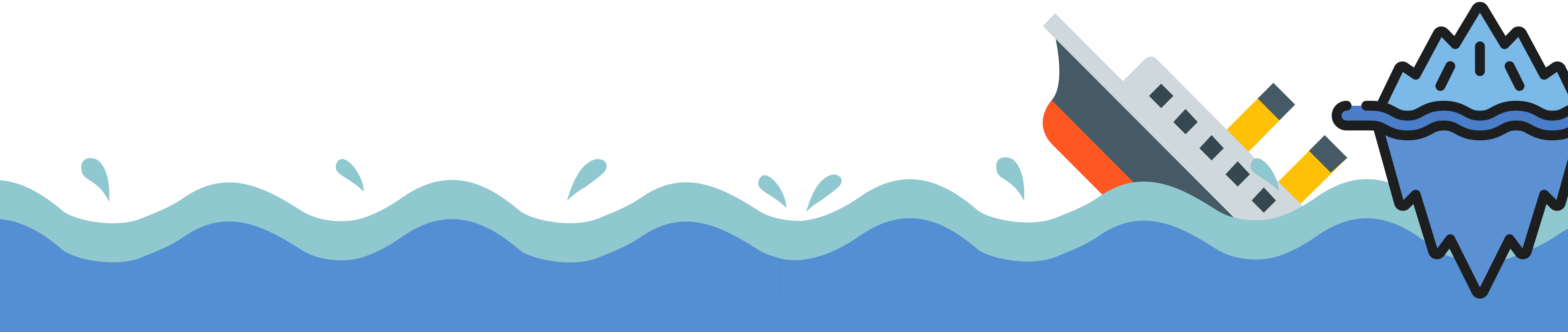
ibimbing

Data Science
Titanic Project

Using python™ kaggle



ibimbing

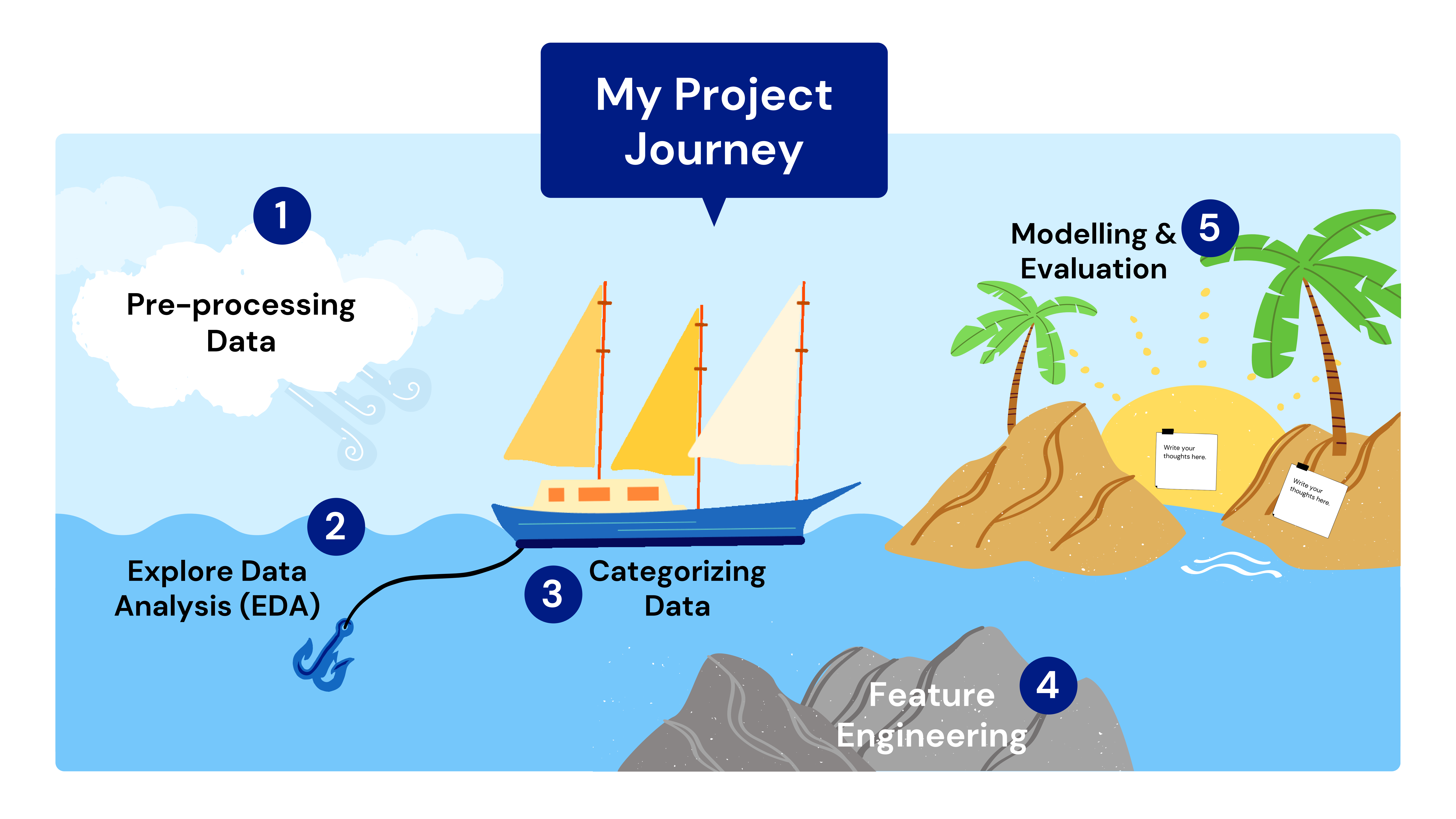
Introduction

As a **fresh graduate** in **Informatics** with a strong passion for data, I am particularly drawn to the fields of data analysis and data science. I have **hands-on experience** in the **end-to-end process** of data handling, from **collecting data** using web scraping techniques to creating **insightful visualizations** that highlight key patterns and trends. My portfolio showcases my ability to **transform raw data** into **actionable insights**, reflecting my dedication to leveraging data for informed **decision-making**.

Putri Nurrahmah Wear

About Titanic's Dataset

The Titanic dataset was created to **provide a deeper understanding** of the **factors that influenced passenger survival** during the **RMS Titanic's sinking** in 1912. By analyzing data such as passengers' names, ages, genders, and socio-economic classes, this dataset allows us to **build predictive models** that answer the question: **"What types of people were more likely to survive?"** Analyzing patterns within the data helps us better understand the dynamics of safety and evacuation decisions during disasters.


My Project Journey
1
Pre-processing Data
2
Explore Data Analysis (EDA)
3
Categorizing Data
4
Feature Engineering
5
Modelling & Evaluation
Write your thoughts here.
Write your thoughts here.

