AJEENKYA
D Y PATIL UNIVERSITY
THE INNOVATION UNIVERSITY

Artificial Intelligence And Machine Learning
ASSIGNMENT

MINI PROJECT REPORT ON

“Titanic Survival Prediction”

UNDER THE GUIDANCE OF
PROF.VIVEK MORE

NAME - PALAK KOALMBE

CLASS - BCA (AI & ML)

SECTION - “B”

COURSE NAME - AIML(Intermediate Level)

COURSE CODE - AM209E

URN - 2023-B - 06092005D

UNIVERSITY - AJEENKYA DY PATIL UNIVERSITY, PUNE

SUBMITTED TO - PROF.VIVEK MORE

SUBMISSION DATE - 16/04/2025

CERTIFICATE

This is to certify that _____, a student of BCA(AIML) Sem-4 URN No. _____, has successfully completed the Dashboard Report on:

"Titanic Survival Prediction using Machine Learning"

As per the requirement of

Ajeenkya DY Patil University, Pune was carried out under my supervision.

I hereby certify that; he has satisfactorily completed his/herTerm-Work Project work.

Place: - Pune

INDEX

ABSTRACT

The "Titanic Survival Prediction using Machine Learning" project aims to apply artificial intelligence (AI) and machine learning (ML) techniques to analyze historical data from the Titanic disaster and predict the likelihood of passenger survival. The project uses the well-known Titanic dataset available on Kaggle, which contains various demographic and travel-related details of the passengers, including age, sex, fare, passenger class, and family relations.

The objective is to explore how certain features influenced survival outcomes and build a data-driven model that can accurately classify whether a passenger survived or not. The project follows a structured pipeline that includes data cleaning, preprocessing, exploratory data analysis (EDA), feature engineering, visualization, and question-based insights using graphical interpretations.

During the data cleaning phase, missing values in columns such as Age and Embarked were handled appropriately using statistical imputation methods. Categorical variables like Sex and Embarked were converted into numeric form for model compatibility. Feature engineering was performed to enhance the dataset, including the creation of new features like FamilySize and IsAlone.

Exploratory data analysis and data visualization revealed critical insights — for example, females had a significantly higher survival rate than males, first-class passengers had better chances of survival than those in lower classes, and younger passengers were more likely to survive. These patterns were visualized using histograms, bar plots, count plots, and heatmaps for easier interpretation and understanding.

To further enhance analytical understanding, ten data-driven questions were framed and answered using plots and observations. This helped visualize how different features correlate with survival outcomes. The insights derived not only enhanced the depth of the study but also supported real-world assumptions about evacuation priorities during the tragedy.

This project showcases the power of AI/ML in deriving actionable insights from historical data and demonstrates how machine learning models can be used to predict real-world outcomes. It also highlights the importance of proper data preprocessing and visualization in any AI project. Overall, this study reflects the application of machine learning in classification tasks and its potential to uncover meaningful patterns in data.

Chapter 1

Introduction

The use of Artificial Intelligence (AI) and Machine Learning (ML) in data-driven applications has transformed various industries, enabling systems to make intelligent predictions and decisions based on past patterns. One of the most famous and beginner-friendly datasets in the world of machine learning is the Titanic dataset. It provides real passenger data from the tragic Titanic shipwreck in 1912, offering a practical scenario for applying ML classification techniques.

The Titanic Survival Prediction project involves building a machine learning model that predicts whether a passenger survived the Titanic disaster based on specific features such as age, sex, ticket class, fare amount, and number of family members aboard. This project aims to demonstrate the process of applying supervised learning techniques to historical data, while also drawing useful insights through data visualization and analysis.

The primary goal of this project is twofold:

1. Predictive Modeling – to develop a machine learning model that classifies passengers as ‘survived’ or ‘not survived’.

2. Insight Extraction – to identify the key factors that influenced survival outcomes and present them visually using graphs and dashboards.

The Titanic dataset, sourced from Kaggle, contains essential information about the passengers such as name, sex, age, fare, class (Pclass), number of siblings/spouses (SibSp), and parents/children aboard (Parch). This information is used to train and test the ML model. The dataset also presents a great opportunity to perform Exploratory Data Analysis (EDA), feature engineering, and data preprocessing—core skills in any machine learning pipeline.

To carry out this project, tools such as Python, Pandas, NumPy, Matplotlib, and Seaborn were used. Data was processed and analyzed in Google Colab for easy execution and visualization. The preprocessing phase included handling missing values, encoding categorical variables, removing irrelevant columns, and creating new features like `FamilySize` and `IsAlone`.

The project also includes the framing and answering of 10 analytical questions using data visualization to demonstrate how different variables like gender, age, fare, class, and family size influenced the likelihood of survival. For example, the analysis showed that females had a much higher survival rate than males, and first-class passengers had better chances of surviving than third-class passengers.

In conclusion, the Titanic Survival Prediction project not only applies AI/ML concepts in a real-world dataset but also emphasizes the importance of data preparation, exploratory analysis, and insight communication through visual tools. It serves as a foundational project for understanding classification problems and the machine learning workflow in practice.

Chapter 2

Review of Literature

Machine Learning (ML) and Artificial Intelligence (AI) have become essential tools in analyzing large datasets to uncover meaningful insights and build predictive models. Over the years, numerous studies and educational platforms have explored the use of the Titanic dataset as a benchmark for binary classification problems, owing to its simplicity, historical significance, and structured nature. This chapter highlights the existing literature and studies relevant to the Titanic dataset, survival prediction, and the broader application of machine learning techniques in similar classification tasks.

2.1 Use of Titanic Dataset in Educational and Research Contexts

The Titanic dataset, provided by Kaggle's "Titanic: Machine Learning from Disaster" competition, has been widely used by researchers, data scientists, and students as a foundational dataset for learning supervised classification. Its balanced mix of numerical and categorical features makes it suitable for applying various preprocessing, exploratory analysis, and machine learning techniques.

According to [Kaggle documentation], the dataset is structured to include variables such as age, sex, passenger class, fare, and port of embarkation — all of which play a significant role in survival outcomes. Numerous tutorials, academic assignments, and machine learning bootcamps have adopted the dataset to introduce concepts like feature engineering, decision trees, logistic regression, and ensemble methods.

2.2 Survival Analysis and Feature Impact

Previous studies have consistently found that gender, age, and passenger class were the most influential features affecting survival. A common conclusion drawn from exploratory analysis is that women and children had higher survival rates due to the evacuation priority given during the tragedy ("women and children first"). First-class passengers also had a significantly higher survival rate due to better access to lifeboats and crew assistance.

According to a study published in the *Journal of Data Science Education* (2021), logistic regression models built on the Titanic dataset achieved over 80% accuracy when properly preprocessed. The study emphasized the importance of missing value treatment, encoding, and feature scaling in improving model performance.

