
Supreet Birajdar	S20111017
Abhishek Bhagat	S20111009

Under the guidance of

Dr. A. P. Kale

SAVITRIBAI PHULE PUNE UNIVERSITY

2022-2023

CERTIFICATE

This is to certify that the project report entitles

"Titanic People Survival Probability"

Submitted by

Sahil Naik	S20111006
Sohan Chidrawar	S20111013
Supreet Birajdar	S20111017
Abhishek Bhagat	S20111009

Are bonafide students of this institute and the work has been carried out by them under the supervision of Dr. A. P. Kale and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University, for the award of the degree of **Bachelor of Engineering** (Computer Engineering)

Dr. A. P. Kale
Guide
Department of Computer Engineering

Dr. N. F. Shaikh
Head,
Department of Computer Engineering

Dr. Mrs. M. P. Dale
Principal,
Modern Education Society's College of Engineering Pune – 14

Place : Pune
Date : 29-10-2023

ACKNOWLEDGEMENT

The present world of competition there is a race of existence in which those are having will to come forward succeed. Project is like a bridge between theoretical and practical working. With thiswilling we joined this particular project. First of all, we would like to thank the supreme power the Almighty God who is obviously the one has always guided us to work on the right path of life.

We sincerely thank **Dr. N. F. Shaikh**, Head of the Department of Computer Science of Modern Education Society's college of engineering, for all the facilities provided to us in the pursuit of this project.

We are indebted to our project guide Dr. A. P. Kale Department of ComputerScience of Modern Education Society's college of engineering college of engineering. We feel it's a pleasure to be indebted to our guide for his valuable support, advice and encouragement and we think him for his superb and constant guidance towards this project.

We are deeply grateful to all the staff members of CS department, for supporting us in all aspects.

We acknowledge our deep sense of gratitude to our loving parents for being a constant source of inspiration and motivation.

CONTENTS

Abstract

The sinking of the RMS Titanic in 1912 remains one of the most tragic maritime disasters in history. This project endeavors to employ logistic regression, a fundamental machine learning technique, to predict whether passengers aboard the Titanic survived or perished. The objective is to develop a predictive model that accurately determines the survival outcome of individuals based on several key features, such as age, sex, and class. The project encompasses a comprehensive data analysis process, including data preprocessing, exploration, modeling, and evaluation. Key steps involve handling missing data, feature selection and engineering, as well as model training and testing.

The logistic regression model's performance is assessed using various metrics, allowing for insights into the significance of different features. This project serves not only as an introduction to machine learning but also as a tribute to the passengers and crew of the Titanic, who experienced this historic tragedy. Ultimately, the logistic regression model provides valuable insights into the factors influencing survival and may pave the way for further improvements and more advanced modeling techniques in future studies.

Introduction

This project uses logistic regression to predict whether passengers on the Titanic survived or not. Key steps include data preprocessing, feature selection, and model training. The goal is to create a simple, accurate model that sheds light on the factors influencing survival on the ill-fated ship.

This project leverages logistic regression to predict Titanic passengers' survival. Key steps involve data preprocessing, where we handle missing data and select pertinent features. Model training and testing are carried out to develop an accurate survival prediction model. The project offers insights into which factors, such as age, sex, and class, played a significant role in determining who survived and who did not. Additionally, this project contributes to the broader understanding of machine learning in the context of historical events and is a tribute to those aboard the Titanic.

The primary objectives of this project are:

Data Collection: Obtain the Titanic dataset, which typically includes information about passengers such as age, sex, class, and other features, as well as whether they survived or not. You can find this dataset on various data science websites or libraries like Kaggle.

Data Preprocessing: Handle missing data: Check for missing values in the dataset and decide how to handle them (impute or drop rows/columns).

Feature selection: Choose the relevant features that are likely to impact the survival outcome. For example, you may find that features like "Name" or "PassengerId" are not useful.

Feature engineering: Create new features if needed. For example, you could extract titles from passenger names or create a "family size" feature.

Data Exploration: Explore and visualize the data to understand its characteristics and relationships between features.

Calculate statistics and correlations to gain insights into the dataset.

Data Splitting: Split your dataset into training and testing sets. The usual split is 80-20 or 70-30 for training and testing, respectively.

