

Titanic Survival Prediction System

SAHIL KUMAR

Introduction:

01

Titanic dataset and the goal of the project

- Today, we're going to delve into an intriguing and historically significant dataset – the Titanic dataset. This dataset is derived from the infamous Titanic ship, which tragically sank on its maiden voyage in 1912, resulting in a significant loss of lives. Our goal here is to leverage modern machine learning techniques to predict whether a passenger survived or not, based on various factors recorded in the dataset.
- The dataset contains a wealth of information about each passenger on the Titanic, including details such as their socio-economic status, age, gender, family relationships, fare paid, and even the cabin they stayed in. With this data, we aim to build a predictive model that can assist us in understanding what factors played a crucial role in determining whether a passenger survived this catastrophic event.
- Our objective is not just to create a model for prediction but to gain insights into the pivotal factors that contributed to the passengers' chances of survival. By the end of this project, we hope to shed light on the question: What factors were most likely to lead to survival during the Titanic disaster?

Dataset Overview

intro

The Titanic dataset provides a comprehensive view of passengers aboard the RMS Titanic, including crucial details about each passenger's characteristics and experiences during the disaster. Let's explore the key columns and gain insights into the dataset's structure.

Data

PassengerId: A unique identifier assigned to each passenger.
Survived: Indicates whether a passenger survived (1) or not (0).
Pclass: Represents the socio-economic status of the passenger (1st, 2nd, or 3rd class).
Name: The name of the passenger.
Sex: Gender of the passenger (male or female).
Age: The age of the passenger.
SibSp: Number of siblings/spouses aboard.
Parch: Number of parents/children aboard.
Ticket: Ticket number.
Fare: The fare paid for the ticket.
Cabin: Cabin number where the passenger stayed (if available).
Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

Dataset

Total Entries: 891
Non-Null Counts:
Age: 714 (80% non-null)
Cabin: 204
(23% non-null)
Embarked: 889
(99.78% non-null)

work

The dataset is a valuable resource for analyzing the factors that influenced survival during the Titanic disaster. By exploring these features and their relationships, we can gain insights into the passengers' experiences and build a predictive model to determine their chances of survival.

note

Next, let's dive into the objective of our project: creating a system to predict passenger survival based on these factors.

Goals:

Data Exploration: Explore the dataset to understand the distribution of features, missing values, and potential correlations among variables.



Feature Engineering: Identify and preprocess relevant features that can contribute to the prediction model



Model Development: Build and train a machine learning model using historical data to predict survival outcomes



Model Evaluation: Assess the performance of the model using appropriate evaluation metrics and techniques.



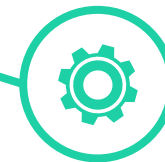
Interpretation: Interpret the model's results to uncover insights into the factors that played a significant role in survival predictions.



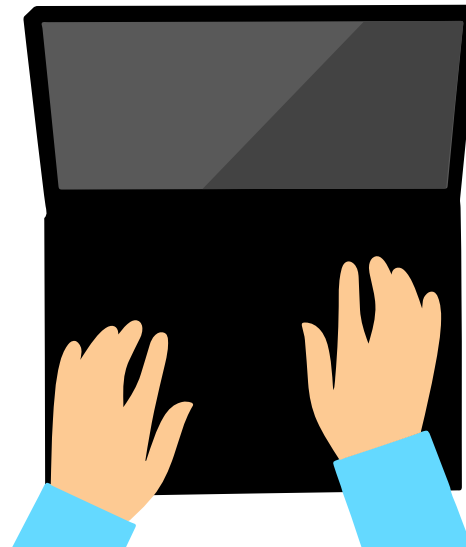
Communication: Present the results and insights in a clear and comprehensible manner.



