MINI PROJECT REPORT

On

"Titanic Survival Prediction Using Machine Learning"

A Report Submitted for a mini project for : Laboratory Practice-III in 7th Semester Computer Engineering

Fourth Year (COMPUTER ENGINEERING)

Academic Year 2025-26

Submitted by-

Student Name	Roll No.
Mr. Anuj Sachin Dhole	B21042
Mr. Mohan Ganesh Dinkar	B21046

Zeal Education Society's

Zeal College of Engineering & Research Narhe, Pune

– 411041

Department of Computer Engineering

Zeal College of Engineering & Research Narhe,

Pune-411041.

Zeal Education Society's

Zeal College of Engineering & Research Department of

Computer Engineering

CERTIFICATE

This is to certify that Project Entitled **"Titanic Survival Prediction Using Machine Learning"** is Bonafide work carried out by **Mohan Ganesh Dinkar (B21046) and Anuj Sachin Dhole (B21042)** of this institute and the work has been carried out by him under the supervision **of Prof. Kalpana D. Sonval** and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University, for the award of Fourth Year Engineering (Computer Engineering).It is certified that all corrections/suggestions indicated for internal assignment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Prof. Kalpana D. Sonval — Project Guide

Prof. A. V. Mote — H. O. D

ACKNOWLEDGEMENT

I take this opportunity to thank our project guide **Prof. Kalpana D. Sonval** and Head of Department **Prof. A. V. Mote** for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. I am also thankful to all the staff members of the Computer Engineering Department for their valuable time, support, comments, suggestions and persuasion. I would also like to thank the institute for providing the required facilities, Internet access and important books.

INDEX

ABSTRACT

The Titanic disaster remains one of the most studied maritime tragedies in history, offering valuable insights into human behavior and survival dynamics. This project focuses on building a machine learning model to predict the likelihood of passenger survival based on demographic and socio-economic attributes such as age, gender, class, and family status. Using the Titanic dataset, data preprocessing techniques such as handling missing values, feature encoding, and normalization are applied to prepare the data for model training. Various classification algorithms—including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines—are evaluated to identify the model with the highest predictive accuracy. The analysis reveals that factors such as gender and passenger class played a crucial role in determining survival chances. The developed model not only demonstrates effective predictive capability but also highlights the importance of data-driven decision-making in understanding historical and social phenomena.

SOFTWARE REQUIREMENT

- **Operating System:** Windows 10 or 11
- **Programming Language:** Python 3.8 or above
- **Development Environment:** Jupyter Notebook, Google Colab, or VS Code
- **Libraries/Frameworks:** NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn
- **Dataset:** Titanic Dataset (from Kaggle)
- **Data Visualization Tools:** Matplotlib and Seaborn
- **Model Evaluation Tools:** Scikit-learn (train-test split, accuracy, confusion matrix, etc.)
- **Version Control (Optional):** GitHub
- **System:** Windows 11

INTRODUCTION

The sinking of the RMS Titanic on April 15, 1912, remains one of the most infamous maritime tragedies in history. The ship, believed to be "unsinkable," collided with an iceberg on its maiden voyage from Southampton to New York City, resulting in the loss of over 1,500 lives out of more than 2,200 passengers and crew. This disaster not only shocked the world but also highlighted the social and economic inequalities that influenced survival during the tragedy. Over the years, the Titanic dataset has become a benchmark for data scientists and machine learning practitioners due to its rich combination of numerical, categorical, and textual data.

The primary objective of this project is to develop a machine learning model capable of predicting the likelihood of survival for Titanic passengers based on features such as age, gender, passenger class, fare, and family relationships. By analyzing these variables, we can gain valuable insights into which factors played the most significant roles in determining survival probabilities.

The project involves several key stages, including data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model selection, and performance evaluation. Popular algorithms such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines are used to train and compare predictive models. Evaluation metrics like accuracy, precision, recall, and F1-score are employed to assess model performance.

In addition to its technical relevance, this project also serves an educational purpose by demonstrating how real-world data can be transformed into actionable insights using modern analytical tools. It exemplifies how predictive analytics can be applied not only to historical datasets but also to broader fields such as risk analysis, decision support, and human behavior modeling. Ultimately, this project showcases the power of machine learning in uncovering meaningful patterns from data and emphasizes the importance of data-driven decision-making in understanding complex social phenomena.

PROBLEM STATEMENT

Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data (i.e. name, age, gender, socio-economic class, etc.).

OBJECTIVE:

The objective of this project is to build a machine learning model that predicts the survival of Titanic passengers based on features such as age, gender, class, and family size. It aims to preprocess and analyze the dataset, apply various classification algorithms, and evaluate their performance to identify key factors influencing survival. The project demonstrates the use of data science techniques to extract insights and make accurate, data-driven predictions from historical data.

OUTCOME:

The outcome of this project is a trained machine learning model capable of predicting the survival of Titanic passengers with high accuracy based on their demographic and socio-economic features. The project successfully identifies key factors influencing survival, such as gender, passenger class, and age. It demonstrates how data preprocessing, feature engineering, and model evaluation techniques can be combined to produce meaningful insights. The results highlight the effectiveness of machine learning in analyzing historical data and serve as a foundation for understanding predictive analytics in real-world scenarios.