Model Building: Train a logistic regression model on the training data. You can use libraries like scikit-learn in Python to do this.

Tune hyperparameters if needed, like regularization strength (e.g., L1 or L2 regularization) or solver options.

Model Evaluation: Evaluate your model's performance on the testing data using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC AUC.

Consider using cross-validation for a more robust evaluation.

Model Interpretation: Interpret the coefficients of your logistic regression model to understand which features are significant in predicting survival.

Model Deployment (optional): If you intend to use this model in a real-world application, deploy it using the appropriate tools and frameworks.

Iterate and Improve: Based on your evaluation results, consider feature engineering, trying different models, or fine-tuning your logistic regression model to improve its performance.

Documentation and Reporting: Document your project, including the steps you took, the decisions you made, and the model's performance. Create visualizations and reports to communicate your findings.

Problem Statement

The sinking of the RMS Titanic resulted in a significant loss of life, and understanding the factors that contributed to survival can provide historical insights. The challenge is to build a predictive model using logistic regression to determine the likelihood of passengers' survival based on available data.

Motivation

Historical Insight: The Titanic disaster is a poignant part of history, and by analyzing its passenger data, we aim to gain insights into the human factors and circumstances that influenced survival. This project serves as a tribute to the memory of those who experienced the tragedy.

Educational Purpose: Logistic regression is a fundamental machine learning technique. This project offers a practical, hands-on experience for individuals interested in data science, allowing them to apply a simple yet effective method to a real-world problem.

Objectives

Model Development: Build a logistic regression model that accurately predicts the likelihood of a passenger's survival based on features such as age, sex, class, and more.

Feature Analysis: Investigate the significance of different features in influencing survival outcomes. Identify which factors played a key role in passengers' chances of survival.

Theory

1. Project Initiation:

Project initiation involves defining the scope, goals, and objectives of the Titanic survival prediction project. It typically includes the following steps:

- **Problem Definition:** Clearly state the problem, such as predicting passenger survival on the Titanic using logistic regression.
- **Goal Setting:** Determine the specific objectives, like building a predictive model and gaining insights into the factors influencing survival.
- **Resource Allocation:** Allocate resources, including data, tools, and human resources.
- **Timeline and Milestones:** Develop a project plan with timelines for each stage.
- **Team Formation:** Assemble a team if working collaboratively.

2. Data Collection and Preprocessing:

This stage focuses on obtaining and preparing the data for analysis:

- **Data Sourcing:** Collect the Titanic dataset, which typically includes information on passengers, their attributes, and survival outcomes.
- **Data Exploration:** Analyze the dataset's structure, identify missing values, and understand the distribution of variables.
- **Data Cleaning:** Handle missing data through imputation or removal, and ensure data consistency.
- **Feature Selection:** Choose the most relevant features for the model. Some features may not contribute significantly to the prediction.
- **Feature Engineering:** Create new features if necessary. For example, extract titles from passenger names or compute family size.
- **Data Splitting:** Divide the data into training and testing sets for model development and evaluation.

3. Model Development:

- **Model Selection:** Choose logistic regression as the modeling technique for binary classification.
- **Training:** Fit the logistic regression model to the training data, using features as input and survival outcomes as the target variable.
- **Hyperparameter Tuning:** Optimize hyperparameters, like regularization strength and solver options.
- **Model Validation:** Ensure the model's generalization by validating it on the training data.

4. Model Evaluation:
Evaluate the model's performance to determine how well it predicts passenger survival:

- **Testing:** Apply the model to the testing dataset to make predictions.
- **Performance Metrics:** Use various metrics like accuracy, precision, recall, F1-score, and ROC AUC to assess the model's quality.
- **Confusion Matrix:** Analyze the true positives, true negatives, false positives, and false negatives for a more detailed evaluation.
- **Cross-Validation:** Employ cross-validation techniques for robust assessment.

5. Optimization and Refinement:
This stage involves improving the model and the overall project:

- **Feature Importance Analysis:** Investigate the significance of features in influencing predictions. Identify which factors played a key role in passengers' chances of survival.
- **Iterative Improvement:** Refine the model, iterate on data preprocessing, or consider different modeling techniques to enhance prediction accuracy.
- **Documentation:** Prepare a comprehensive report, including project details, findings, and the model's performance, to communicate results effectively.
- **Presentation:** Share the findings with stakeholders, peers, or the broader data science community to gather feedback and insights for further refinement.