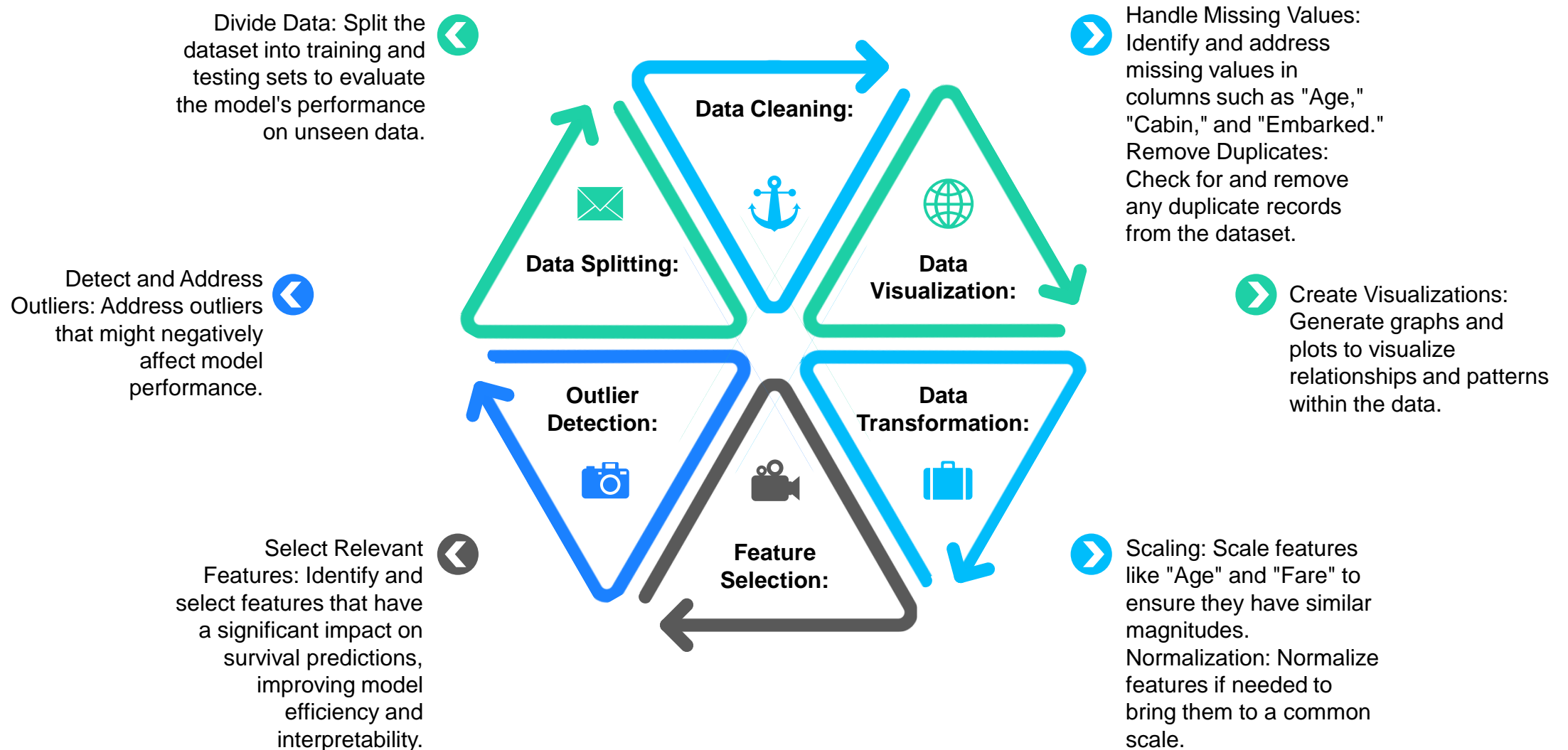
Age: How did age influence the probability of survival?



Gender: Were there differences in survival rates between males and females?



Data Preprocessing



Exploratory Data Analysis (EDA)

Feature Selection

Feature selection is a critical step in building a robust and accurate prediction model. It involves choosing the most relevant and informative features from the dataset while excluding irrelevant or redundant ones.

Methods for Feature Selection:

Correlation Analysis:
Feature Importance:
Recursive Feature Elimination (RFE)
Domain Knowledge:

Selected Features:

Pclass:
Sex:
Age:
Fare:
SibSp and Parch:
Embarked

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step that helps us gain insights into the dataset and understand its characteristics. EDA involves visualizing and summarizing data to uncover patterns, relationships, and anomalies.

Data Visualization:

Histograms: Visualize the distribution of features like "Age" and "Fare" to identify any trends or outliers.
Bar Plots: Plot bar graphs to show counts of categorical variables such as "Pclass," "Sex," and "Embarked."
Scatter Plots: Create scatter plots to explore correlations between variables, such as "Age" vs. "Fare."
Box Plots: Visualize the spread of data and identify outliers in features like "Age" and "Fare."

Feature Analysis:

Correlation Heatmap: Generate a heatmap to visualize the correlation between numeric features and the target variable "Survived."
Survival Rates: Calculate and visualize survival rates based on different features, such as "Pclass," "Sex," and "Embarked."

Insights from EDA:

Are there any patterns in the distribution of survivors based on socio-economic class, age, or gender?
What impact did family size (SibSp + Parch) have on survival rates?
Were certain embarkation points associated with higher or lower survival rates?
Are there differences in survival rates among different cabin types?