Data Overview

The dataset contains 891 rows and 12 columns.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Pre-Processing Data

Before

```
#show if there are any missing values
titanic.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

After

```
#Check missing values
titanic.isnull().sum()
```

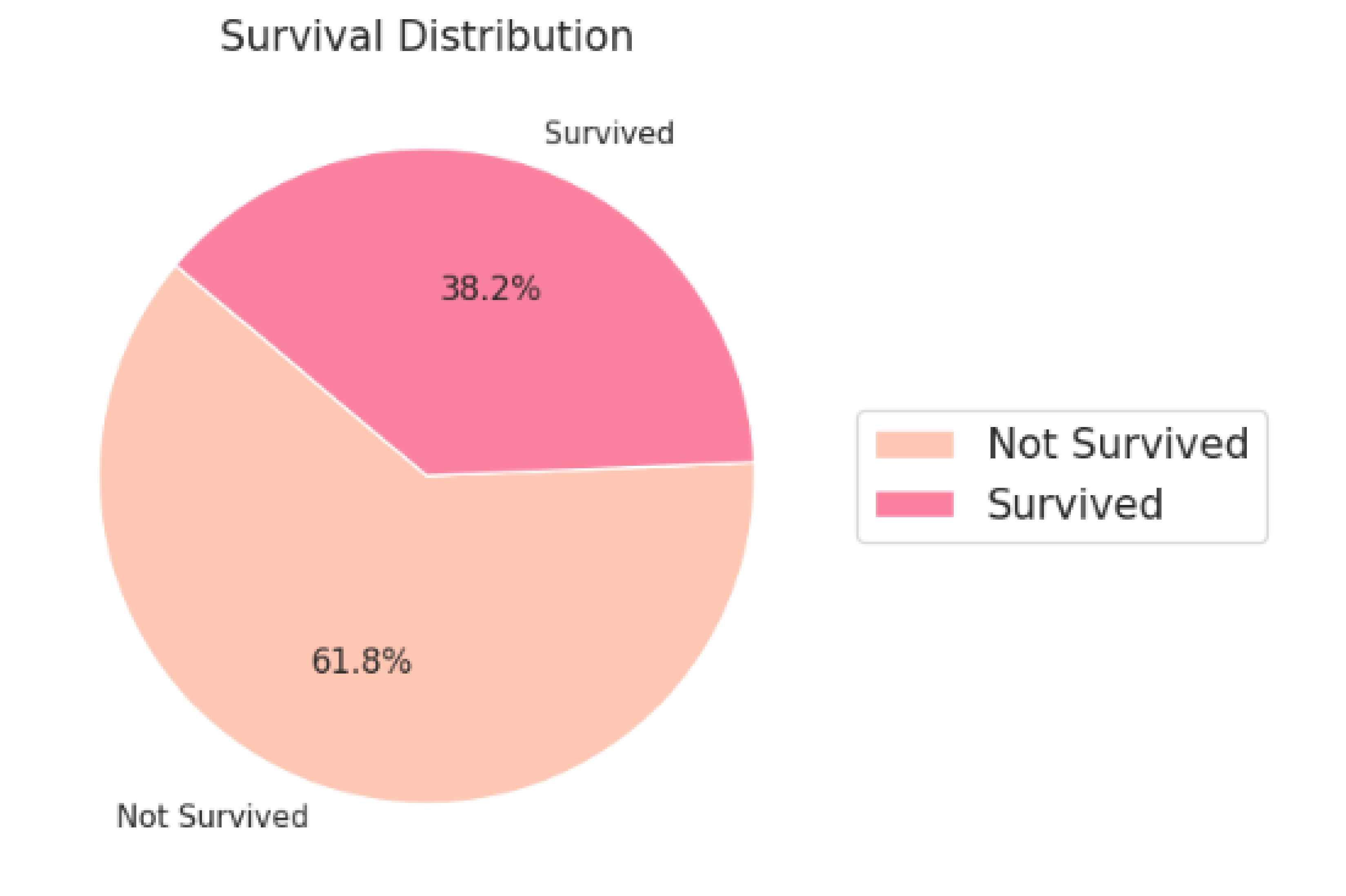
```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Embarked    0
dtype: int64
```

1. The dataset **has missing values** in the **Age, Cabin, and Embarked** columns.
2. I **dropped** the **'PassengerID' & 'Ticket'** columns because they're just a unique codes for each passengger, and **removed the 'Cabin'** column due to **many missing values**. I also **dropped** the rows with **missing values in 'Embarked'** since there were only two.

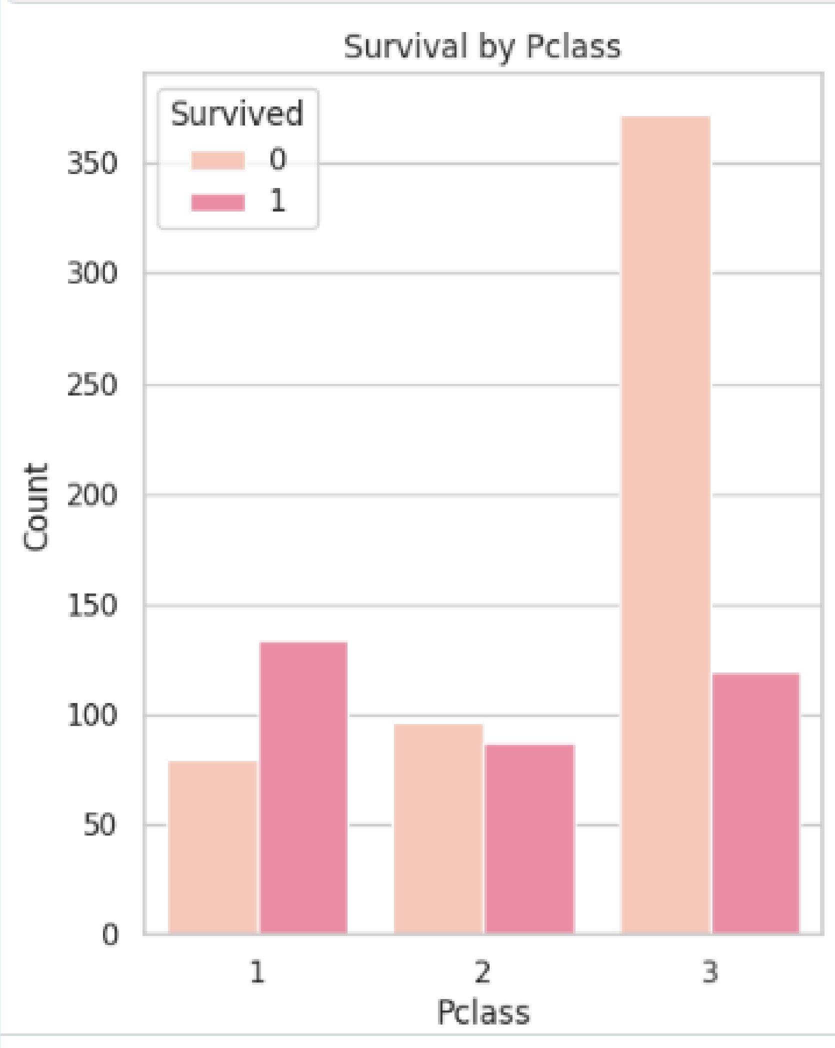
Explore Data Analysis (EDA)

Passenger Survival Rate

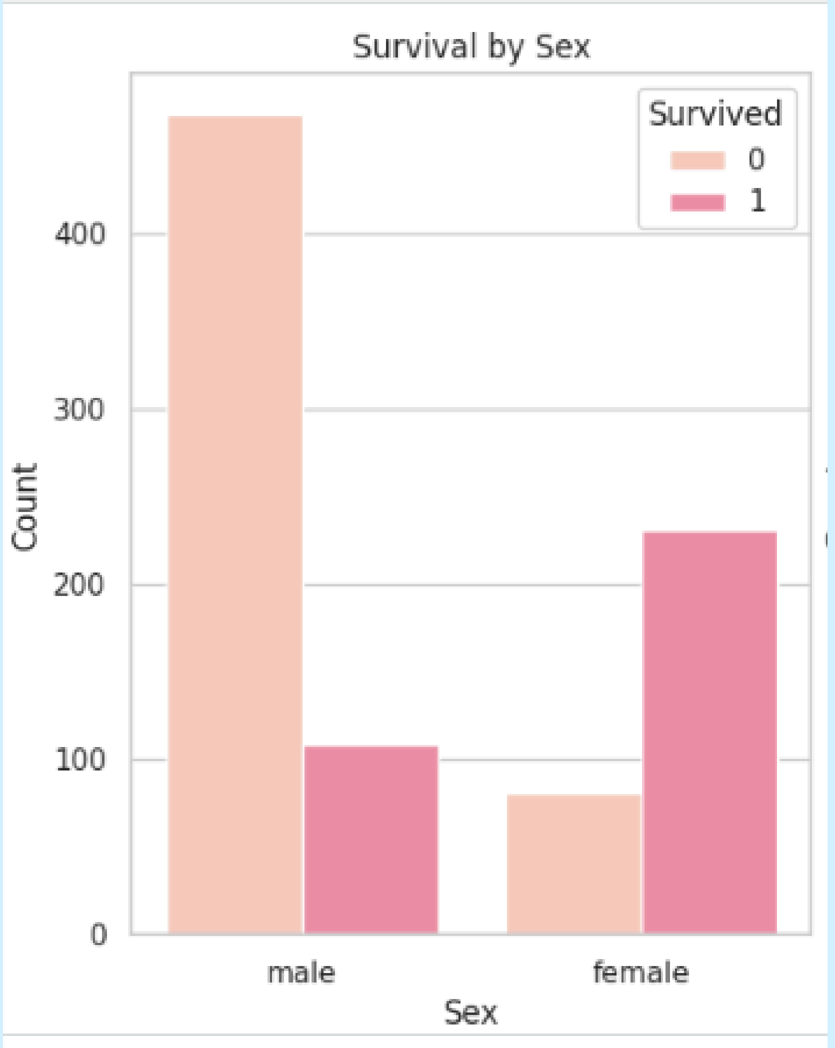