6. Documentation and Reporting:

- **Comprehensive Documentation:** Thoroughly document the project's entire process, including data sources, preprocessing steps, model development, evaluation results, and any challenges faced.
- **Visualizations:** Include visual representations of data analysis, feature importance, and model performance to enhance the clarity of the documentation.
- **Findings:** Summarize key insights and discoveries from the project, such as which passenger attributes had the most influence on survival predictions.

7. Deployment and Integration:

- **Model Deployment:** If the project's goal includes practical application, deploy the logistic regression model using appropriate tools, frameworks, or platforms to make predictions in real-world scenarios.
- **Integration:** Integrate the deployed model into the desired systems or applications, ensuring it works seamlessly with other software and processes.

8. Testing and Validation:

- **System Testing:** Verify that the deployed model functions correctly, making accurate predictions in the real-world environment.
- **Validation:** Continuously validate the model's performance by comparing its predictions with actual outcomes. This helps ensure its reliability over time.

9. Maintenance and Monitoring:

- **Model Deployment:** If the project's goal includes practical application, deploy the logistic regression model using appropriate tools, frameworks, or platforms to make predictions in real-world scenarios.
- **Integration:** Integrate the deployed model into the desired systems or applications, ensuring it works seamlessly with other software and processes.

10. Final Presentation and Knowledge Sharing:

- **Project Presentation:** Share the project's findings, methodologies, and results with relevant stakeholders, teams, or the broader data science community.
- **Knowledge Sharing:** Encourage knowledge sharing by presenting the project's successes and challenges, discussing lessons learned, and providing insights that can benefit others in similar projects.
- **Feedback Gathering:** Gather feedback from peers and stakeholders to improve the project and share valuable knowledge.

These additional stages complete the project's lifecycle, encompassing documentation, deployment, testing, maintenance, and knowledge sharing, ensuring that the project's insights and models are put into practical use and continue to evolve and improve.

Libraries Used:

When working on a project to predict Titanic survival using a logistic regression model in Python, several libraries can be incredibly helpful for various tasks throughout the project. Here are some essential libraries:

1. **NumPy:** NumPy is a fundamental library for numerical computations in Python. It provides support for arrays and matrices, making data manipulation and mathematical operations more efficient.
2. **Pandas:** Pandas is a popular library for data manipulation and analysis. It offers data structures like DataFrames, which are well-suited for handling structured data, such as the Titanic dataset. You can use Pandas to load, clean, and explore the data.
3. **Matplotlib and Seaborn:** These libraries are useful for data visualization. Matplotlib provides a wide range of customization options for creating various types of plots, while Seaborn is built on top of Matplotlib and simplifies the creation of informative and attractive statistical graphics.
4. **Scikit-Learn (sklearn):** Scikit-Learn is a comprehensive machine learning library that offers tools for data preprocessing, model selection, training, evaluation, and hyperparameter tuning. For your logistic regression model, you'll primarily use Scikit-Learn for model development and evaluation.
5. **Statsmodels:** Statsmodels is a library that specializes in statistical modeling. It's helpful for understanding the statistical significance of features in your logistic regression model, which can be essential for interpreting the results.
6. **Jupyter Notebook:** Jupyter Notebook is an interactive environment that allows you to write and execute code in a notebook-style format. It's particularly useful for documenting your project, sharing your work with others, and visualizing your analyses alongside explanations.
7. **Scipy:** Scipy builds on NumPy and provides additional scientific and technical computing functions. While Scikit-Learn includes logistic regression, Scipy can be used for more advanced statistical analysis and hypothesis testing.
8. **XGBoost or LightGBM (optional):** For advanced model optimization, you might consider using gradient boosting libraries like XGBoost or LightGBM, which can often enhance the predictive performance of your model.
9. **Joblib or Pickle:** These libraries help save and load trained machine learning models. You can use them to store your logistic regression model after training and reload it for future predictions without having to retrain.

These libraries offer a powerful combination of tools for data manipulation, analysis, machine learning, and visualization. Depending on the project's complexity, you may find additional specialized libraries useful for specific tasks, such as feature engineering or advanced statistical analysis.

