Digital Skill Fair 31.0

Titanic Passanger Survival Analysis

Linking Safety SOPs and Predicting Outcomes with Supervised Learning Models

Presented by **Nabila Putri Dian Luthfiyyah**

Introduction ✨

I'm Nabila Putri Dian Luthfiyyah, a fresh graduate of Mathematics major from Sepuluh Nopember Institute of Technology. Over the past four years, I have developed a strong passion for data-related fields, particularly in mathematical modeling, statistics, and data analysis. I have gained valuable experience in managing and analyzing data through coursework, internships, and my final project. I would love to deepen my skills in these areas to provide insightful solutions and make data-driven decisions that can impact organizations positively.

Problem ⁉️

The tragedy of the Titanic sinking has triggered many studies to understand the factors that influence passenger safety. One important approach is to analyse passenger data to find patterns relating to the implementation of shipboard rescue Standard Operating Procedures (SOPs). Factors such as age, gender, and ticket class may be related to rescue priorities and their role in SOPs. In this project, various classification models from supervised learning are used to predict passenger safety based on patterns from historical data. Through this analysis, it is expected to **find the relationship between passenger data and the implementation of rescue SOPs in determining the possibility of passenger safety, and develop an accurate predictive model.**

1. What was the key factor that contributed to the safety of Titanic passengers?
2. How do the results of the data analysis relate to the safety SOPs implemented on the Titanic?
3. Which supervised learning model provides the best accuracy in predicting the survival of Titanic passengers?

Data Dictionary 

Variable	Definition	Key
Survived	Survival	0 = No, 1= Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3= 3rd
Sex	Gender	Male, Female
Age	Age in years	
SibSp	# of siblings/spouses abroad in Titanic	
Parch	# of parents/childern abroad in Titanic	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Dataset 📒

	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

Tools 🛠️

➡️ Main Tools

Package Tools ⬅️

Data Cleaning ⏳

Before

```
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
```

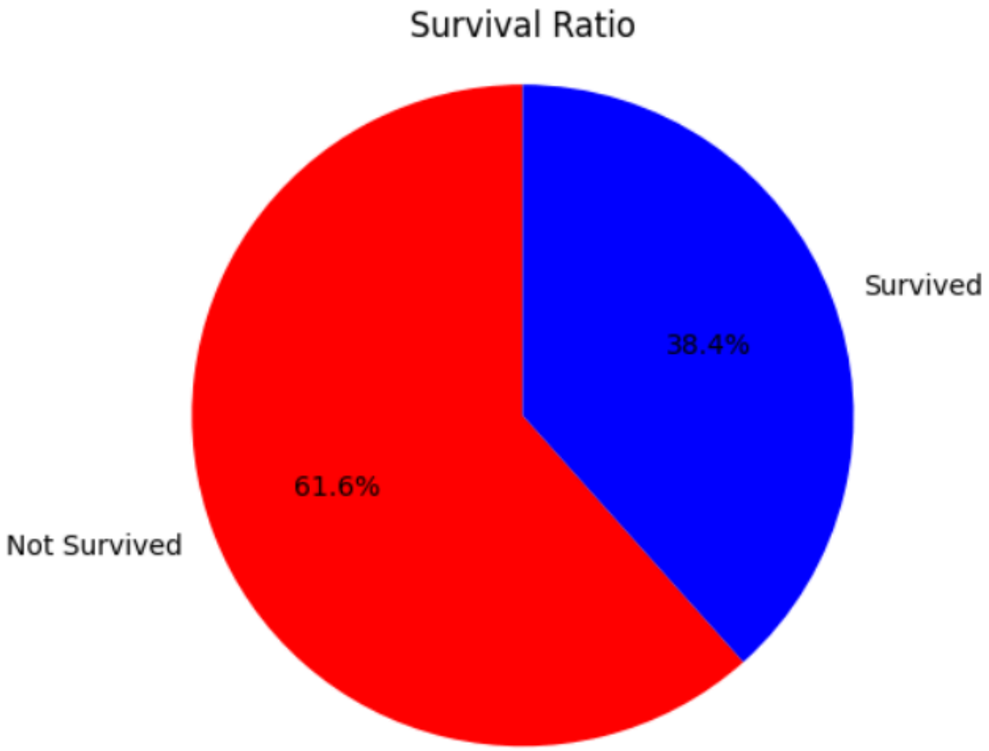
- There are **177 missing values** in the 'Age' feature.
- There are **687 missing values** in the 'Cabin' feature.
- There are **2 missing values** in the 'Embarked' feature.