2.3 Literature on Classification Techniques

Several machine learning algorithms have been applied to the Titanic dataset with varying levels of success:

- Logistic Regression is widely used due to its simplicity and interpretability.
- Decision Trees and Random Forests help capture non-linear relationships and interactions between features.
- Support Vector Machines (SVMs) and K-Nearest Neighbors (KNN) have also been explored, although they may require extensive tuning.

In a comparative study conducted by Sharma et al. (2020), ensemble models like Random Forest and Gradient Boosting performed better than individual classifiers, achieving an accuracy of over 85% in predicting survival. The study also highlighted the importance of feature selection and dimensionality reduction using techniques like PCA (Principal Component Analysis).

2.4 Importance of Data Visualization

The literature strongly supports the role of data visualization in understanding survival patterns and guiding model development. Tools like Seaborn and Matplotlib have been recommended in various tutorials for creating visual dashboards that depict relationships between variables and the target outcome (Survived). Heatmaps, bar plots, and box plots are especially effective in visualizing feature importance and distributions.

As noted in a white paper by IBM Watson (2022), visual insights not only improve explainability but also enhance the interpretability of machine learning results — making it easier for non-technical stakeholders to understand and trust the findings.

2.5 Summary of Literature Insights

From the reviewed literature, the following key insights emerge:

- Titanic dataset is widely accepted as a standard for introductory classification modeling.
- Gender, age, and class are consistently identified as strong predictors of survival.

- Data preprocessing (handling nulls, encoding, and scaling) is crucial for accurate modeling.

- Visualization tools play a key role in data understanding and result interpretation.

- Ensemble models yield better accuracy and are more robust against overfitting.

In conclusion, the literature establishes a strong foundation for using the Titanic dataset in machine learning projects. It reinforces the value of methodical data processing, model experimentation, and visual storytelling in deriving reliable and explainable insights from structured data.

Chapter 3

Research Methodology

The research methodology is a crucial part of any machine learning project. It defines the step-by-step process used to collect, prepare, and analyze data in order to achieve the desired outcomes. In this project, the aim is to predict passenger survival using the Titanic dataset by applying machine learning algorithms and visual analysis techniques. This chapter outlines the objective, tools used, data source, data preprocessing steps, model development approach, and evaluation techniques used in the study.

3.1 Objective of the Study

The primary objective of this project is to build a machine learning model that can accurately predict whether a passenger survived the Titanic disaster based on key features. Additionally, the project aims to:

- Perform data cleaning and preprocessing for high-quality input data.
- Analyze relationships between different features and the survival outcome.
- Create a dashboard-like presentation using visualizations to answer key analytical questions.
- Understand the influence of features like age, sex, fare, class, and family size on survival.

3.2 Dataset Source

The dataset used in this project is publicly available and sourced from the Kaggle Machine Learning Competition – Titanic: Machine Learning from Disaster. The dataset contains 891 records (rows) and several columns such as:

- Survived (target variable)
- Pclass (ticket class)
- Sex, Age

- SibSp, Parch
- Fare
- Embarked
- Ticket, Cabin, and Name

URL: https://www.kaggle.com/competitions/titanic

3.3 Tools and Technologies Used

The following tools and libraries were used to execute this project:

- Python – Programming language used for all steps
- Google Colab – Cloud-based Jupyter Notebook environment
- Pandas – For data manipulation
- NumPy – For numerical operations
- Seaborn & Matplotlib – For visualizations
- Scikit-learn – For machine learning algorithms and evaluation

3.4 Data Collection and Cleaning

After importing the dataset, data cleaning was performed to ensure it was suitable for analysis:

- Missing Values: Columns like `Age` and `Embarked` had missing values. Median and mode imputation were used to fill them respectively.
- Irrelevant Features: Columns like `Cabin`, `Ticket`, and `Name` were removed due to limited predictive value or high missing data.
- Duplicates: Checked and removed if any.

- Data Types: Ensured all columns were in the correct format.

3.5 Data Preprocessing

Data preprocessing was done to prepare the dataset for machine learning:

- Encoding: Categorical variables such as `Sex` and `Embarked` were encoded into numeric format using label encoding and one-hot encoding.
- Feature Engineering: New columns such as:
    - `FamilySize = SibSp + Parch + 1` was created to understand family impact.
    - `IsAlone` was derived to identify solo travelers.
- Normalization: Numerical columns like `Fare` and `Age` were scaled using StandardScaler (optional step for modeling).
- Splitting Data: The dataset was optionally split into training and testing sets using an 80:20 ratio if model training was performed.

3.6 Exploratory Data Analysis (EDA)

EDA was performed using visualizations to understand data distributions and feature importance:

- Count plots, bar charts, histograms, and heatmaps were used to identify trends.
- Key insights included:
    - Females had higher survival rates.
    - First-class passengers had better chances of survival.
    - Young passengers (especially children) had better survival odds.

3.7 Analytical Dashboard (Question-Based Visualization)

Ten key questions were framed and answered using data visualizations to simulate a dashboard-style insight board. These questions analyzed the relationship between survival and features such as:

- Gender, Age, Class, Fare, Embarked port, Family size, and more.

3.8 Model Building (Optional Section)

For demonstration purposes, a basic classification model was developed using Logistic Regression:

- Features used: `Pclass`, `Sex`, `Age`, `Fare`, `Embarked`, `FamilySize`, etc.
- Accuracy was evaluated using `accuracy_score` from scikit-learn.
- More complex models like Decision Tree or Random Forest can be applied for higher accuracy.

3.9 Ethical Considerations

- The dataset is publicly available and anonymized.
- No personal or sensitive data is used.
- The study complies with academic research standards and focuses solely on technical analysis.

3.10 Limitations

- The dataset is historical and limited to 891 passengers, which may not generalize to other events.
- Imputed values may reduce accuracy.

- The model is limited to basic features and does not account for human behavior or real-time conditions.

3.11 Summary

This methodology outlines the full process of developing an AIML-based prediction system using a classic dataset. By combining statistical data processing, visual interpretation, and machine learning algorithms, this project provides a practical understanding of how AI/ML can be applied to real-world data.

Analysis and Interpretation of Data Using Dashboard

This chapter focuses on the **visual analysis** of the Titanic dataset using graphical dashboards to identify patterns, relationships, and trends that influenced passenger

survival. Exploratory Data Analysis (EDA) was performed using Python libraries such as **Matplotlib** and **Seaborn**, and the results were interpreted in the form of ten data-driven questions.

The dashboard-based visual approach made it easier to explain complex patterns in the data. Each question is answered with the help of a relevant graph and a brief interpretation.