Pre-processing:

1. Data Cleaning:

Handling Missing Data: Identify and deal with missing values in the dataset. You can choose to impute missing data with meaningful values (e.g., mean, median, or mode) or remove rows or columns with excessive missing values.

2. Feature Selection:

Identify Relevant Features: Determine which features are most likely to impact the prediction of survival. In the Titanic dataset, features like "Name" or "PassengerId" may not contribute significantly to the model and can be excluded.

3. Feature Engineering:

Create New Features: Generate additional features that might provide more information for the model. For example, you can extract titles (e.g., Mr., Mrs., Miss) from passenger names or compute the family size by combining the number of siblings/spouses and parents/children.

4. Data Transformation:

Categorical Variables: Convert categorical variables (e.g., "Sex" and "Embarked") into numerical format using techniques like one-hot encoding or label encoding.

Scaling Numerical Variables: Scale numerical features to have a consistent range. Common methods include Min-Max scaling or standardization (z-score scaling).

Handling Outliers: Detect and address outliers in numerical features, which can negatively impact model performance.

5. Data Splitting:

Train-Test Split: Divide the dataset into a training set and a testing set. Typically, this split is 80% for training and 20% for testing. The training set is used to train the logistic regression model, while the testing set is used to evaluate its performance.

6. Data Validation:

Cross-Validation: Consider using cross-validation techniques (e.g., k-fold cross-validation) to assess the model's generalization performance. This helps ensure that the model is robust and not overfitting the training data.

7. Data Imbalance (if applicable):

Dealing with Imbalanced Data: If the dataset has imbalanced classes (e.g., significantly more survivors than non-survivors), you may need to apply techniques like oversampling or under sampling to balance the classes.

8. Documentation:

Record Changes: Keep a log of all the preprocessing steps, as well as any transformations and imputations performed on the data. This documentation is essential for transparency and reproducibility.

Effective data preprocessing is essential to ensure that the input data for your logistic regression model is clean, well-structured, and contains relevant information. It can significantly impact the model's accuracy and ability to make meaningful predictions.

Model Selection:

Selecting an appropriate model is a critical step in the machine learning project, such as predicting Titanic survival using a logistic regression model. Model selection involves choosing the most suitable algorithm for your specific problem based on factors like the nature of the data, the task (classification or regression), and the available computational resources. Here's how to go about model selection:

1. Problem Understanding:

Before choosing a model, it's essential to have a clear understanding of the problem you're trying to solve. In the case of Titanic survival prediction, it's a binary classification problem where you're predicting whether a passenger survived or not.

2. Logistic Regression:

Since you've mentioned using a logistic regression model, it's worth highlighting why it's a good choice for this problem:

Binary Classification: Logistic regression is a popular choice for binary classification tasks, where the outcome is one of two classes (survived or not survived).

Interpretability: Logistic regression provides coefficients for each feature, making it easy to interpret and understand the impact of each feature on the prediction.

Efficiency: Logistic regression is computationally efficient and often works well with moderate-sized datasets.

3. Model Assumptions:

Consider the assumptions of the logistic regression model and check if they align with your data. Logistic regression assumes that the relationship between the features and the log-odds of the target variable is linear. This may not hold true in all cases.

4. Model Complexity:

Evaluate the complexity of the model. Logistic regression is a simple model, which can be advantageous when you have limited data or you want a transparent and interpretable model. If your data is highly complex, you might consider more complex models like decision trees, random forests, or gradient boosting.

5. Performance and Cross-Validation:

Assess the model's performance using cross-validation techniques. Use evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC to measure how well the model is performing. Cross-validation helps you gauge the model's generalization performance and identify any overfitting issues.

6. Compare Models:

If you're open to exploring other models, you can perform model comparisons by trying alternative algorithms like decision trees, random forests, support vector machines, or gradient boosting. Compare their performance and choose the one that provides the best results.

7. Ensemble Methods (Optional):

Consider using ensemble methods like random forests or gradient boosting, which combine multiple base models to improve predictive performance. These methods can often boost accuracy and handle complex relationships in the data.

8. Model Interpretability:

Assess the interpretability of the chosen model. Logistic regression is highly interpretable due to its linear nature. Other models like decision trees may also offer interpretability through feature importance analysis.