IMPLEMENTATION CODE

In [27]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

In [28]:
```
data = pd.read_csv('train.csv')  # Use the training dataset
```

In [29]:
```
data
```

Out[29]:

	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

891 rows × 12 columns

In [30]:
```
# Select relevant features
features = ['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare']
data = data[features + ['Survived']]
```

In [31]:
```
data
```

Out[31]:

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

891 rows × 7 columns

891 rows × 7 columns

In [51]:
```
# Convert categorical to numeric
data =data.copy()  # or train_df[condition].copy()
data['Sex'] = data['Sex'].map({'male': 0, 'female': 1})
data
```

Out[51]:

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

891 rows × 8 columns

891 rows × 7 columns

In [32]:
```
data.loc[:, 'Age'] = data['Age'].fillna(data['Age'].median())
data
```

Out[32]:

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

891 rows × 7 columns

890	3	NaN	32.0	0	0	7.7500	0	0	

891 rows × 8 columns

In [34]:
```
X = data[features]
y = data['Survived']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [35]:
```
model = RandomForestClassifier()
model.fit(X_train, y_train)
```

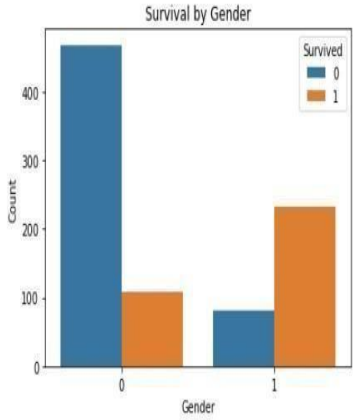
Out[35]: RandomForestClassifier()

In [36]:
```
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
```

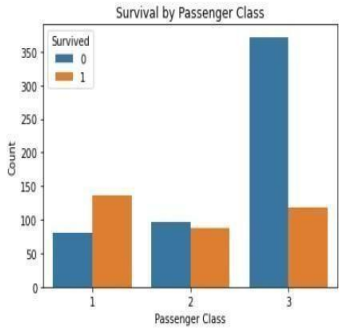
Accuracy: 0.8044692737430168

In [39]:
```
#Survived by Gender
sns.countplot(x='Sex', hue='Survived', data=data)
plt.title('Survival by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

In [39]:
```
#Survived by Gender
sns.countplot(x='Sex', hue='Survived', data=data)
plt.title('Survival by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

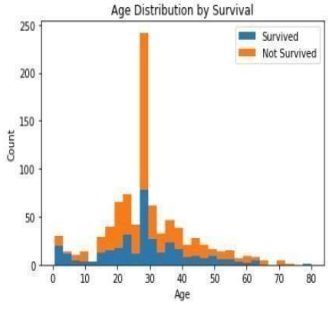


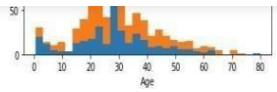
In [41]:
```
#Survived by passenger
sns.countplot(x='Pclass', hue='Survived', data=data)
plt.title('Survival by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Count')
plt.show()
```



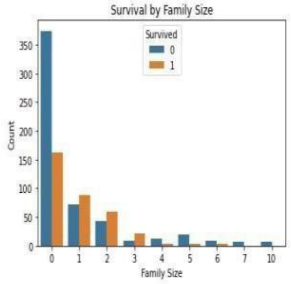
In [43]:
```
#Age Distribution of Survivors
survived = data[data['Survived'] == 1]['Age']
not_survived = data[data['Survived'] == 0]['Age']

plt.hist([survived, not_survived], bins=30, stacked=True, label=['Survived', 'Not Survived'])
plt.title('Age Distribution by Survival')
plt.xlabel('Age')
plt.ylabel('Count')
plt.legend()
plt.show()
```





In [48]:
```
#Family Size vs. Survival
sns.countplot(x='FamilySize', hue='Survived', data=data)
plt.title('Survival by Family Size')
plt.xlabel('Family Size')
plt.ylabel('Count')
plt.show()
```



In [ ]:

CONCLUSION

The Titanic Survival Prediction project demonstrates the practical application of machine learning in analyzing real-world data and deriving meaningful insights. By exploring passenger information such as gender, age, class, and family relations, the project effectively identifies the factors that most influenced survival during the disaster. Among the models tested, classification algorithms like Logistic Regression and Random Forest provided reliable results in predicting survival outcomes.

This project highlights the importance of data preprocessing, feature selection, and model evaluation in building accurate predictive systems. Beyond its technical value, it also offers a deeper understanding of how socio-economic conditions and human characteristics can impact survival in critical situations. Overall, the project reinforces the power of data-driven approaches in uncovering patterns, supporting decision-making, and solving complex analytical problems.

REFERENCES:

- Sharma, R., & Kumar, S. (2021). Blockchain for Electronic Voting System—Review. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC8434614/

- El Khatib, F., & et al. (2024). Blockchain-Based Electronic Voting Systems: A Case Study. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2666603024000046

- Singh, A., & et al. (2023). Blockchain-Based E-Voting System on Ethereum Private Network. ResearchGate.