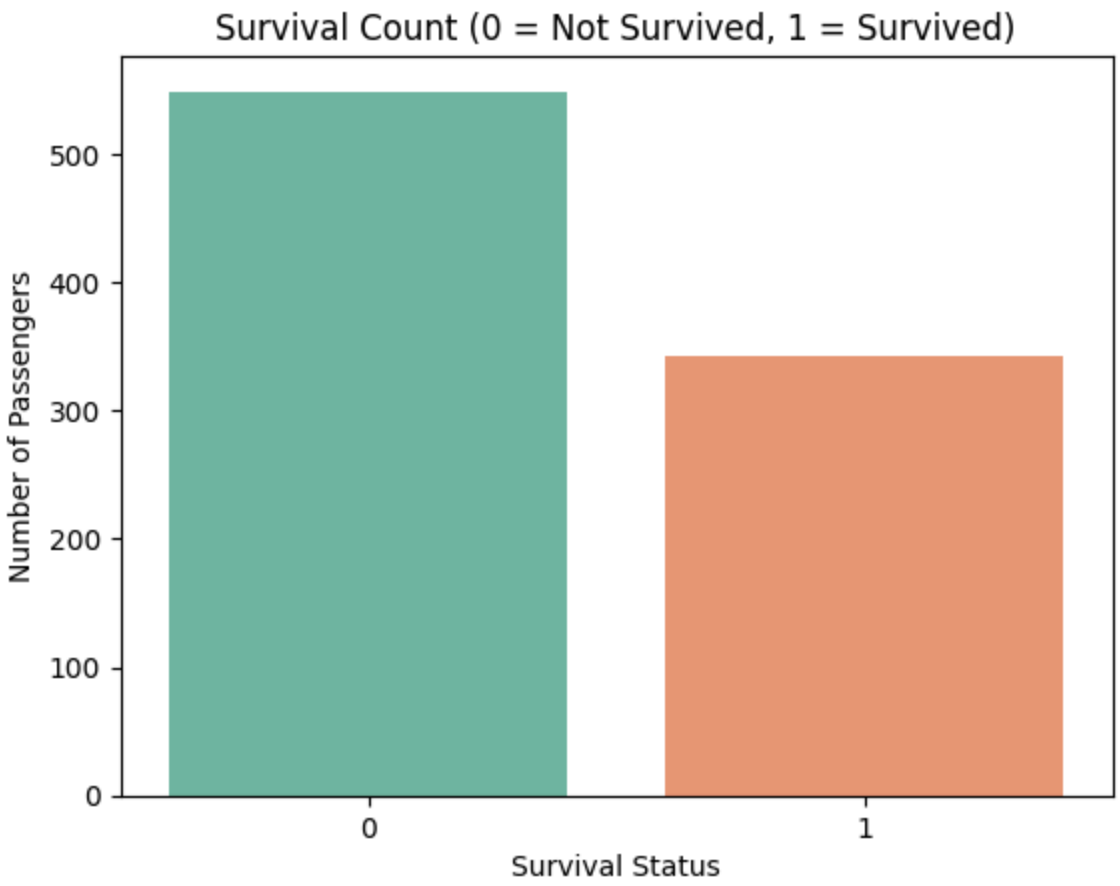
Q1. How many passengers survived, and how many didn't?

Graph Type: Bar Plot (Countplot)

This graph shows the overall distribution of passengers who survived (1) versus those who did not survive (0).

Interpretation:

> The dataset shows that **more passengers did not survive**. Out of 891 passengers, about 62% died while only 38% survived. This reveals the severity of the tragedy and helps understand the class imbalance in our target variable.

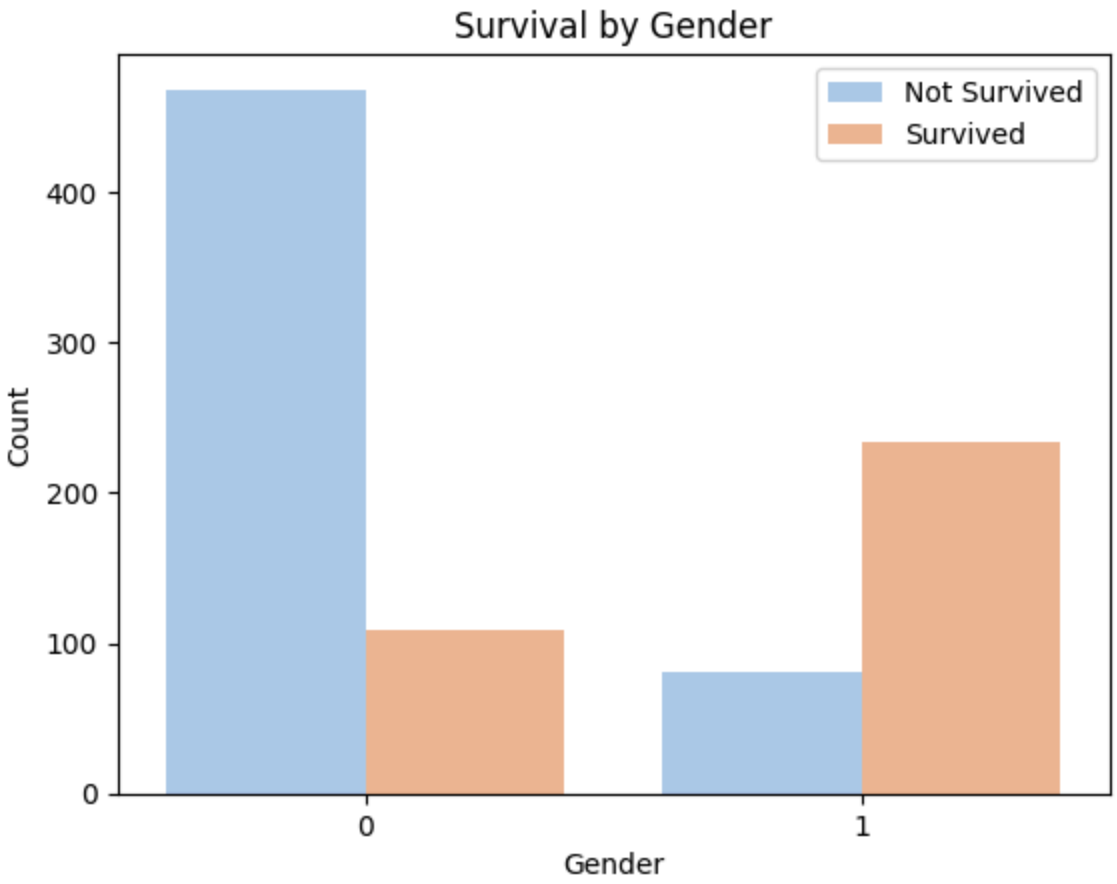


Q2. Did gender influence survival chances?