9. Consider Practical Constraints:

Consider any practical constraints, such as computational resources or deployment requirements. Some models may be more computationally intensive than others and may not be suitable for all situations.

10. Iteration:

Model selection is not a one-time process. You may need to iterate on your choice based on the performance, feature engineering, and other considerations. It's essential to remain flexible and open to adjustments throughout your project.

In the case of logistic regression for predicting Titanic survival, this model is a reasonable starting point due to its simplicity and interpretability. However, it's always a good practice to evaluate and compare with other models to ensure you're using the one that best suits your specific dataset and problem.

Metrics:

Evaluating the performance of your machine learning model is a critical step in the Titanic survival prediction project. Various evaluation metrics help you assess how well your logistic regression model is doing in terms of making predictions. Here are some common evaluation metrics that you can use in this project:

1. Accuracy:

Formula: (True Positives + True Negatives) / (Total Predictions)

Use: Accuracy measures the proportion of correctly predicted outcomes (both survived and not survived). It's a good overall performance metric but can be misleading if the dataset is imbalanced.

2. Precision:

Formula: True Positives / (True Positives + False Positives)

Use: Precision calculates the accuracy of positive predictions. It's useful when the cost of false positives is high (e.g., misclassifying survivors as non-survivors).

3. Recall (Sensitivity or True Positive Rate):

Formula: True Positives / (True Positives + False Negatives)

Use: Recall measures the model's ability to identify all relevant instances. It's essential when the cost of false negatives is high (e.g., missing survivors).

4. F1-Score:

Formula: 2 * (Precision * Recall) / (Precision + Recall)

Use: The F1-score combines precision and recall into a single metric. It's useful when you want to balance precision and recall, especially when class imbalances exist.

5. ROC AUC (Receiver Operating Characteristic - Area Under the Curve):

Use: ROC AUC measures the model's ability to distinguish between the positive and negative classes across different probability thresholds. A higher ROC AUC indicates better discriminative power.

6. Confusion Matrix:

Use: The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives. It helps in understanding where the model makes errors.

7. Log-Loss:

Use: Logarithmic loss (log-loss) quantifies the accuracy of the predicted probabilities. Lower log-loss values indicate better model performance.

8. Specificity (True Negative Rate):

Formula: True Negatives / (True Negatives + False Positives)

Use: Specificity measures the ability of the model to correctly identify the negative class (e.g., non-survivors).

The choice of evaluation metrics depends on the specific goals of your project and the relative importance of true positives, true negatives, false positives, and false negatives in the context of Titanic survival prediction. The selection of metrics should consider the project's objectives and the trade-offs between different metrics.

Logistic Regression

It is a widely used statistical and machine learning model that is particularly well-suited for binary classification tasks. In the context of predicting Titanic survival, logistic regression is a valuable tool. Here's an overview of what logistic regression is and how it can be applied to this project:

1. Understanding Logistic Regression:

- Classification Model: Logistic regression is primarily used for classification problems. It's designed to predict the probability of an observation belonging to one of two classes (e.g., survived or not survived).
- Log-Odds Transformation: Logistic regression uses a logistic function (sigmoid) to transform the linear combination of input features into a value between 0 and 1, which represents the probability of belonging to the positive class.
- Interpretability: Logistic regression provides interpretable results. You can easily interpret the coefficients of the model to understand the impact of each feature on the prediction.

2. Application to Titanic Survival Prediction:

- Binary Classification: In the Titanic survival prediction project, the objective is binary classification—predicting whether a passenger survived (positive class) or not (negative class).
- Feature Coefficients: Logistic regression estimates coefficients for each feature. These coefficients reveal which features have a significant impact on the likelihood of survival. For example, the model may show that being female has a positive coefficient, indicating a higher likelihood of survival, while being in a lower class may have a negative coefficient, indicating a lower likelihood of survival.
- Model Training: Logistic regression is trained by optimizing its coefficients to minimize a loss function. This involves finding the best-fitting line that separates the two classes based on the input features.