After

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

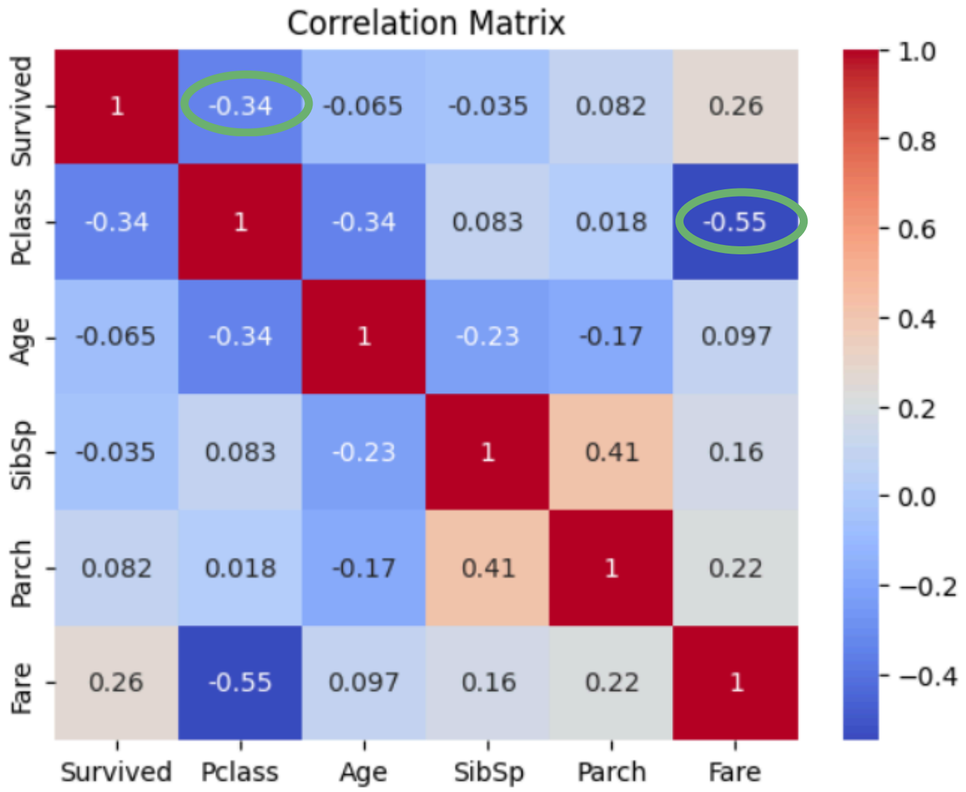
- The missing values in the **'Age' and 'Embarked'** feature have been resolved by inputting the **median and mode.**
- The '**Passenger Id**' and '**Ticket**' features are **dropped** because they are just **unique codes**.
- The '**Embarked**' feature was dropped because it has **too many missing values**.

Survival Rate of Passengers 🔍



💡 The number of passengers who **did not survive** was **greater** than those who survived, with a **ratio** of **61.6 : 38.4**.

Correlation Matrix 🔍



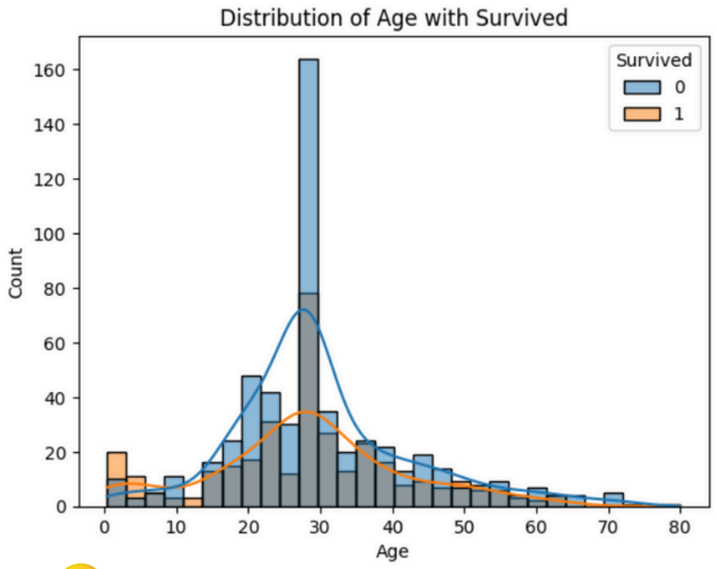
The **'Pclass'** feature has **the greatest effect on 'Survived'** with a **negative correlation** of 0.34.

➡️ The higher Pclass, the greater the chance of survival.