The data indicates that the number of passengers who **survived** is **smaller than** those **who did not**, with 38.2% of passengers surviving and 61.8% not surviving.



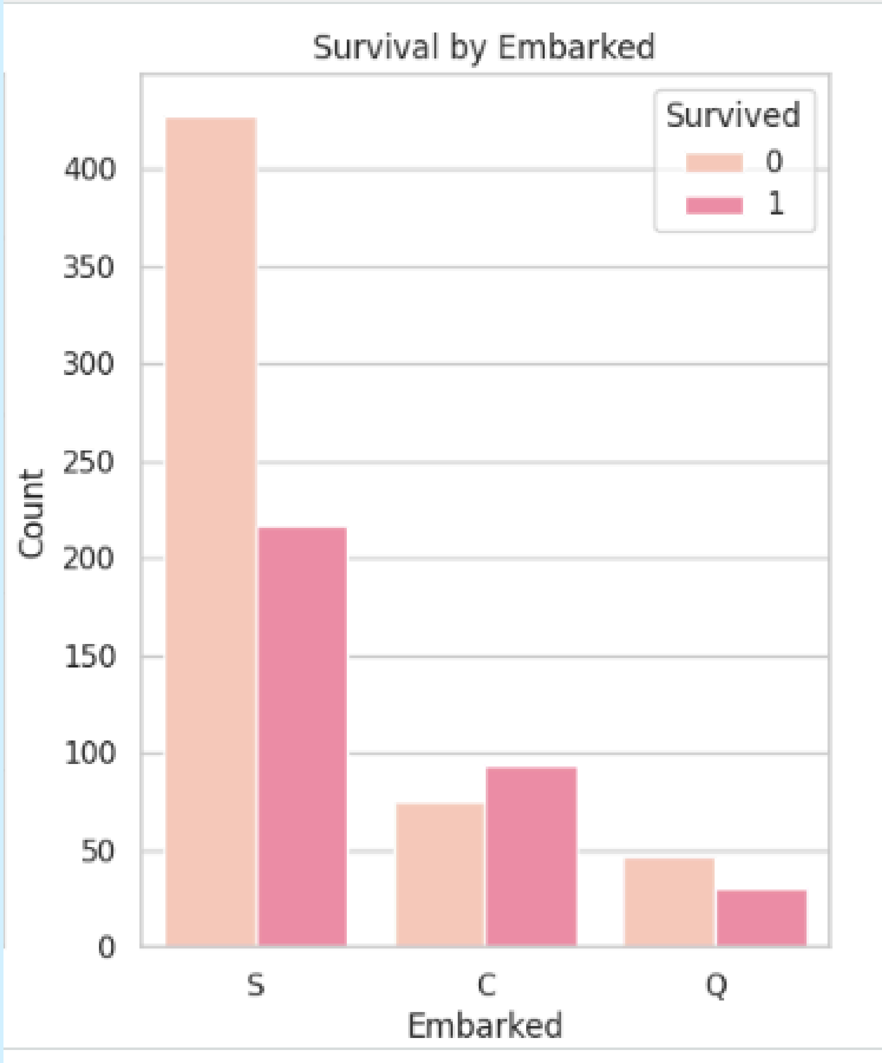
Survival Rate by Pclass, Sex, & Embarked



Higher-class passengers had a significantly **higher survival rate**.

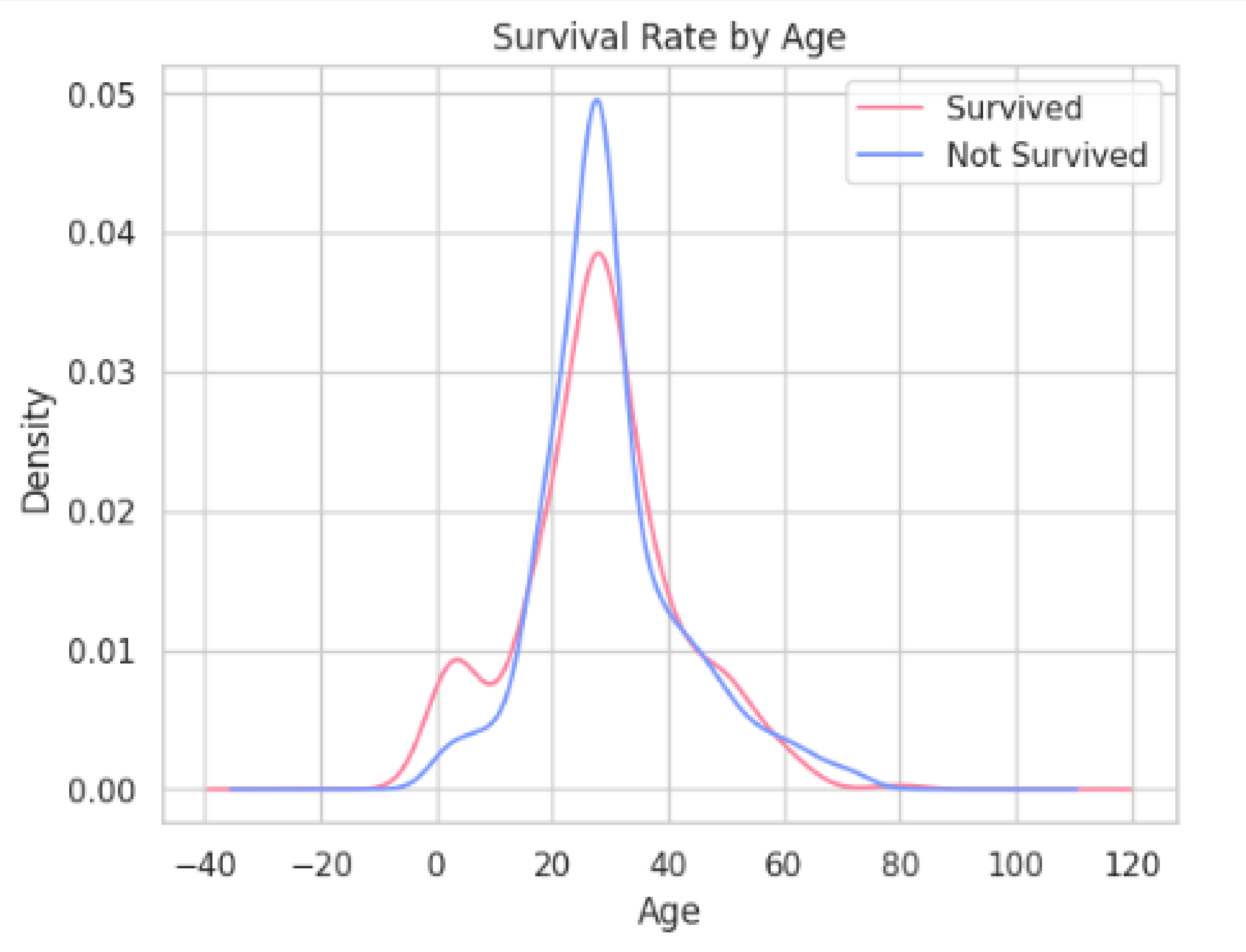


There were more male passengers than female, but the **survival rate** for **female passengers was higher.**

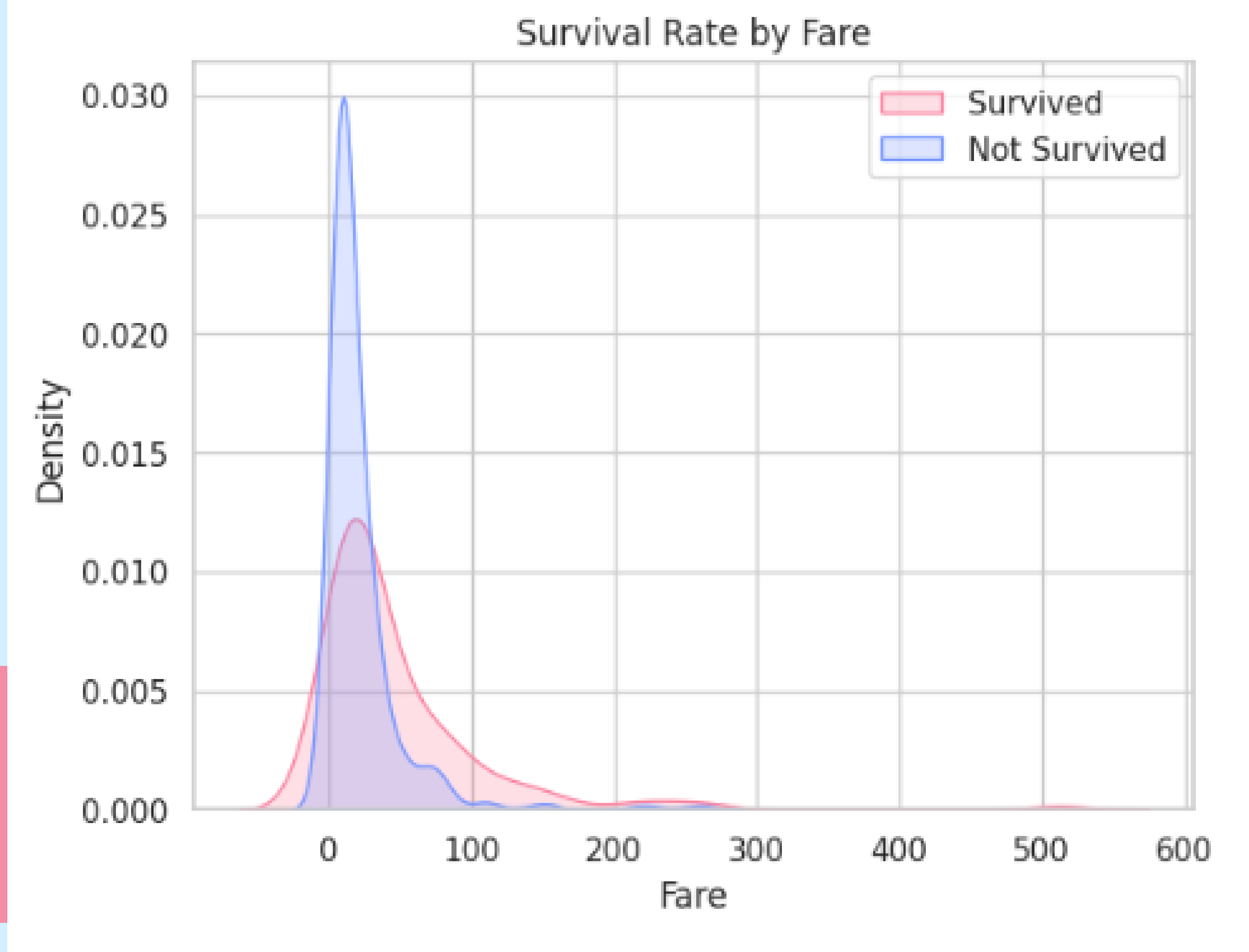


Passengers from **Southampton had the best chance of survival** compared to those from Cherbourg & Queenstown.

Survival Rate by Age & Fare



- The density plot shows that children aged 0-5 had a higher survival rate. This suggests that **children** were **prioritized for survival**.



- Passengers with **higher fares** also had a **higher survival rate.**

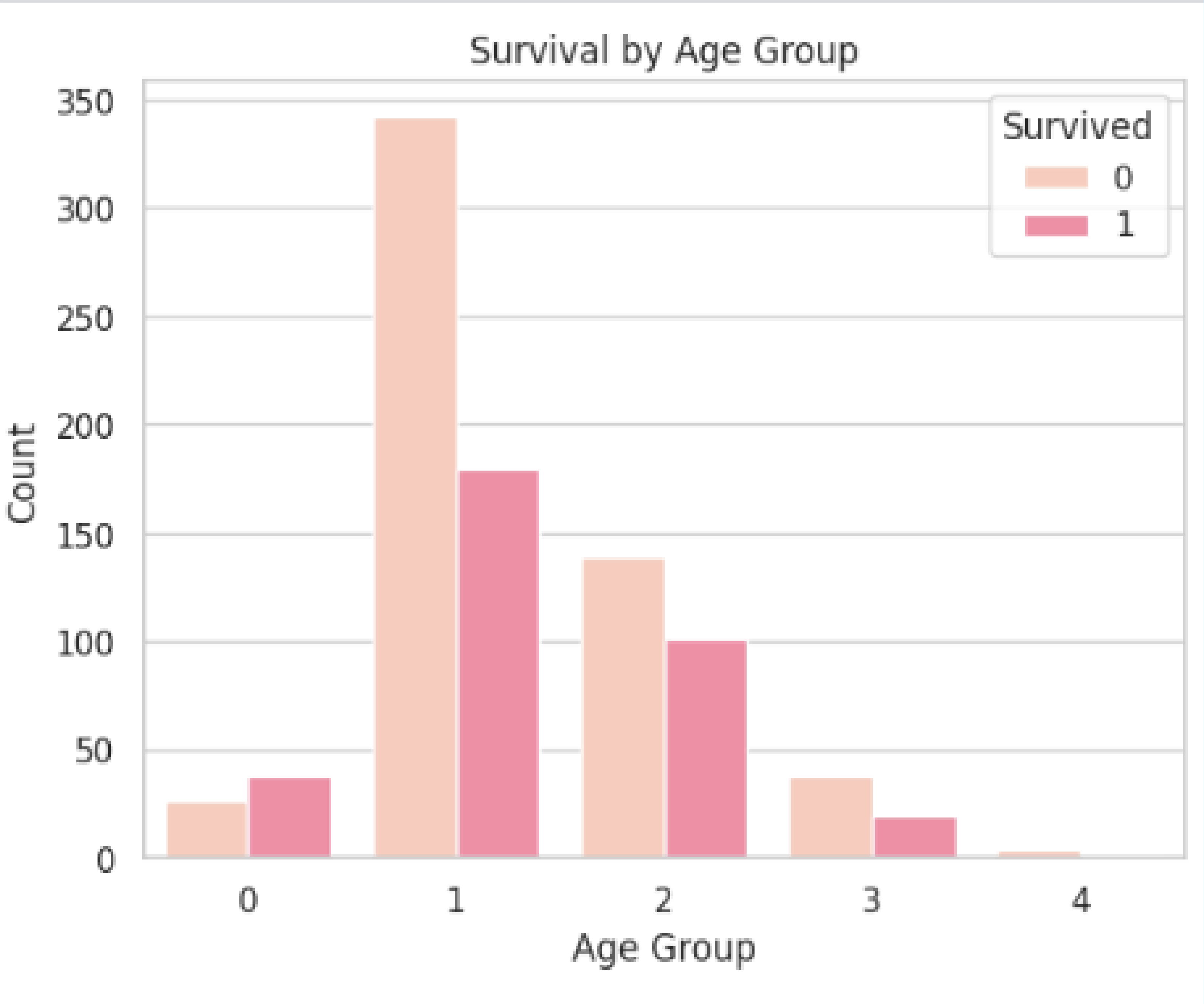
Categorizing Data

Survival by Age Group