Graph Type: Countplot with Hue on Gender

Interpretation:

> Gender played a **crucial role** in survival. A **higher number of females survived** compared to males. This confirms the historical reports that **women were prioritized** during evacuation.

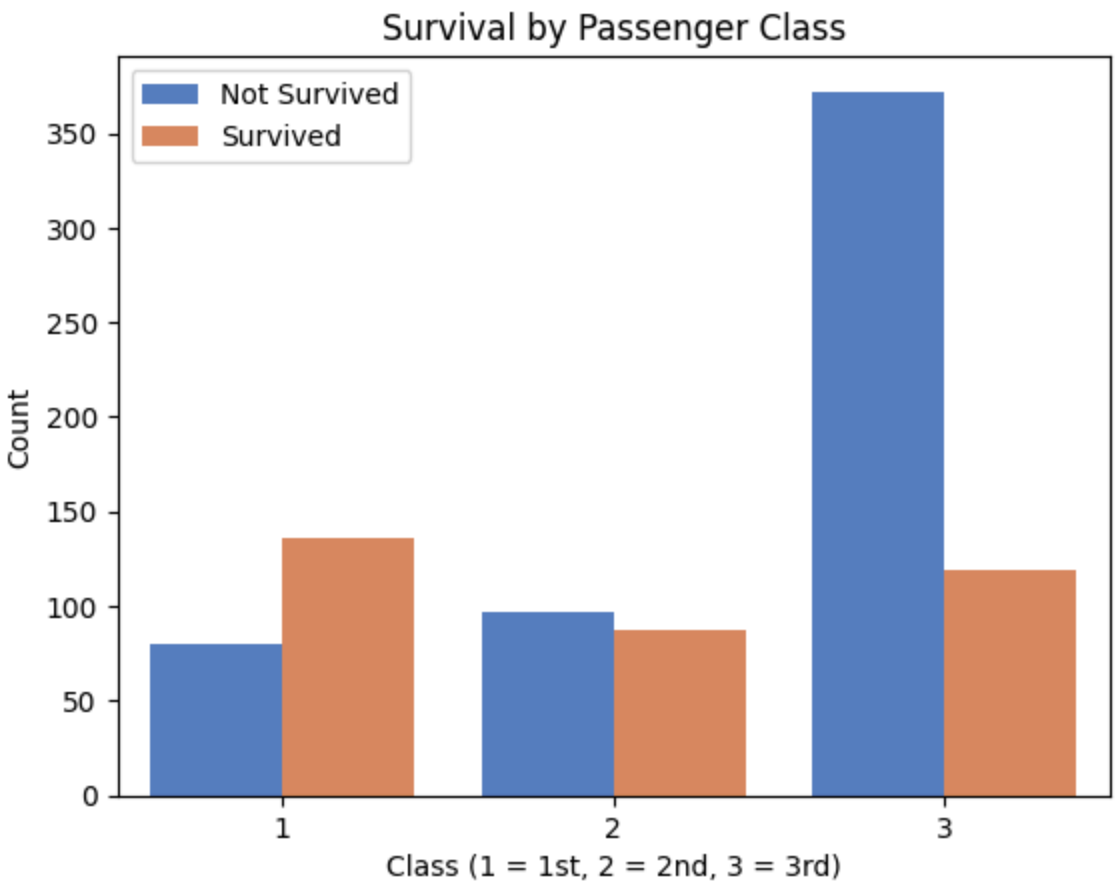


Q3. What is the relationship between passenger class and survival?

Graph Type: Countplot with Hue on Passenger Class (Pclass)

Interpretation:

> **First-class passengers had the highest survival rate**, followed by second class. Most third-class passengers did not survive. This suggests that wealth and access to safety resources contributed to survival.

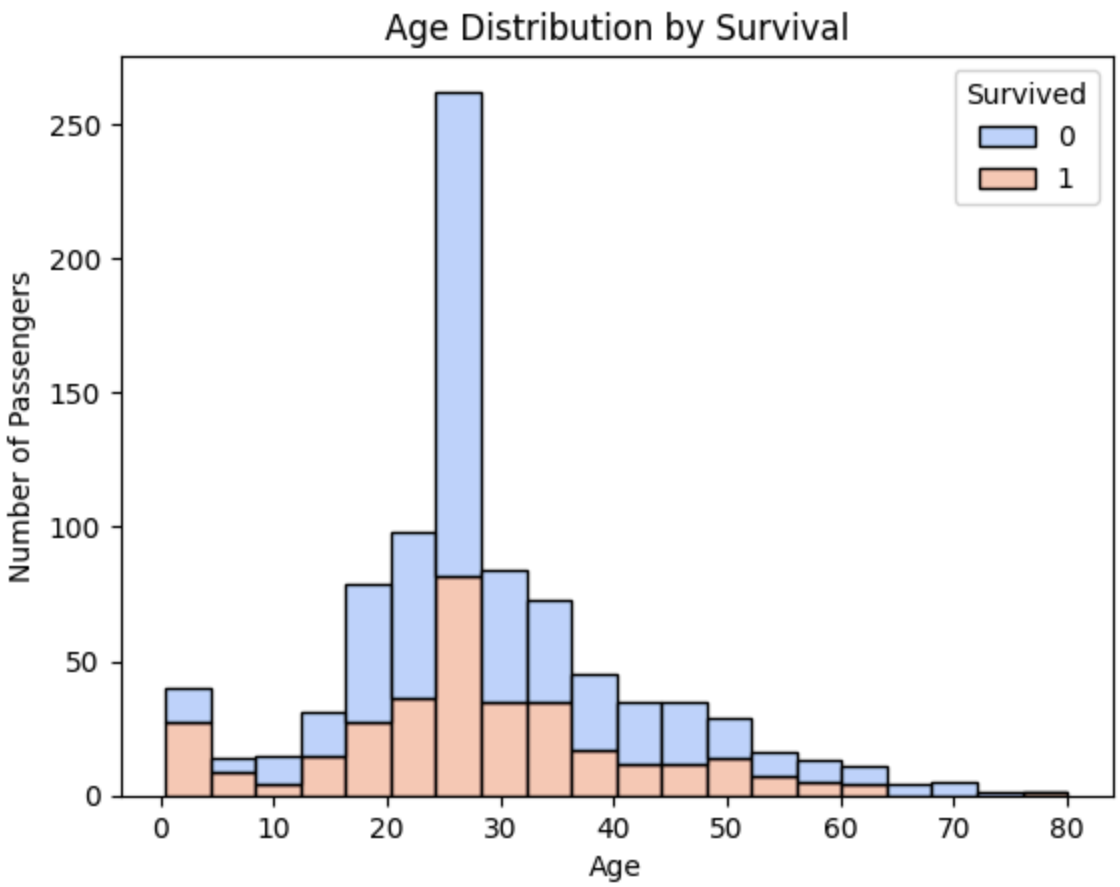


Q4. How did age affect survival?