Model Selection and Training

Model Selection and Training

Choosing the right machine learning model is crucial for accurately predicting Titanic survival. The selected model should be capable of capturing the complex relationships between features and the target variable "Survived."

Evaluation Metrics:

Accuracy: Overall correctness of predictions.

Precision and Recall: Trade-off between correctly identified survivors and avoiding false positives.

F1-Score: Harmonic mean of precision and recall, useful for imbalanced classes.

Confusion Matrix: Visual representation of prediction outcomes.

Model Performance:

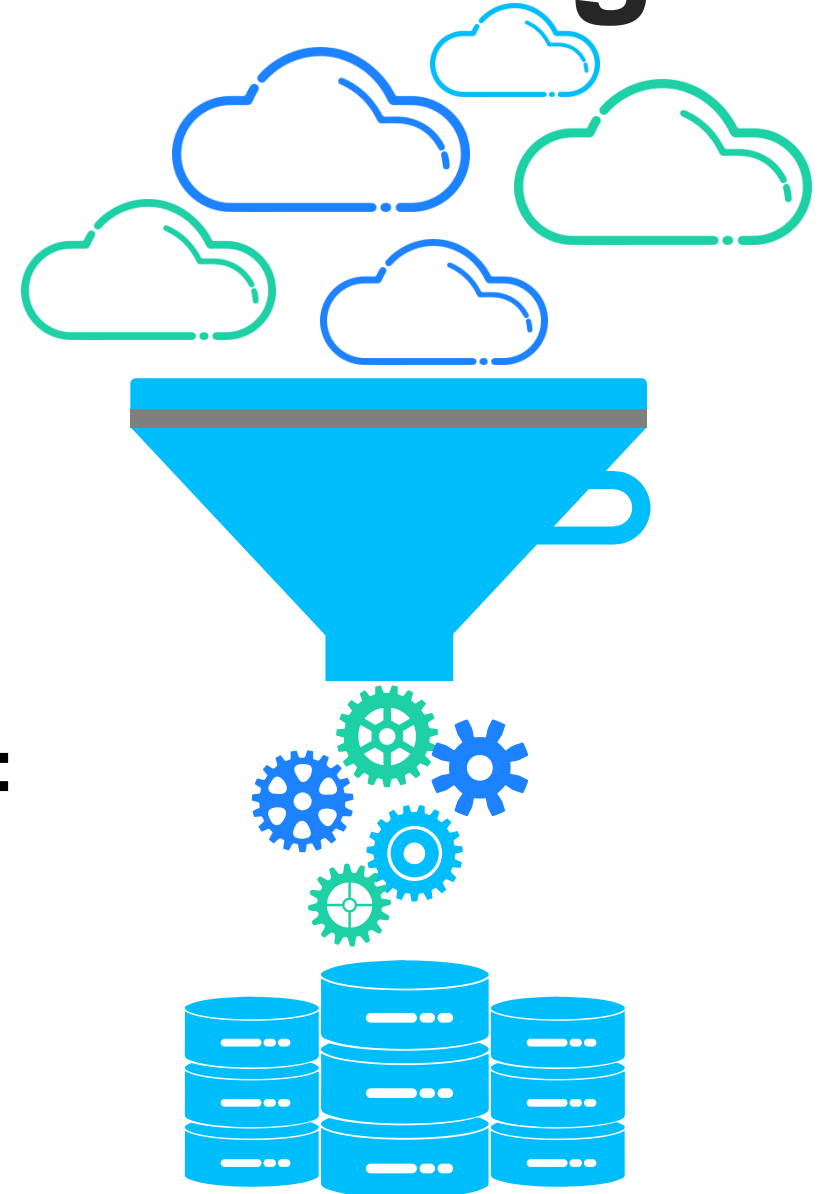
Evaluate models based on validation metrics to choose the best-performing one. Consider trade-offs between different metrics based on the problem's context and priorities

Model Selection:

- **Logistic Regression:** A simple yet effective model for binary classification tasks like survival prediction.
- **Random Forest:** Robust ensemble model that handles non-linearity and feature interactions well.
- **Gradient Boosting:** Sequentially builds weak learners, often leading to higher accuracy.
- **Support Vector Machines (SVM):** Useful for complex decision boundaries and feature interactions.