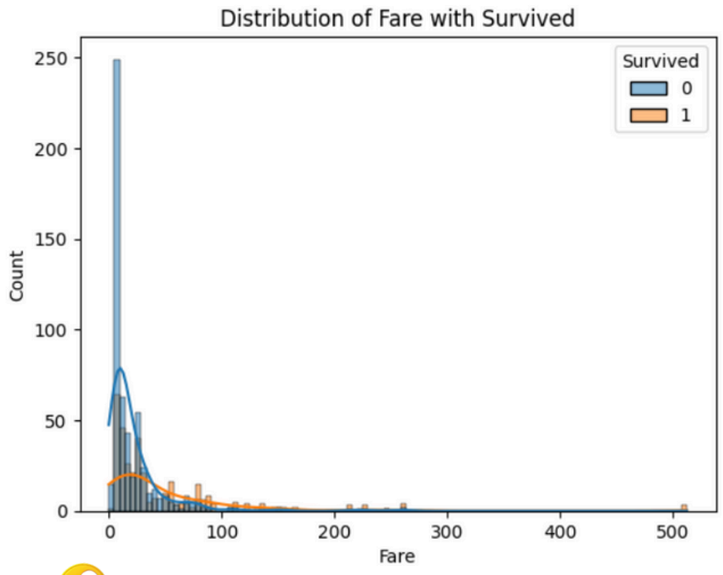
The **'Fare'** feature has **the greatest effect on 'PClass'** with a **negative correlation** of 0.55.

➡️ The higher Pclass, the higher the Fare.

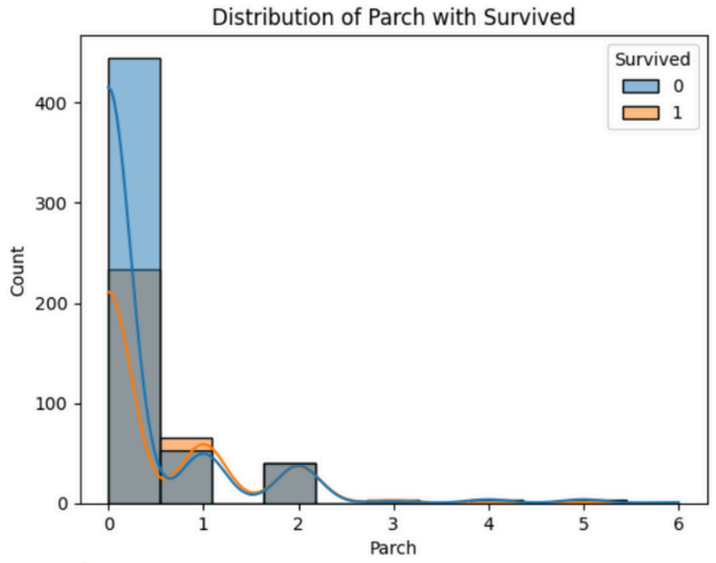
Survival Rate by Age, Fare, and Parch 🔍



🔑 **Children under 10 are more probable to survive**, possibly because they are prioritised for rescue.

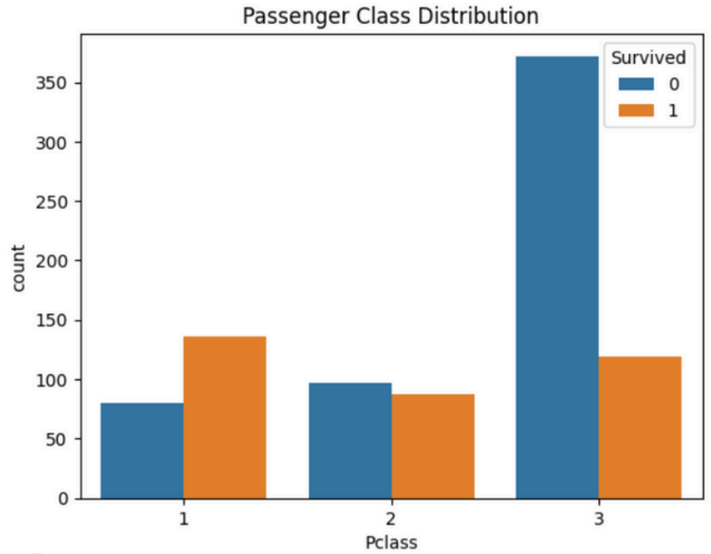


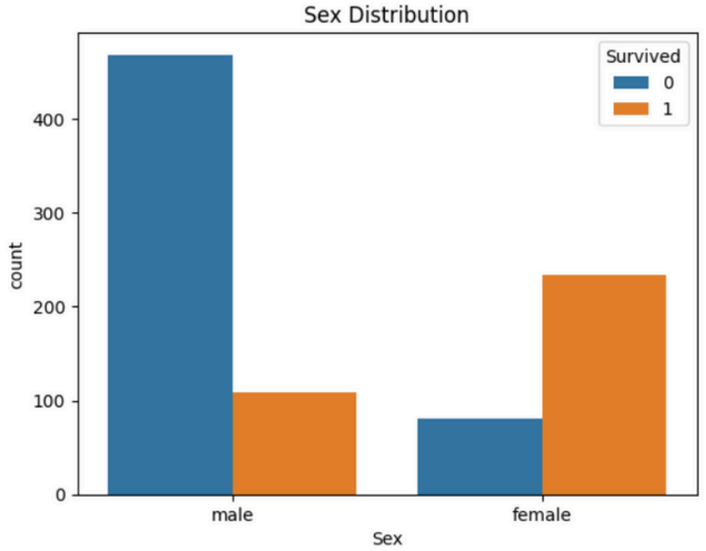
🔑 **Passengers with high fares are more likely to survive** than those with low fares.