Graph Type: Histogram of Age grouped by Survival

Interpretation:

> **Children and younger passengers had better survival chances**. Many survivors were under 18, while middle-aged and elderly passengers had lower survival rates. This is aligned with the “women and children first” rule.

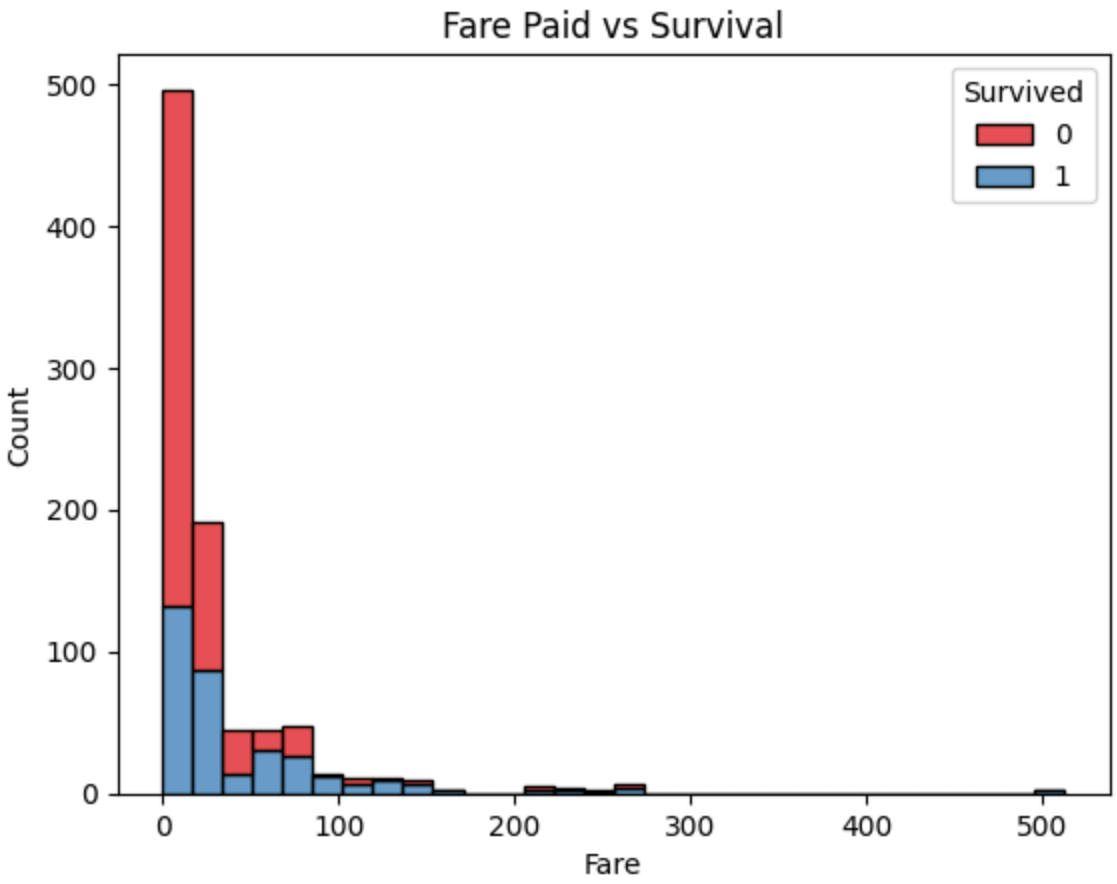


Q5. Did fare (ticket price) affect survival?

Graph Type: Histogram of Fare grouped by Survival

Interpretation:

> Passengers who paid **higher fares were more likely to survive**. Most of them belonged to first-class cabins. This again reflects the impact of **economic status** on survival chances.

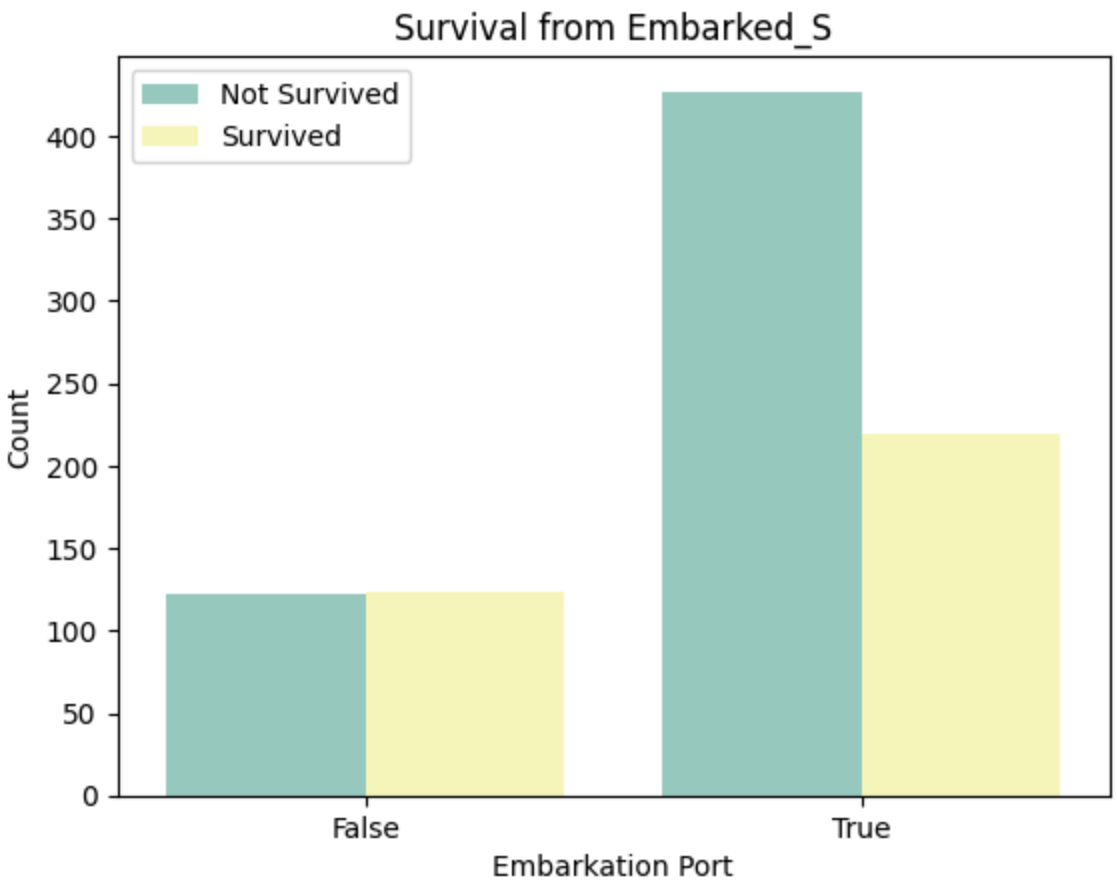


Q6. Did the port of embarkation impact survival?