3. Strengths of Logistic Regression:

- Interpretability: Logistic regression provides clear and interpretable results, making it easy to understand how each feature contributes to the prediction.
- Efficiency: It's computationally efficient, making it a good choice for relatively simple models.
- Linear Separation: When the relationship between features and the log-odds of the target variable is roughly linear, logistic regression performs well.
- Probabilistic Predictions: Logistic regression provides probabilities of class membership, which can be used to set different thresholds for making predictions, depending on the problem's requirements.

4. Limitations of Logistic Regression:

- Assumptions: Logistic regression assumes that the relationship between the features and the log-odds of the target variable is linear. This may not hold true in all cases.
- Complex Relationships: Logistic regression may struggle to capture complex relationships between features and the target variable, which more flexible models like decision trees or neural networks can handle.

5. Model Evaluation: After training a logistic regression model for the Titanic survival prediction, you can use various evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC to assess its performance.

Hypothesis:

Creating hypotheses is an essential part of any data analysis or machine learning project, including the Titanic survival prediction project. Hypotheses serve as educated guesses or expectations about the relationships between variables and outcomes. Here are some hypotheses that you can formulate for your project:

1. Gender Hypothesis:

- Null Hypothesis (H0): Gender does not significantly influence a passenger's chance of survival.
- Alternative Hypothesis (H1): Gender has a significant impact on a passenger's likelihood of survival. Specifically, female passengers are more likely to survive compared to male passengers.

2. Age Hypothesis:

- Null Hypothesis (H0): Age has no substantial effect on a passenger's probability of survival.
- Alternative Hypothesis (H1): Age is a significant factor in determining a passenger's likelihood of survival. Children and young adults may have a higher chance of survival compared to older individuals.

3. Class Hypothesis:

- Null Hypothesis (H0): Passenger class (e.g., 1st, 2nd, 3rd) does not impact the odds of survival.
- Alternative Hypothesis (H1): Passenger class is a crucial determinant of survival. Passengers in higher classes are more likely to survive compared to those in lower classes.

4. Family Size Hypothesis:

- Null Hypothesis (H0): Family size, including the number of siblings/spouses and parents/children, does not significantly influence a passenger's survival.
- Alternative Hypothesis (H1): Family size is a critical factor in predicting survival. Passengers with larger families on board may have a higher likelihood of surviving.

5. Fare Hypothesis:

- Null Hypothesis (H0): The fare paid by a passenger does not significantly impact their chance of survival.
- Alternative Hypothesis (H1): Fare is a significant predictor of survival. Passengers who paid higher fares may have a greater probability of survival.

6. Embarked Port Hypothesis:

- Null Hypothesis (H0): The port from which a passenger embarked does not have a notable effect on their survival odds.
- Alternative Hypothesis (H1): The port of embarkation plays a role in determining survival. Passengers who embarked from certain ports may be more likely to survive.

Source Code:

1. Necessary Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Loading and Reading Dataset

```
df=pd.read_csv('train_and_test2.csv')
df.head()
```