Based on the previous insights, 'Age' is a significant feature affecting passenger survival rates. I **re-categorized** it, as fewer categories generally improve machine learning performance. It is clear from the data that the **age group 10 years & under** has a **higher survival rate**.

Age Group

- 0 : <= 10 years
- 1 : <= 30 years
- 2 : <= 50 years
- 3 : <= 70 years
- 4 : else

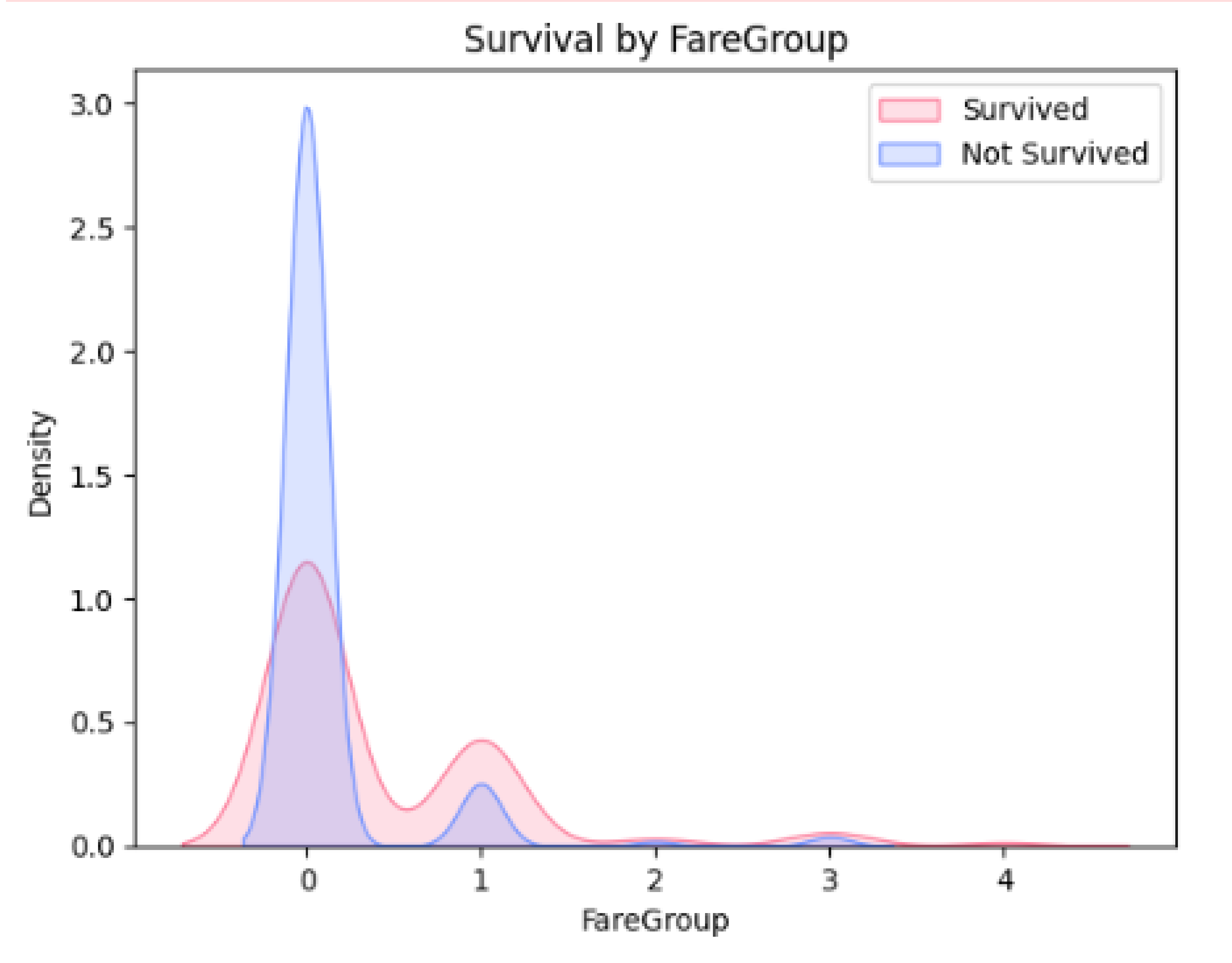


Survival by Fare Group

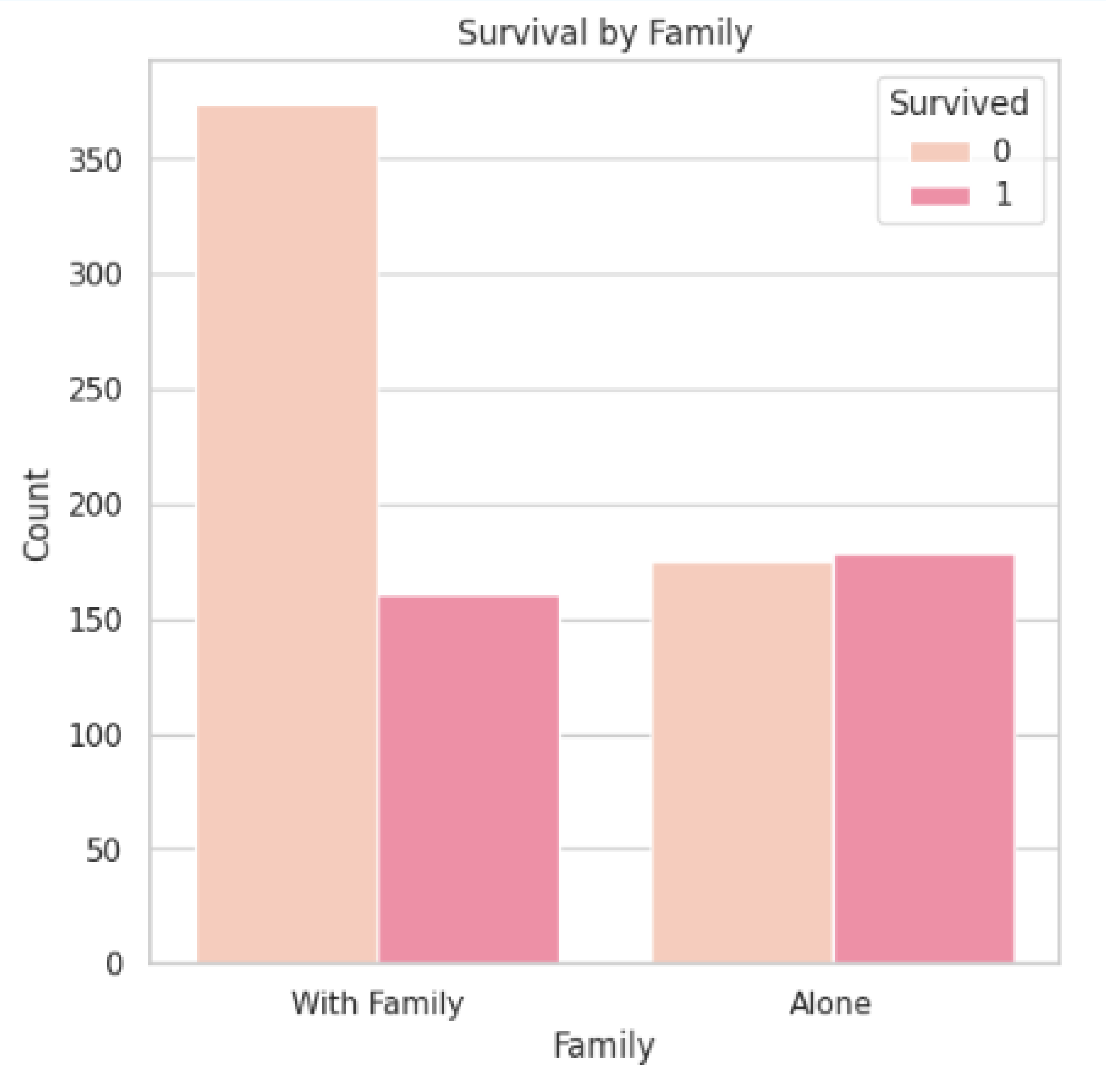
'Fare' is another feature that influences passenger survival chances. The data shows that **higher fare categories** are associated with **higher survival rates**.

Fare Group

- 0 : <= 50
- 1 : <= 150
- 2 : <= 200
- 3 : <= 300
- 4 : else



Survival by Family



- In this dataset, family members are categorized into **'SibSp' and 'Parch'**. I merged these **into** a single **'Family'** category.

- Passengers **without family** members had a **higher survival rate compared to** those **with family** members.

Feature Engineering

ibimbing

Label Encoding