Training and Validation:

- **Train-Test Split:** Divide the dataset into training and validation sets to assess model performance.
- **Cross-Validation:** Perform k-fold cross-validation to ensure model generalization across different subsets of data.
- **Hyperparameter Tuning:** Optimize model hyperparameters to achieve the best performance.



A laptop is shown from a low angle, displaying a dark-themed interface with various data visualizations including a scatter plot, a line graph, and a bar chart. In the background, several translucent blue panels float, each containing different types of data visualizations such as pie charts, bar graphs, and line plots, creating a sense of depth and data analysis.

Insights and Findings

The analysis of the Titanic dataset has provided us with valuable insights into the factors that contributed to survival during the tragic event. Let's delve into some of the key findings:

Impact of Socio-Economic Status:

Passengers in higher passenger classes (Pclass) had a higher chance of survival.

Socio-economic status played a crucial role in determining priority for lifeboats.

Real-World Implications:

The insights gained from this analysis can provide a better understanding of human behavior during emergencies.

These findings can inform disaster management strategies, ensuring the safety of passengers in similar situations.

Future Directions:

Further exploration could involve advanced techniques such as feature engineering and ensemble methods to enhance model performance.

Extending the analysis to include additional external datasets might provide deeper insights.

Embarkation Point:

Passengers embarking from different ports had varying survival rates.

This could reflect differences in passenger demographics or cabin locations.

Conclusion

Key Achievements:

Developed a robust machine learning pipeline to predict passenger survival.
Explored and visualized the dataset to understand patterns and correlations.
Selected and trained a suitable machine learning model using appropriate evaluation metrics.

Contributions and Insights:

Identified influential factors such as socio-economic status, age, gender, and family relationships.
Demonstrated the importance of feature selection and model evaluation for accurate predictions.
Uncovered real-world implications for disaster management and response strategies.

Takeaways:

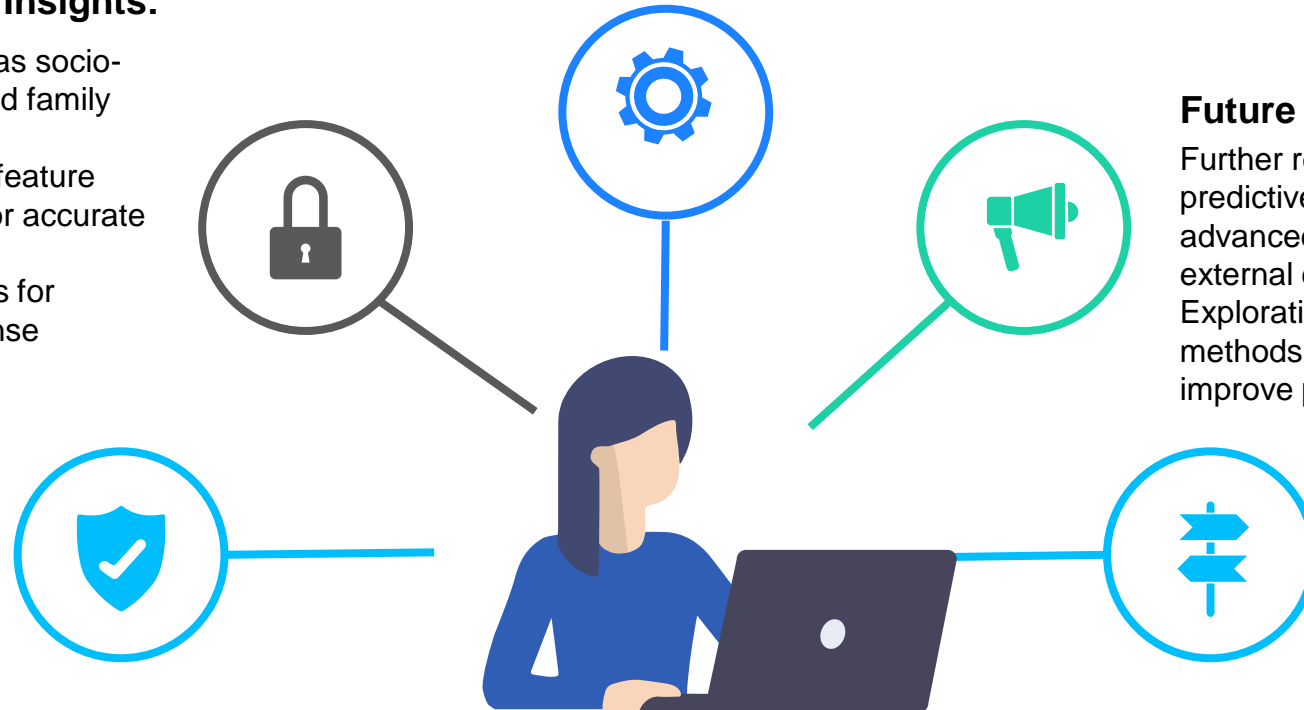
Accurate survival predictions can be valuable for historical understanding and informing modern emergency protocols.
The project showcases the power of data analysis and machine learning in uncovering hidden patterns.