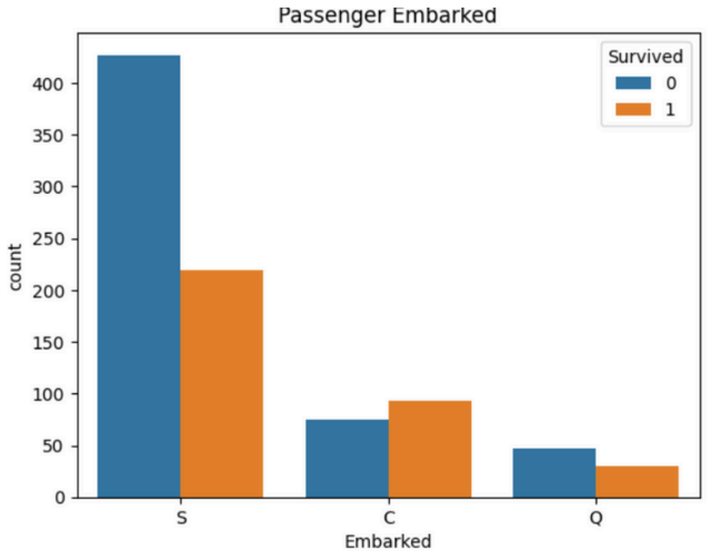


🔑 **Passengers who are accompanied by 1 or 2 parents/children** have a **higher chance of survival**.

Survival Rate by PClass, Sex, and Embarked 🔍







Passengers with **highest ticket classes (1) have the highest chance of survival**.

Female passengers have a higher chance of survival than male passengers.

Passengers who departed from Southampon harbour had the highest chance of survival compared to Cherboug & Queensstown harbour.

Findings and Insights 🔎

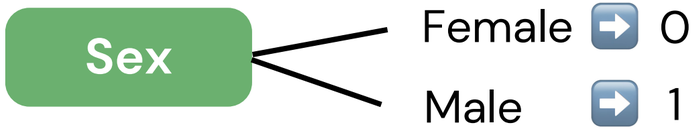
- The factor that contributed most to the survival of Titanic passengers is Ticket Class ('PClass').
- **Insufficient Facilities**: Fewer passengers survived than those who did not may have happened because of the insufficient rescue facilities such as lifeboats.
- **Social Prioritisation**: Upper-class passengers and those who paid for more expensive tickets were given higher priority, suggesting the possibility of a rescue SOP based on economic status.
- **Child and Women Rescue**: Children and women have a greater chance of survival, suggesting a possible application of the rule of "women and children first."
- **Family Coordination**: Passengers with small families are easier to rescue, maybe because the rescue SOP allows them to coordinate more easily.

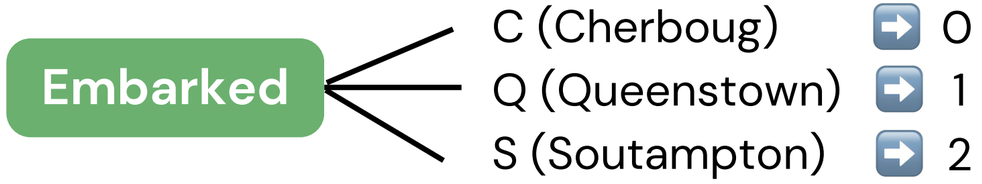
⚠️ Generally, the safety SOPs on the Titanic seem insufficient and show bias in the prioritization of rescue, both in terms of facilities and social roles.

Label Encorder 🖋️

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	7.2500	2
1	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	71.2833	0
2	1	3	Heikkinen, Miss. Laina	0	26.0	0	0	7.9250	2
3	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	53.1000	2
4	0	3	Allen, Mr. William Henry	1	35.0	0	0	8.0500	2

The **Label Encoder function** is used in order to **convert categorical data into numerical data**. In this case, the data to be converted are 'Sex' and 'Embarked'.





Feature Selection 📂

X Variables

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	1	22.0	1	0	7.2500	2
1	1	0	38.0	1	0	71.2833	0
2	3	0	26.0	0	0	7.9250	2
3	1	0	35.0	1	0	53.1000	2
4	3	1	35.0	0	0	8.0500	2
...	...	...	...	...	...	...	...
886	2	1	27.0	0	0	13.0000	2
887	1	0	19.0	0	0	30.0000	2
888	3	0	28.0	1	2	23.4500	2
889	1	1	26.0	0	0	30.0000	0
890	3	1	32.0	0	0	7.7500	1

✨ The features selected as X variables for modelling are **'PClass'**, **'Sex'**, **'Age'**, **'SibSp'**, **'Parch'**, **'Fare'**, and **'Embarked'** columns.