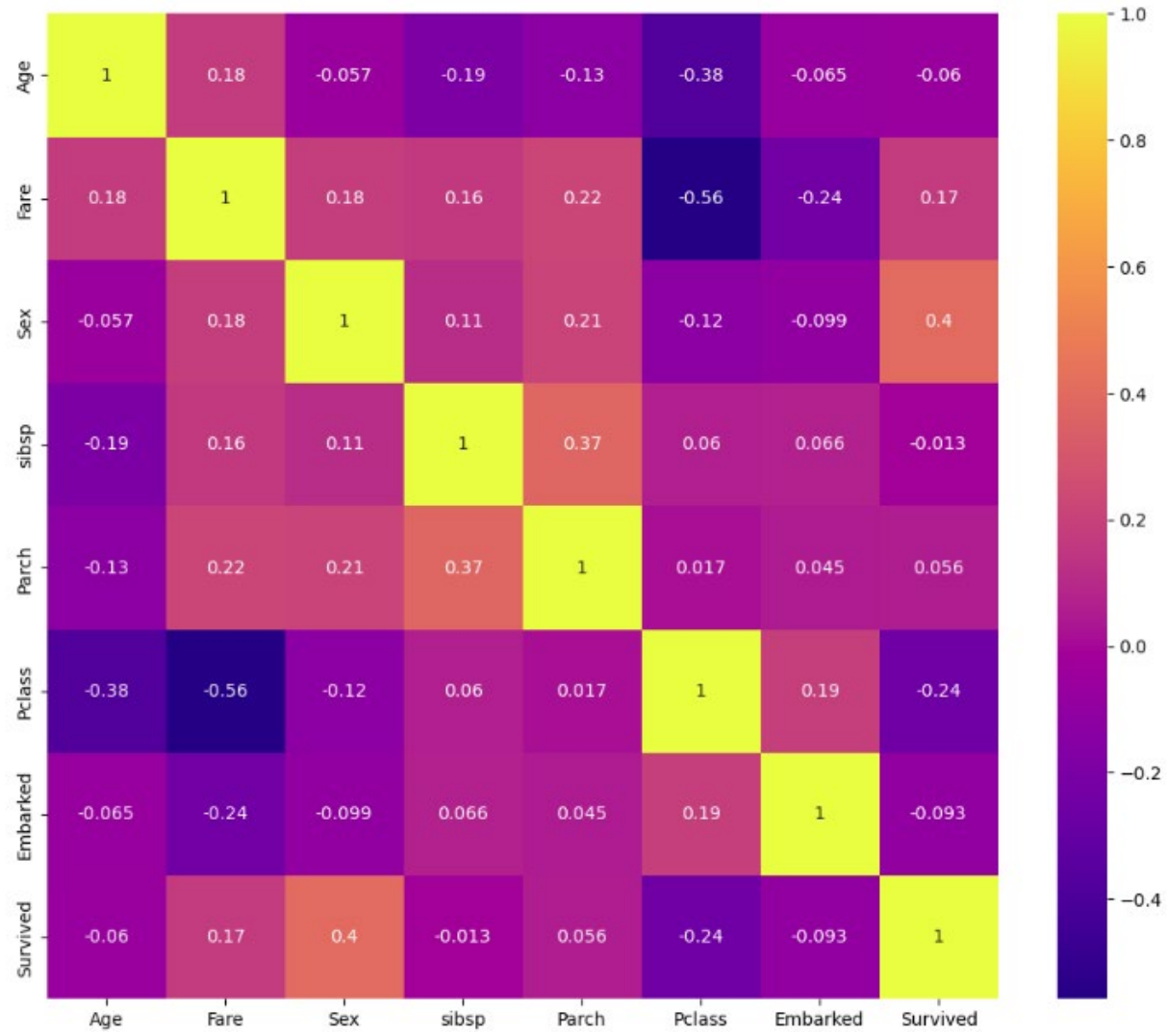
	Passengerid	Age	Fare	Sex	sibsp	zero	zero.1	zero.2	zero.3	zero.4	...
0	1	22.0	7.2500	0	1	0	0	0	0	0	...
1	2	38.0	71.2833	1	1	0	0	0	0	0	...
2	3	26.0	7.9250	1	0	0	0	0	0	0	...
3	4	35.0	53.1000	1	1	0	0	0	0	0	...
4	5	35.0	8.0500	0	0	0	0	0	0	0	...

3. Normalizing Dataset

```
df.drop(['Passengerid','zero','zero.1','zero.2','zero.3','zero.4','zero.5',
df.rename(columns={'2urvived':'Survived'},inplace=True)
df.head()
```

	Age	Fare	Sex	sibsp	Parch	Pclass	Embarked	Survived
0	22.0	7.2500	0	1	0	3	2.0	0
1	38.0	71.2833	1	1	0	1	0.0	1
2	26.0	7.9250	1	0	0	3	2.0	1
3	35.0	53.1000	1	1	0	1	2.0	1
4	35.0	8.0500	0	0	0	3	2.0	0

```
df.dropna(inplace=True)
```



```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
X=df.drop(['Survived'],axis=1)
Y=df['Survived']
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.20,random_state=1)
```

