Exploratory Data Analysis (EDA) Report

1. Overview

The dataset was analyzed to understand its structure, detect missing values, outliers, and discover relationships between variables and the target feature (**Survived**).
The main goal was to gain insights that can help build a strong prediction model.

2. Data Cleaning and Preprocessing

Before analysis, the dataset was cleaned to make it suitable for modeling:

Missing values were detected and handled properly.

Unnecessary columns (like IDs or irrelevant information) were removed.

Categorical features were converted into numerical form using encoding techniques.

Outliers were detected using boxplots and handled by replacement to reduce their negative effect on model performance.

Skewed features such as “Fare” were normalized using replacing techniques to make the data more balanced.

3. Summary Statistics

The dataset contained both **numerical** and **categorical** variables.

The mean, median, and standard deviation were calculated for each numerical column to check data distribution.

A heatmap was used to visualize the **correlation** between variables.
It showed that some features were strongly correlated with the target, while others were not significant.

4. Correlation Findings

The **correlation heatmap** showed that features like **Sex, Pclass, Fare, and Age** were the most relevant to the target variable.

Some features were weakly correlated and could be dropped to simplify the model without losing accuracy.

5. Key Insights

After exploring the data, several important insights were found:

Gender and Survival: Females had a much higher survival rate than males.

→ Gender is a strong predictor for survival.

Passenger Class (Pclass): Passengers in higher classes (1st class) were more likely to survive.

→ Social status or ticket price had an effect on survival chances.

Age: Younger passengers had better chances of survival than older ones.

→ Age has a negative relationship with survival probability.

Fare: Higher ticket fares were related to a higher survival rate

→ Wealthier passengers had more access to safety.

Embarked Port:
The port of embarkation showed small differences in survival rate but was less significant compared to other features.

Modeling and Results Report

1. Objective

After cleaning and exploring the data, the goal was to build machine learning models that can predict whether a passenger survived or not on the Titanic.
Different models were trained and evaluated to find the one with the best accuracy and performance.

2. Data Preparation

Before training:

The **target variable** was defined as Survived.

All **features** were selected and prepared for the models.

Categorical columns were converted into numerical form using **OneHotEncoder**.

The dataset was then **split** into training and testing sets to evaluate model performance.

3. Models Used

The following classification models were tested:

Logistic Regression

→ Used as a baseline model to understand data separability.

Support Vector Machine (SVM)

→ Used to find the optimal hyperplane that separates survivors and non-survivors.

4. Results Summary

Model 1: Logistic Regression	Model 2: SVM
Accuracy: 80%	**Accuracy:** 84%
Precision (0): 0.82	**Precision (0):** 0.88
Recall (1): 0.73	**Recall (1):** 0.83
F1-Score: 0.80	**F1-Score:** 0.84