Y Variable

```
0      0
1      1
2      1
3      1
4      0
      ..
886    0
887    1
888    0
889    1
890    0
Name: Survived
```

✨ The selected target as variable Y is the **'Survived'** column.

Splitting Data ✨

X Train

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
331	1	1	45.5	0	0	28.5000	2
733	2	1	23.0	0	0	13.0000	2
382	3	1	32.0	0	0	7.9250	2
704	3	1	26.0	1	0	7.8542	2
813	3	0	6.0	4	2	31.2750	2
...	...	...	...	...	...	...	...
106	3	0	21.0	0	0	7.6500	2
270	1	1	28.0	0	0	31.0000	2
860	3	1	41.0	2	0	14.1083	2
435	1	0	14.0	1	2	120.0000	2
102	1	1	21.0	0	1	77.2875	2

Y Train

```
331    0
733    0
382    0
704    0
813    0
      ..
106    1
270    0
860    0
435    1
102    0
```

Data Train

Out of the 891 data points, **80% (712 data points) were used for training** the X and Y variables.

Splitting Data 🪄

X Test

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
709	3	1	28.0	1	1	15.2458	0
439	2	1	31.0	0	0	10.5000	2
840	3	1	20.0	0	0	7.9250	2
720	2	0	6.0	0	1	33.0000	2
39	3	0	14.0	1	0	11.2417	0
...	...	...	...	...	...	...	...
433	3	1	17.0	0	0	7.1250	2
773	3	1	28.0	0	0	7.2250	0
25	3	0	38.0	1	5	31.3875	2
84	2	0	17.0	0	0	10.5000	2
10	3	0	4.0	1	1	16.7000	2

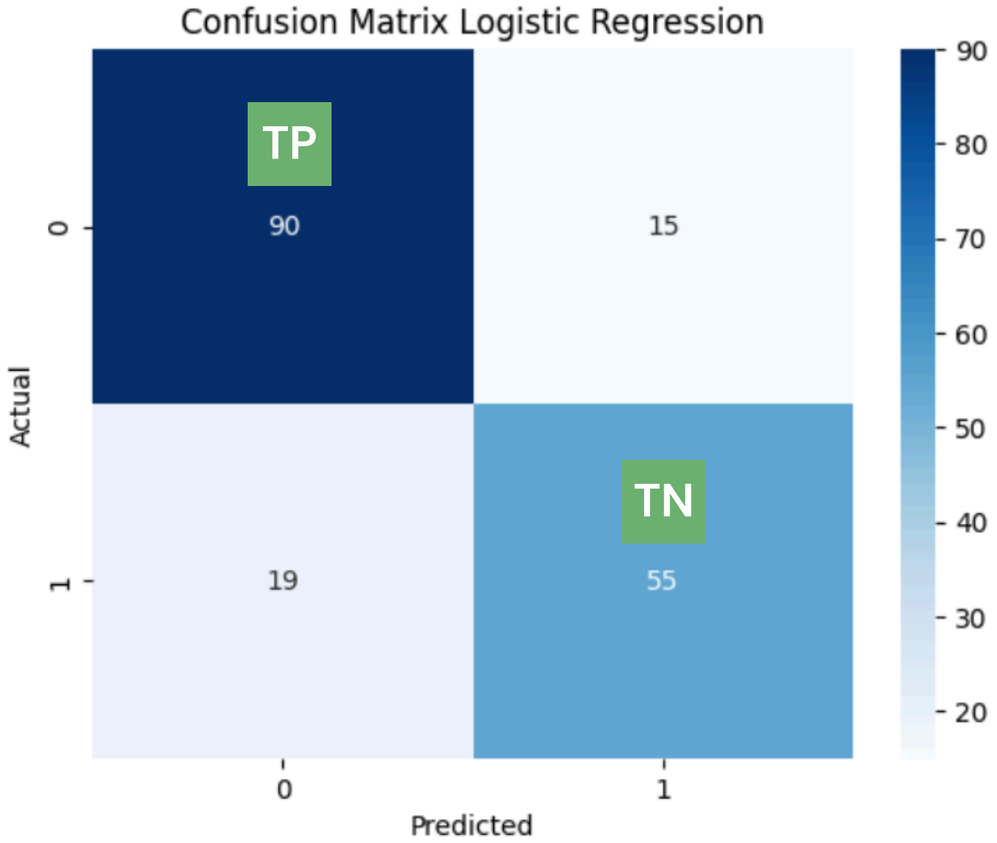
Y Test

```
709    1
439    0
840    0
720    1
39     1
      ..
433    0
773    0
25     1
84     1
10     1
```

Data Test

Out of the 891 data points, **20% (129 data points) were used for training** the X and Y variables.