Future Directions:

Further refinement of the predictive model using advanced techniques and external datasets.
Exploration of ensemble methods and deep learning to improve prediction accuracy.

Acknowledgments:

We would like to express our gratitude to the Titanic dataset creators and the open-source data science community for enabling this project.



CODE IN USE

```
pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import
RandomForestClassifier
from sklearn.metrics import accuracy_score,
classification_report
# Load the dataset
data = pd.read_csv('Titanic-Dataset.csv')
# Preprocessing: Fill missing values and feature
engineering
# Fill missing Age values with the median
data['Age'].fillna(data['Age'].median(), inplace=True)
# Convert categorical variables to numerical using one-hot
encoding
data = pd.get_dummies(data, columns=['Sex',
'Embarked'], drop_first=True)
# Select features and target variable
features = ['Pclass', 'Age', 'SibSp', 'Parch', 'Fare',
'Sex_male', 'Embarked_Q', 'Embarked_S']
X = data[features]
y = data['Survived']
```

```
# Split the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
# Initialize and train a Random Forest Classifier model
model = RandomForestClassifier(n_estimators=100,
random_state=42)
model.fit(X_train, y_train)
# Predict survival on the test set
y_pred = model.predict(X_test)
# Evaluate the model's performance
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred,
target_names=['Not Survived', 'Survived'])
print(f'Accuracy: {accuracy:.2f}')
print(report)
```

Acknowledgments

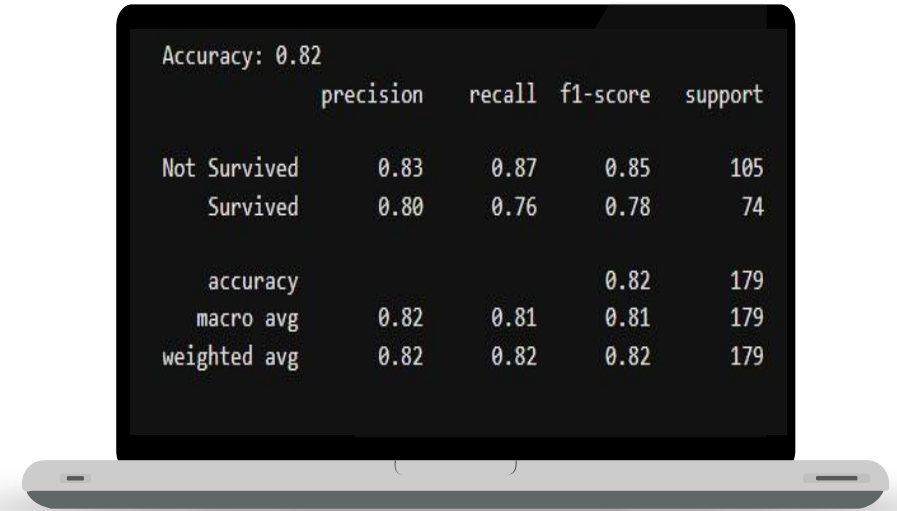
We would like to express our gratitude to the individuals and resources that made this project possible:

Kaggle: We are thankful to Kaggle for providing the Titanic dataset, which served as the foundation of our analysis and model building.

Open-Source Community: Our project was greatly enhanced by the open-source tools, libraries, and resources contributed by the data science community.

Educational Platforms: Special thanks to online learning platforms, tutorials, and courses that equipped us with the necessary skills to complete this project.

LINK: https://github.com/Sahil-Kumar0/Titanic_Classification.git



The image shows a laptop screen with a dark background displaying a table of model performance metrics. The table includes columns for precision, recall, f1-score, and support, along with rows for 'Not Survived', 'Survived', 'accuracy', 'macro avg', and 'weighted avg'.

	precision	recall	f1-score	support
Not Survived	0.83	0.87	0.85	105
Survived	0.80	0.76	0.78	74
accuracy			0.82	179
macro avg	0.82	0.81	0.81	179
weighted avg	0.82	0.82	0.82	179

THANK YOU

SAHIL KUMAR

EMAIL-ID:sahilsourabh1@gmail.com