- **Label Encoder** is a function used to **convert categorical data** into **numeric data**. The data that needs to be converted includes **'Sex' and 'Embarked'**.

Sex

0 : Female
1 : Male

Embarked

0 : C (Cherbourg)
1 : Q (Queenstown)
2 : S (Southampton)

Before

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Family	FareGroup	AgeGroup
0	0	3	male	22.0	1	0	7.2500	S	1	0	1
1	1	1	female	38.0	1	0	71.2833	C	1	1	2
2	1	3	female	26.0	0	0	7.9250	S	0	0	1
3	1	1	female	35.0	1	0	53.1000	S	1	1	2
4	0	3	male	35.0	0	0	8.0500	S	0	0	2

After

	Survived	Pclas	Sex	ge	SibSp	Parch	Fare	Embarked	Family	FareGroup	AgeGroup
0	0		1	2.0	1	0	7.2500	2	1	0	1
1	1		0	8.0	1	0	71.2833	0	1	1	2
2	1		0	6.0	0	0	7.9250	2	0	0	1
3	1		0	5.0	1	0	53.1000	2	1	1	2
4	0		1	5.0	0	0	8.0500	2	0	0	2

Feature Selection

- The selected features **(X)** for modeling **include** the columns **'Pclass', 'Sex', 'Family', 'AgeGroup', 'FareGroup', and 'Embarked'**.
- The target **(Y)** is the **'Survived'** column.