Confusion Matrix Logistic Regression 📈

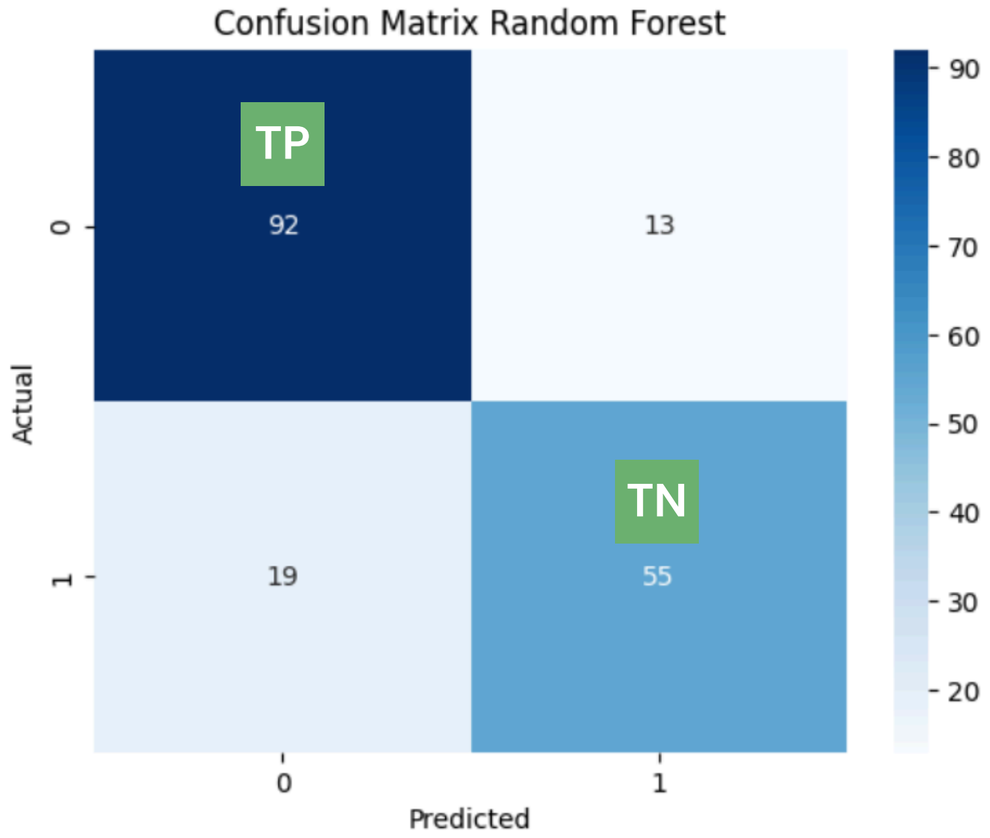


Evaluation

The Logistic Regression model achieved **an accuracy of 81%**, with the following information:

- True Positive (TP) : The model correctly predicted **90 passengers who did not survive.**
- True Negative (TN) : The model correctly predicted **55 passengers who survived.**