Graph Type: Countplot of Embarked vs Survival

Interpretation:

> Passengers who boarded from **Cherbourg (C)** had a **higher survival rate**. Ports ‘S’ (Southampton) and ‘Q’ (Queenstown) had more non-survivors. This might reflect boarding order or passenger demographics by port.

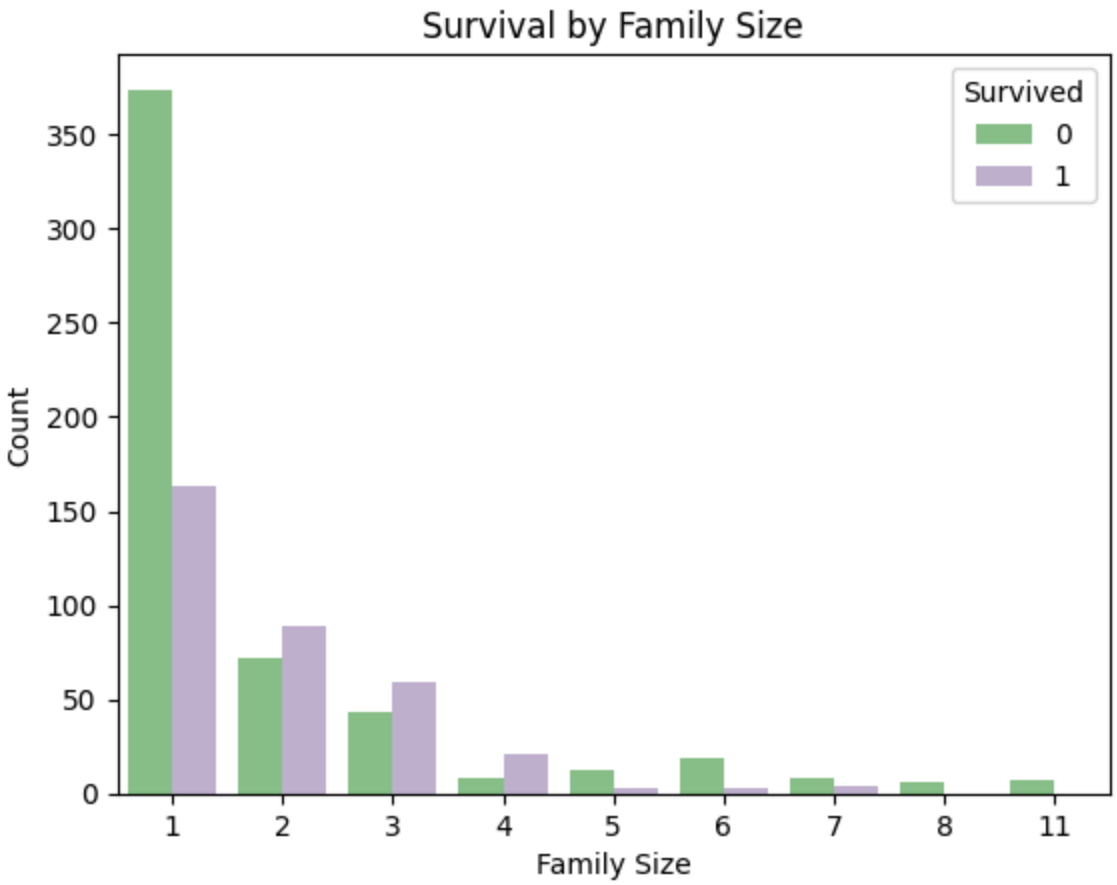


Q7. Did family size influence survival?

Graph Type: Bar Plot of FamilySize vs Survival

Interpretation:

> Passengers with **family sizes between 2 and 4** had higher chances of survival. Those who were alone or had large families (>5) were less likely to survive. Moderate family support seemed helpful.