```
X
```

	Pclass	Sex	Family	AgeGroup	FareGroup	Embarked
0	3	1	1	1	0	2
1	1	0	1	2	1	0
2	3	0	0	1	0	2
3	1	0	1	2	1	2
4	3	1	0	2	0	2
...	...	...	...	...	...	...
886	2	1	0	1	0	2
887	1	0	0	1	0	2
888	3	0	1	1	0	2
889	1	1	0	1	0	0
890	3	1	0	2	0	1

889 rows × 6 columns

```
y
```

```
0      0
1      1
2      1
3      1
4      0
      ..
886    0
887    1
888    0
889    1
890    0
Name: Survived, Length: 889, dtype: int64
```

ibimbing

Splitting Data

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

- The data was split into **80% for training** (711 data) and **20% for testing** (178 data).

```
X_train
```

	Pclass	Sex	Family	AgeGroup	FareGroup	Embarked
708	1	0	0	1	2	2
240	3	0	1	1	0	0
382	3	1	0	2	0	2
792	3	0	1	1	1	2
683	3	1	1	1	0	2
...	...	...	...	...	...	...
107	3	1	0	1	0	2
271	3	1	0	1	0	2
862	1	0	0	2	0	2
436	3	0	1	1	0	2
103	3	1	0	2	0	2

711 rows × 6 columns

Train Data