Confusion Matrix Random Forest 🌲

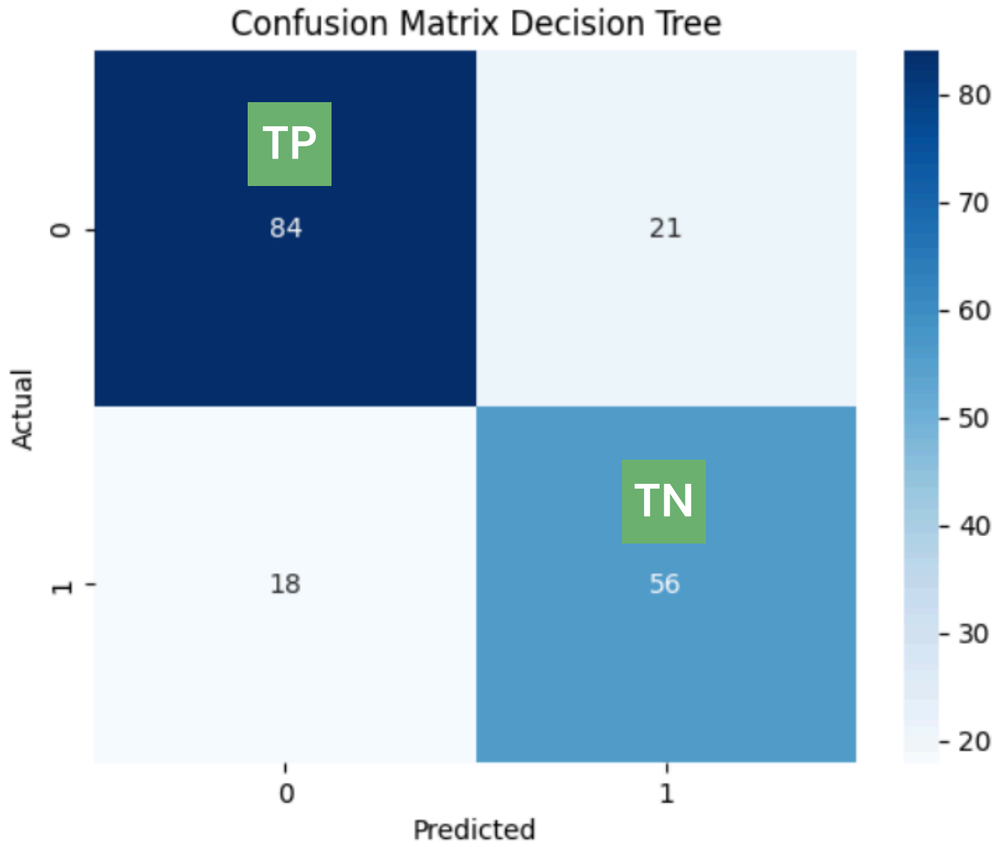


Evaluation

The Random Forest model achieved **an accuracy of 82%**, with the following information:

- True Positive (TP) : The model correctly predicted **92 passengers who did not survive.**
- True Negative (TN) : The model correctly predicted **55 passengers who survived.**

Confusion Matrix Decision Tree 🌳

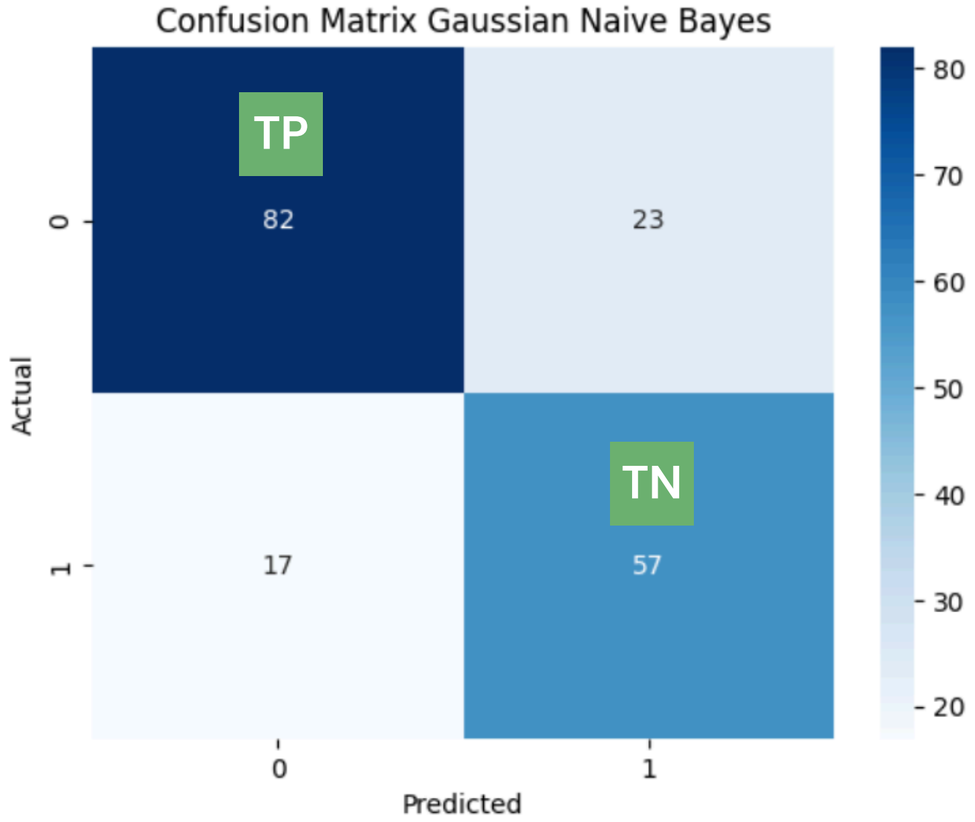


Evaluation

The Decision Tree model achieved **an accuracy of 78%**, with the following information:

- True Positive (TP) : The model correctly predicted **84 passengers who did not survive.**
- True Negative (TN) : The model correctly predicted **56 passengers who survived.**