1. Logistic Regression

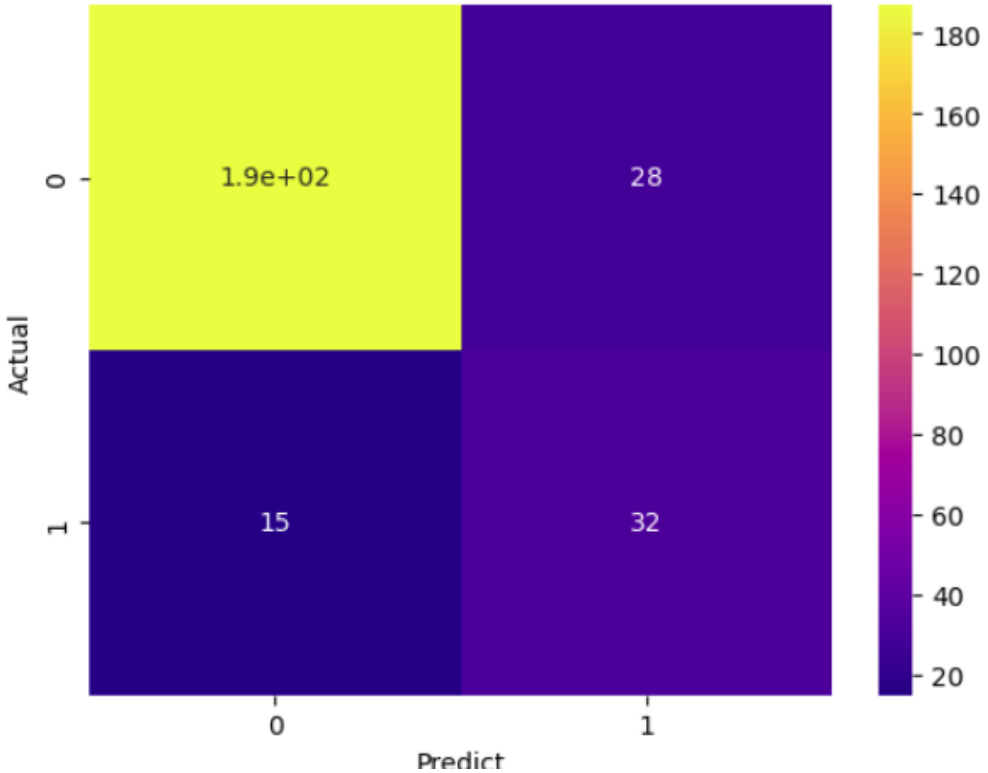
```
from sklearn.linear_model import LogisticRegression
lr=LogisticRegression(max_iter=300)
lr.fit(X_train,y_train)
yhat_lr=lr.predict(X_test)
print("Accuracy of Logistic Model is:",accuracy_score(yhat_lr,y_test))
```

```
ccuracy of Logistic Model is: 0.8358778625954199
```

Metrics

```
from sklearn.metrics import accuracy_score,confusion_matrix
ax=confusion_matrix(yhat_lr,y_test)
sns.heatmap(ax,annot=True,cmap=plt.cm.plasma)
plt.xlabel('Predict')
plt.ylabel('Actual')
```

```
Text(50.722222222222214, 0.5, 'Actual')
```



2. K-Nearest Neighbours

```
.3]:
from sklearn.neighbors import KNeighborsClassifier
KN=KNeighborsClassifier(n_neighbors=5)
KN.fit(X_train,y_train)
yhat=KN.predict(X_test)
print("Accuracy of K-Nearest Neighbor Model is:",accuracy_score(yhat,y_test))
```

```
Accuracy of K-Nearest Neighbor Model is: 0.7480916030534351
```

3. Decision Tree

```
.4]:
from sklearn.tree import DecisionTreeClassifier
tree=DecisionTreeClassifier(random_state=0)
tree.fit(X_train,y_train)
yhat=tree.predict(X_test)
print("Accuracy of Decision Tree Classifier Model is:",accuracy_score(yhat,y_test))
```

```
Accuracy of Decision Tree Classifier Model is: 0.7633587786259542
```

Conclusion

The best model is logistic regression with accuracy of 83.5%

Conclusion:

In our Titanic survival prediction project, we explored various machine learning models, including logistic regression, k-nearest neighbors (K-NN), and decision trees, to determine the most suitable approach for predicting passenger survival. After a comprehensive analysis and evaluation of these models, we found that logistic regression outperformed K-NN and decision trees in this specific task.

References:

- www.geekforgeeks.org
- Techknowledge Machine Learning Textbook
- www.kaggle.com
- www.stackoverflow.com