```
X_test
```

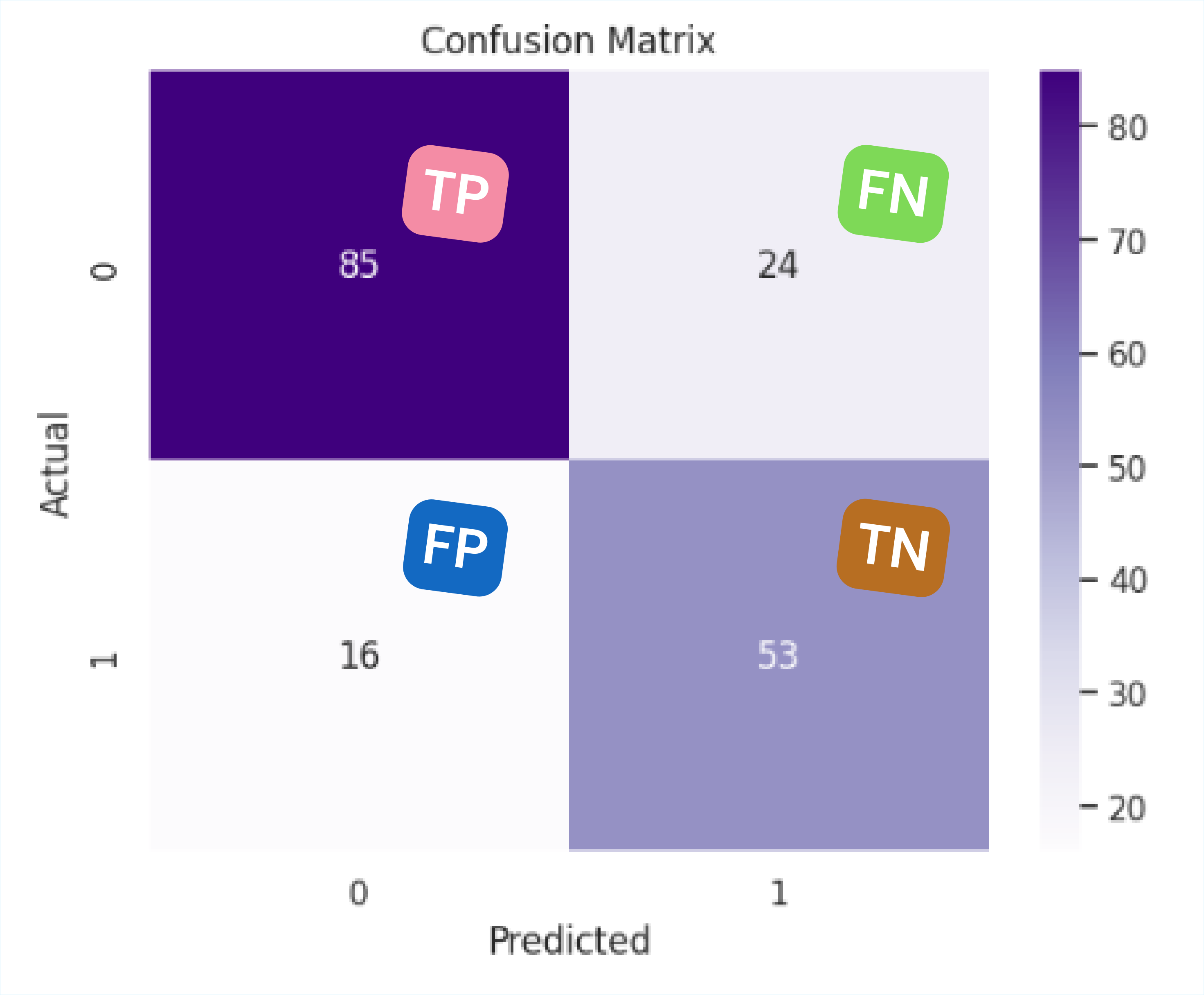
	Pclass	Sex	Family	AgeGroup	FareGroup	Embarked
281	3	1	0	1	0	2
435	1	0	1	1	1	2
39	3	0	1	1	0	0
418	2	1	0	1	0	2
585	1	0	1	1	1	2
...	...	...	...	...	...	...
433	3	1	0	1	0	2
807	3	0	0	1	0	2
25	3	0	1	2	0	2
85	3	0	1	2	0	2
10	3	0	1	0	0	2

178 rows × 6 columns

Test Data

Modelling & Evaluation

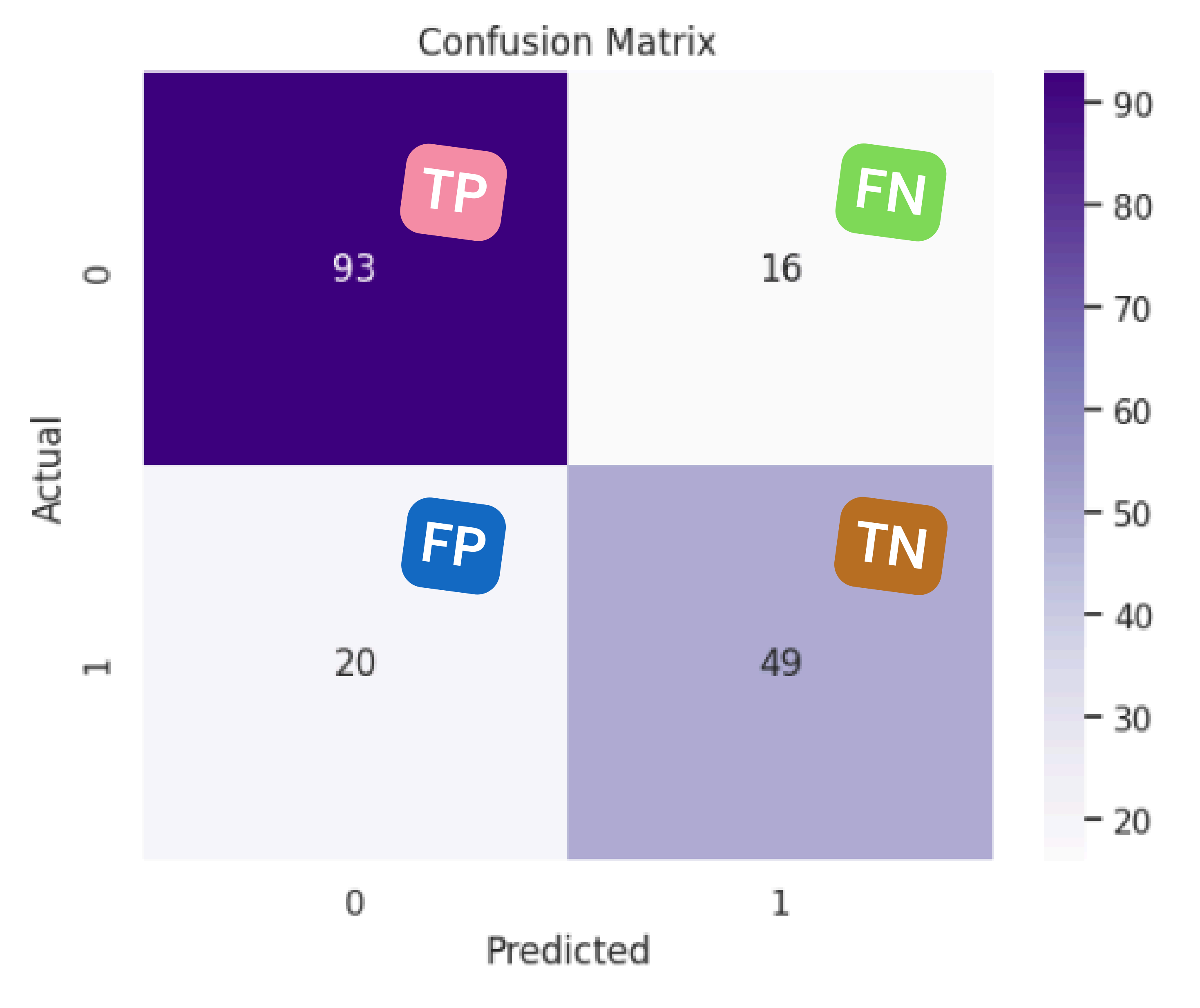
Logistic Regression



The Logistic Regression model achieved an **accuracy of 78%**, with the following results:

- **True Positives (TP):** The model **correctly predicted 85 positive cases**.
- **False Positives (FP):** The model **incorrectly predicted 16 positive cases** as negative.
- **False Negatives (FN):** The model **incorrectly predicted 24 negative cases** as positive.
- **True Negatives (TN):** The model **correctly predicted 53 negative cases**.

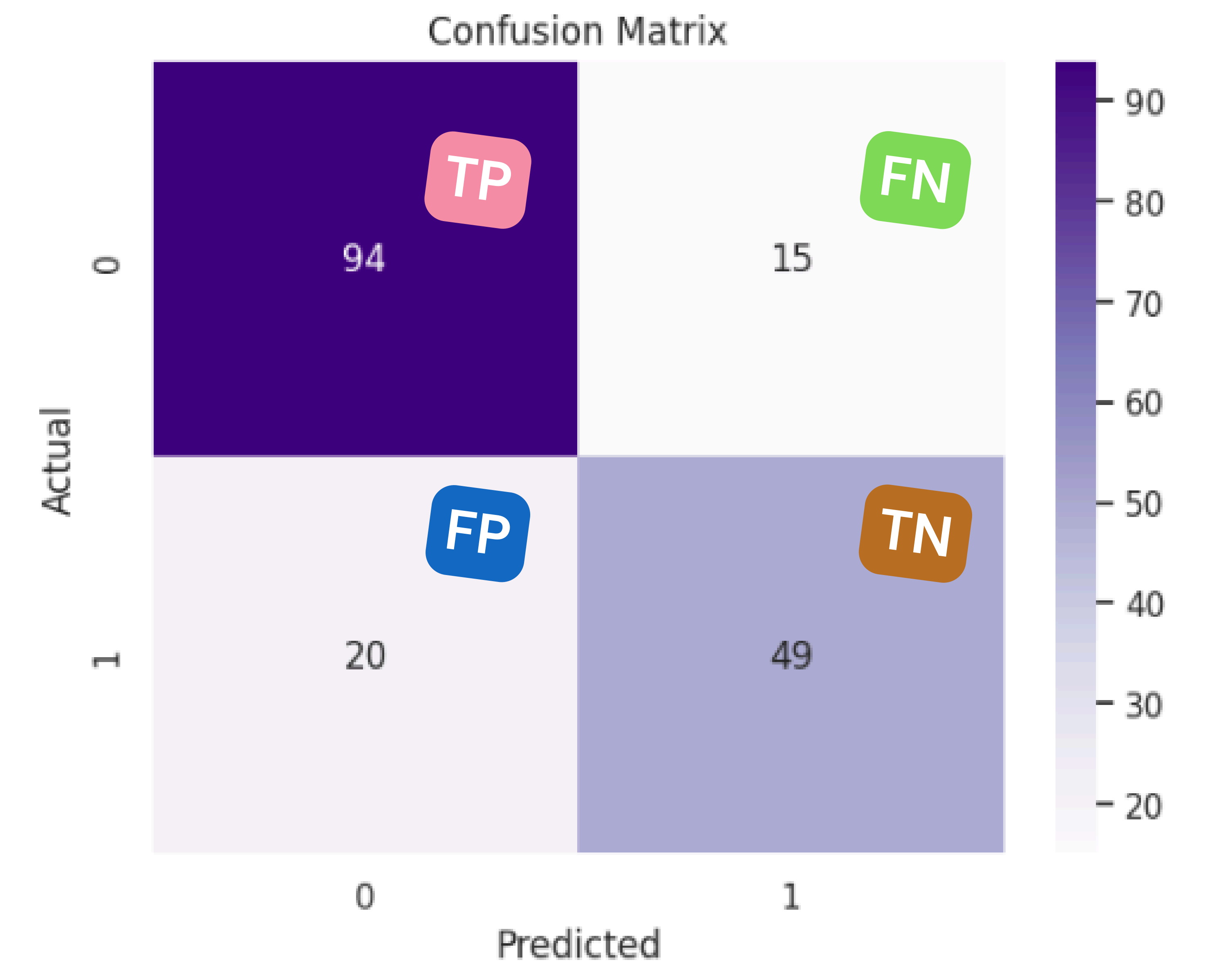
Random Forest



The Logistic Regression model achieved an **accuracy of 80%**, with the following results:

- **True Positives (TP):** The model **correctly predicted 93 positive cases**.
- **False Positives (FP):** The model **incorrectly predicted 20 positive cases** as negative.
- **False Negatives (FN):** The model **incorrectly predicted 16 negative cases** as positive.
- **True Negatives (TN):** The model **correctly predicted 49 negative cases**.

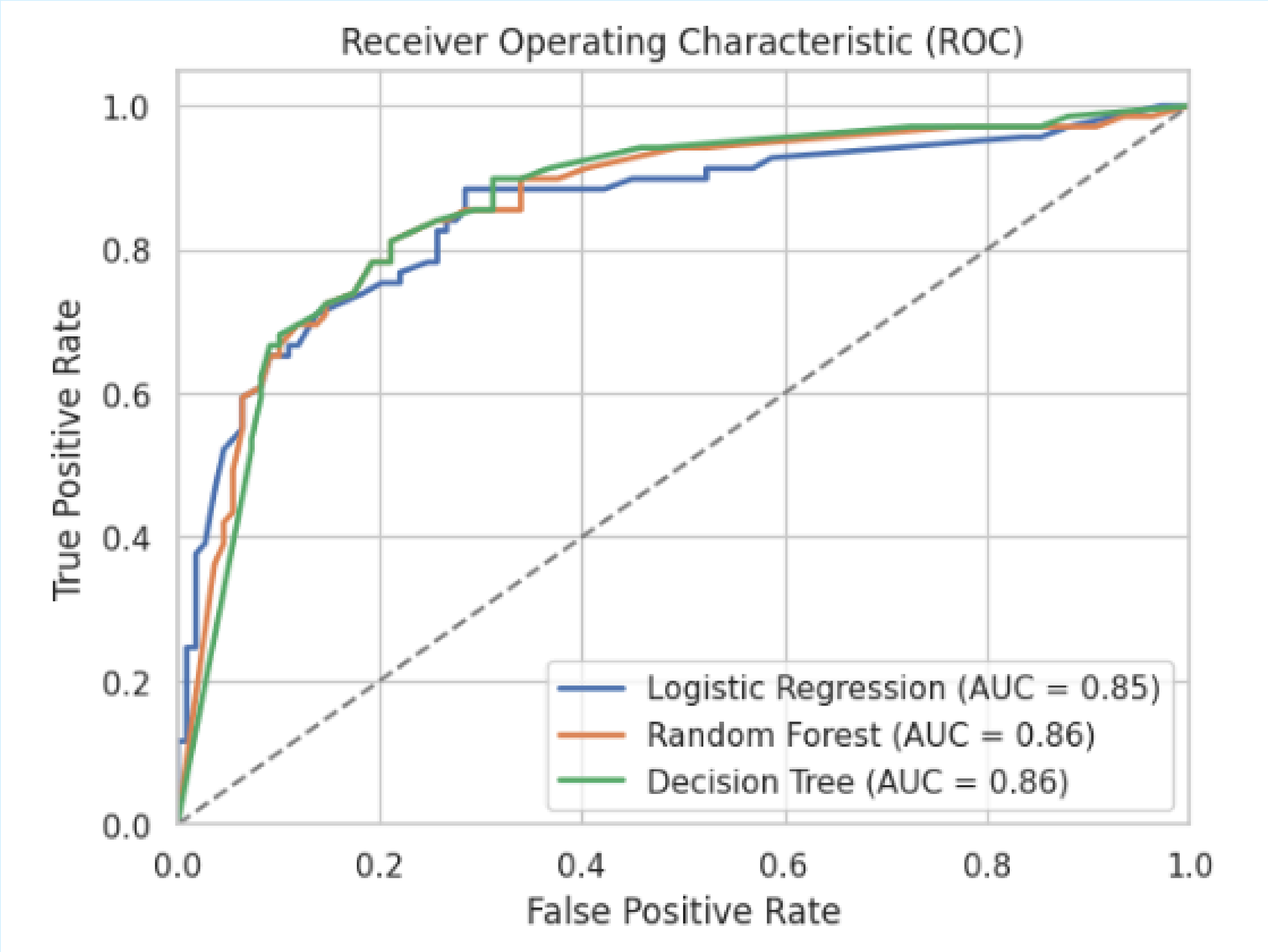
Decision Tree



The Logistic Regression model achieved an **accuracy of 80%**, with the following results:

- **True Positives (TP)**: The model **correctly predicted 94 positive cases.**
- **False Positives (FP)**: The model **incorrectly predicted 20 positive cases** as negative.
- **False Negatives (FN):** The model **incorrectly predicted 15 negative cases** as positive.
- **True Negatives (TN):** The model **correctly predicted 49 negative cases**.

ROC Curve



- Random Forest and Decision Tree exhibit similar performance and slightly better AUC compared to Logistic Regression. **A higher AUC indicates that Random Forest and Decision Tree** are **more effective** at distinguishing between classes **with lower error rates** than **Logistic Regression**.

Conclusion

1. Among the three models, **Random Forest performs best** with an accuracy of 80% and a **strong F1-score for both classes**. It **also** offers an **optimal balance between precision and recall**, outperforming the Decision Tree, which also has an accuracy of 80%.

2. Although Random Forest and Decision Tree have the same AUC, **the choice may depend on specific needs**, such as interpretability (Decision Tree is easier to understand) or other performance metrics.

Thank you!

Have a great day ahead.

putriwear28@gmail.com

linkedin.com/in/putri-nurrahmah-wear

github.com/putriwear