Confusion Matrix Gaussian Naive Bayes 🕶️

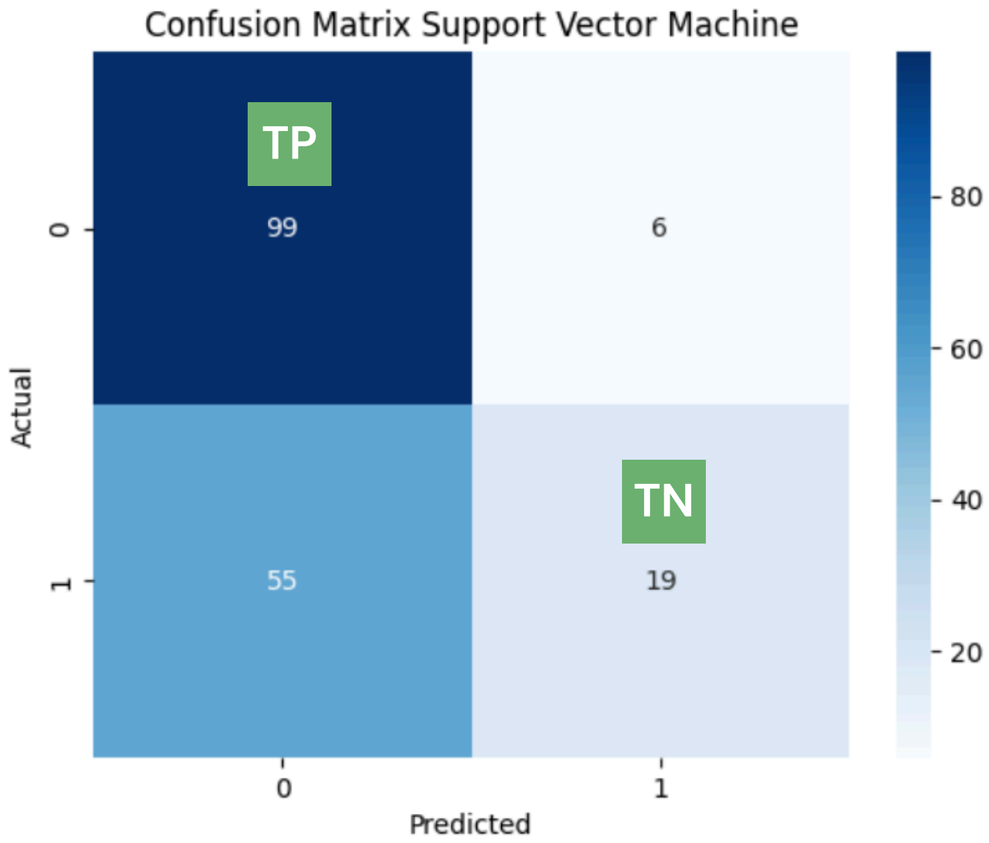


Evaluation

The Gaussian Naive Bayes model achieved **an accuracy of 78%**, with the following information:

- True Positive (TP) : The model correctly predicted **82 passengers who did not survive.**
- True Negative (TN) : The model correctly predicted **57 passengers who survived.**

Confusion Matrix Support Vector Machine ⚙️



Evaluation

The Support Vector Machine model achieved **an accuracy of 66%**, with the following information:

- True Positive (TP) : The model correctly predicted **99 passengers who did not survive.**
- True Negative (TN) : The model correctly predicted **19 passengers who survived.**

Comparison ⚖️

Model	Accuracy	True Positive (TP)	True Negative (TN)
Logistic Regression	81%	90	55
Random Forest	82%	92	55
Decision Tree	78%	84	56
Gaussian Naive Bayes	78%	82	57
Support Vector Machine	66%	99	19

- ✨ The model with a **good balance of TP and TN, such as Random Forest and Logistic Regression**, tends to be more accurate.
- ✨ The **Gaussian Naive Bayes and Decision Tree models have similar accuracy** (78%), but Gaussian Naive Bayes is better at identifying negative samples (passenger who survived).
- ✨ **Accuracy does not necessarily reflect overall performance, such as in the Support Vector Machine model.** This may be due to the very low TN (19).
- ✨ From the table, the **Random Forest model has the highest accuracy** (82%), with a TP of 92 and TN of 55, showing **good performance in classifying both passengers who survived and those who did not.**

Conclusion 💌

- 🚢 The factor that contributed most to the survival of Titanic passengers is Ticket Class.
- 🚢 The safety SOPs on the Titanic appear to be insufficient and exhibit bias in the prioritization of rescue, both in terms of available facilities and social roles.
- 🚢 The supervised learning model that provides the best accuracy in predicting the survival of Titanic passengers is Random Forest.

What's Next? 🚀

- Optimize models like Random Forest and Support Vector Machine that have been used by performing Hyperparameter Tuning to see if the accuracy can be improved.
- To better validate the model, use k-fold cross-validation techniques to ensure that the model has good generalization and more stable performance.
- Model evaluation can be conducted with precision, recall, or F1-score to see the performance of the model from various points of view.

Thank You

Let’s get in touch!

nabilaputri5609@gmail.com

Nabila Putri Dian Luthfiyyah

github.com/nabputttt