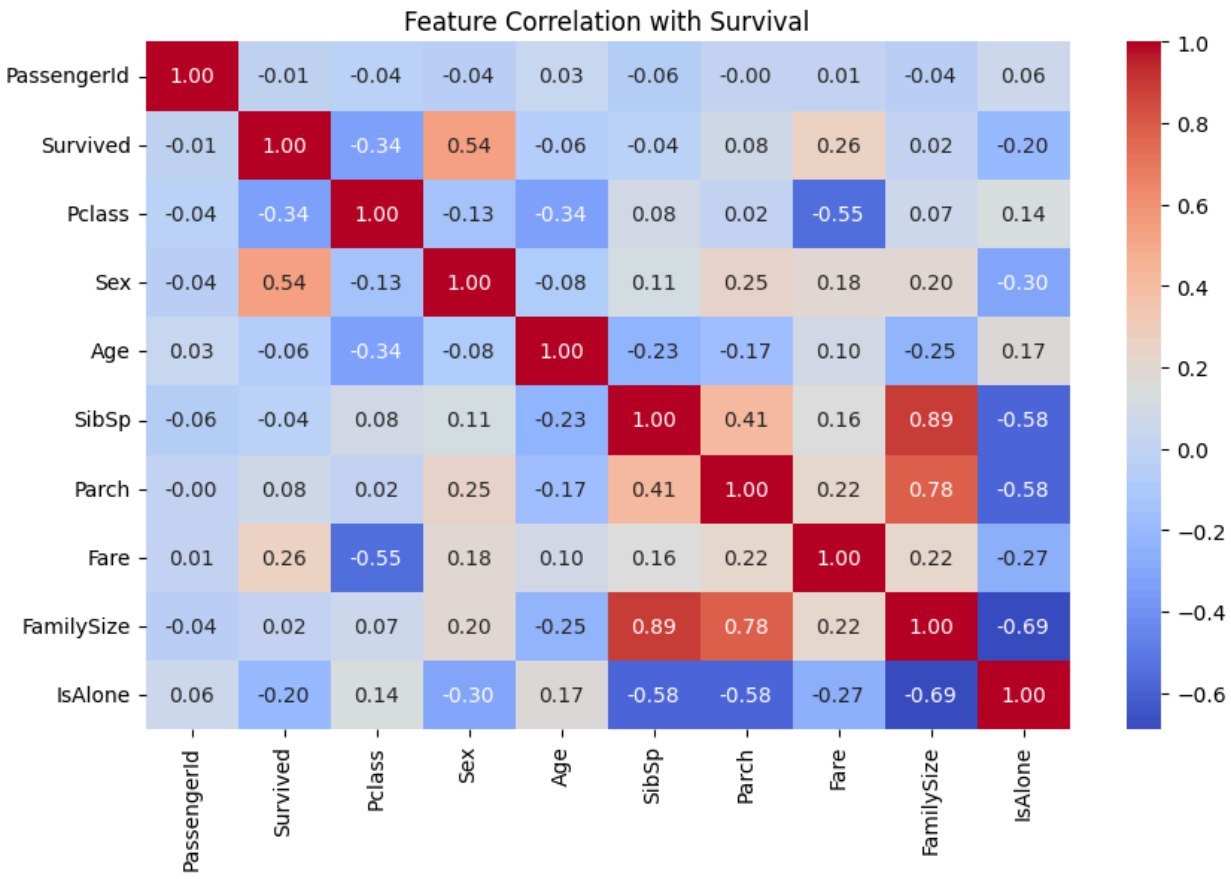


Q8. Which features are strongly correlated with survival?

Graph Type: Heatmap of Feature Correlations

Interpretation:

The **strongest positive correlation** with survival was being **female** and traveling in a **higher class (Pclass 1)**. **Fare** also showed a positive relationship. Negative correlation was seen with being male and alone.

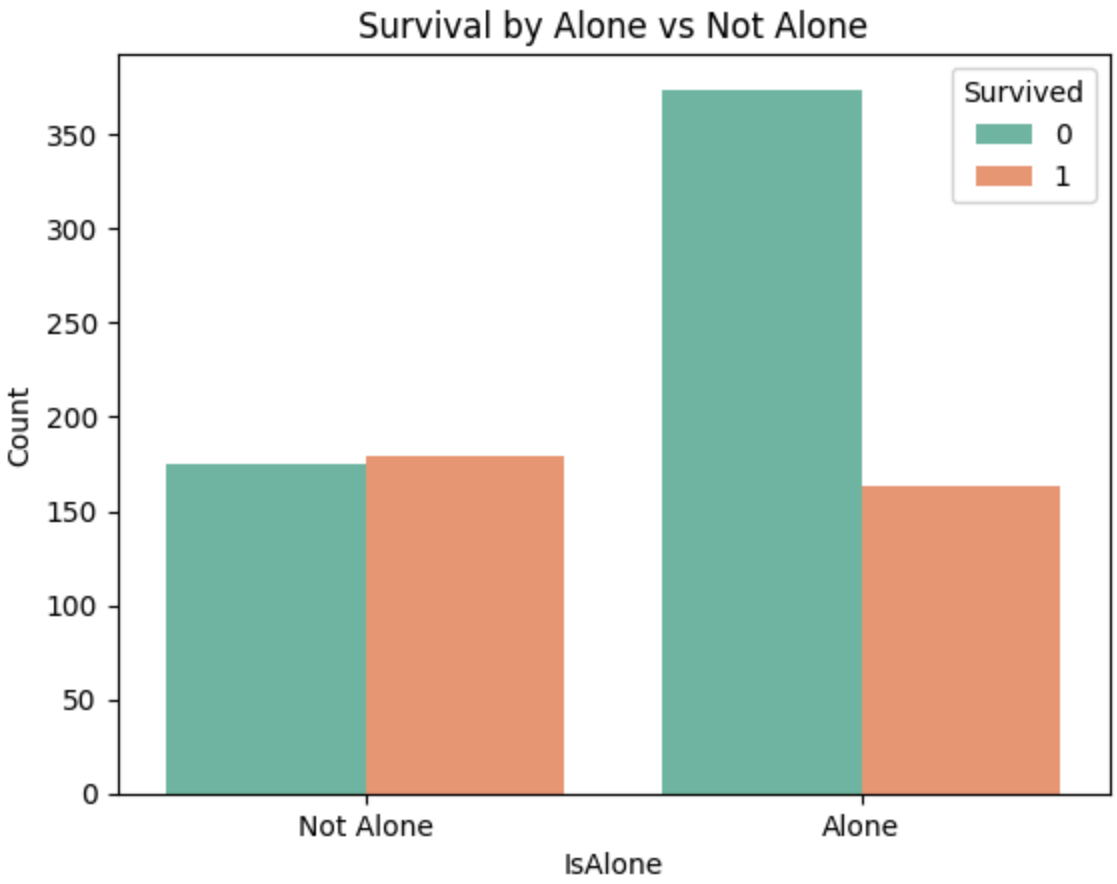


Q9. Did being alone reduce survival chances?

Graph Type: Countplot of IsAlone vs Survival

Interpretation:

> Passengers who were **traveling alone were less likely to survive** than those with family members. Emotional support or group decision-making may have contributed to this outcome.



Q10. Which age group had the highest survival rate?

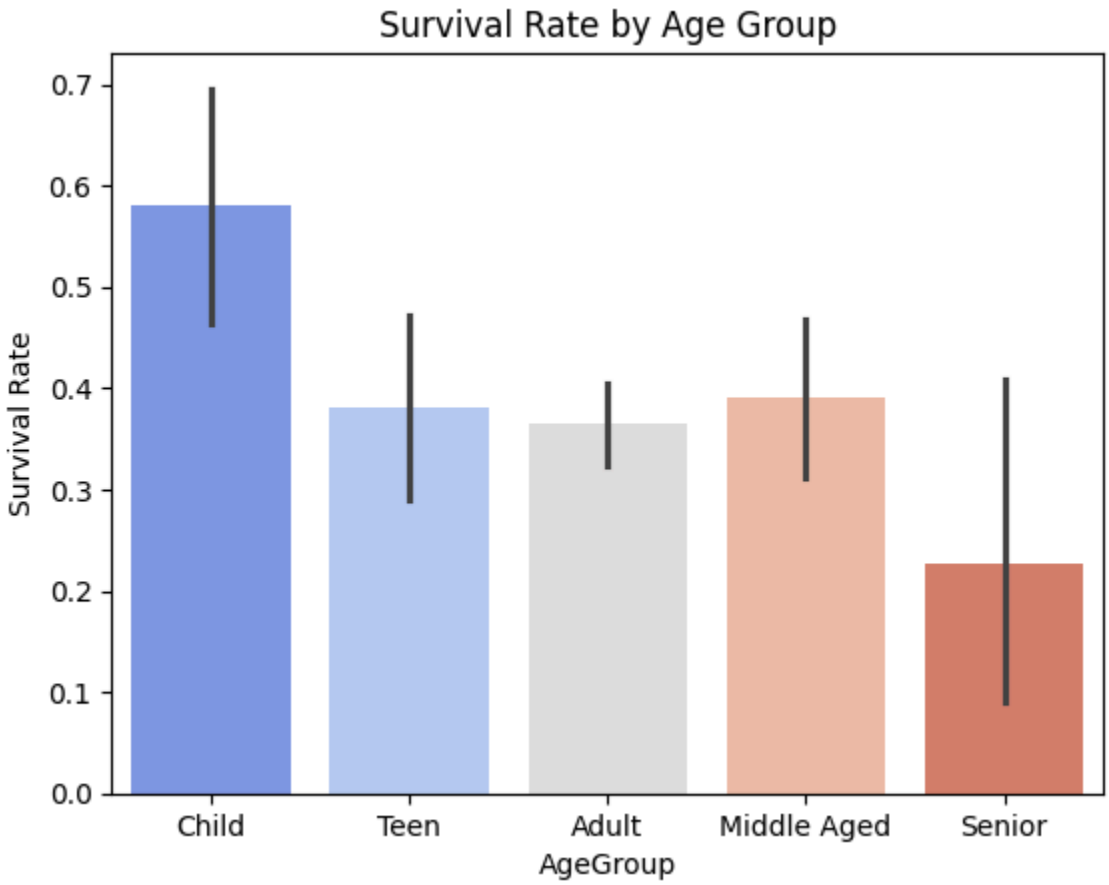
Graph Type: Bar Plot of Age Group vs Survival

Passengers were categorized into:

- **Child (0-12)**
- **Teen (13-20)**
- **Adult (21-40)**
- **Middle-Aged (41-60)**
- **Senior (61+)**

Interpretation:

Children had the highest survival rate, followed by teens. Seniors had the lowest. This emphasizes that **younger passengers were given priority** during evacuation.



Overall Dashboard Observations:

- **Women, children, and first-class passengers** had significantly higher survival rates.
- **Fare paid, passenger class, and family size** were key predictors of survival.
- Visualizations made it easier to understand and communicate data insights clearly and quickly.

🎯 Conclusion from the Dashboard:

Using visual analysis tools, the project was able to derive **10 critical insights** that align with historical knowledge while also revealing data-specific trends. These findings helped shape the predictive model and showed how data visualization can be used effectively in AI/ML-based decision systems.

Chapter 5

Conclusions, Summary and Recommendations

This chapter presents the key conclusions drawn from the project, a summary of all processes and findings, and practical recommendations based on the results. Additionally, it outlines future research scope and offers suggestions for improving the project or extending it further.

5.1 Summary

The objective of this project was to apply machine learning techniques to the Titanic dataset to predict passenger survival based on features such as age, gender, fare, class, and family size. The project followed a complete AI/ML workflow which included:

- **Data Collection**: The dataset was sourced from Kaggle's Titanic challenge.
- **Data Cleaning & Preprocessing**: Handled missing values, encoded categorical data, and created engineered features like `FamilySize` and `IsAlone`.
- **Exploratory Data Analysis (EDA)**: Statistical summaries and visualizations revealed patterns and insights, such as higher survival rates among women, children, and first-class passengers.
- **Dashboard-Based Analysis**: Ten analytical questions were framed and answered using visual tools such as countplots, histograms, and heatmaps.
- **Model Building (Optional Step)**: A basic classification model using Logistic Regression was implemented to demonstrate how AI could be used to predict survival.

Through this step-by-step process, the project successfully demonstrated how machine learning can be applied to historical data to predict outcomes and derive insights.

5.2 Conclusions

Based on the data analysis and dashboard visualizations, the following key conclusions were drawn:

- **Gender had a significant impact** on survival: Female passengers had a much higher survival rate than males.

- **Passenger class was a strong predictor**: First-class passengers were more likely to survive due to better access to safety resources.

- **Age influenced survival**: Children and teenagers had higher survival rates compared to older passengers.

- **Fare amount correlated with survival**: Higher-paying passengers, generally from upper classes, had better chances of survival.

- **Being alone reduced survival probability**: Passengers traveling alone were less likely to survive than those with family members.

- **Port of embarkation showed trends**: Passengers from Cherbourg (C) had a higher survival rate than those from Southampton or Queenstown.

These results align with historical evacuation priorities and highlight how data science can extract meaningful knowledge from past events.

5.3 Recommendations

Based on the findings and technical process followed in this project, the following recommendations can be made:

- **In future projects, explore advanced ML models** such as Random Forest, Gradient Boosting, or Neural Networks for better prediction accuracy.

- **Apply hyperparameter tuning** and cross-validation to improve model robustness.

- **Use more complex feature engineering**, such as extracting titles from passenger names or analyzing family groups.

- **Deploy the model as a web app** using Flask or Streamlit for interactive predictions and dashboard sharing.

- **Automate the EDA** using tools like Sweetviz, Pandas-Profiling, or Tableau for a more professional presentation.

5.4 Scope for Further Research

There are several ways this project can be extended:

- **Compare multiple ML models** with evaluation metrics like ROC-AUC, Precision, Recall, and F1-Score.
- **Incorporate external datasets** such as crew lists, weather conditions, or rescue boat data for deeper insights.
- **Use deep learning models** like Artificial Neural Networks (ANNs) for complex pattern recognition.
- **Explore NLP techniques** on the Name or Ticket fields to extract latent features.
- **Simulate real-time dashboards** using tools like Power BI, Dash, or Tableau.

5.5 Suggestions

- Beginners in AI/ML should use Titanic as a practice project due to its clean structure and balanced classification problem.
- Educators can use this dataset to demonstrate the full machine learning lifecycle.
- For students, adding model deployment (e.g., via Heroku or GitHub Pages) can significantly enhance project value and resume strength.

Chapter 6

Sample Code & Links

This chapter includes the essential Python code snippets used in the implementation of the Titanic Survival Prediction project. These code blocks illustrate how data was processed, analyzed, and visualized using popular Python libraries such as Pandas, NumPy, Seaborn, and Scikit-learn. Each snippet has been selected to represent a key phase in the machine learning workflow — from data loading to preprocessing, visualization, and modeling.

6.5 External Links and Resources

- **Dataset Source (Kaggle):**
  https://www.kaggle.com/competitions/titanic

- **Google Colab Notebook:**
  https://colab.research.google.com/drive/1fyEI9WNGQ6rkiJOH1i4FINZ0MM3BhGqW#scrollTo=ZUoQIwED5sKe

- **Python Documentation:**
  https://docs.python.org/3/

- **Pandas Documentation:**
  https://pandas.pydata.org/

- **Seaborn Documentation:**
  https://seaborn.pydata.org/

- **Scikit-learn Documentation:**
  https://scikit-learn.org/

Summary

This chapter showcased the essential code elements that powered the Titanic Survival Prediction project. It highlighted how Python libraries were used for data preprocessing, analysis, and visualization, and provided references to external resources that supported the development process.