TITANIC CLASSIFICATION

-PREDICTING PASSENGER SURVIVAL

An Application Development – 1 (Project) Report Submitted

In partial fulfillment of the requirement for the award of the degree of

# Bachelor of Technology in

**Computer Science and Engineering (Data Science)**

**by**

**ENUMULA MANOJ KUMAR 21N31A6713**

### GADIDAS THARUN 21N31A6714

**GAJJELA SANJAY 21N31A6715**

Under the guidance of

### A.NAVEEN KUMAR

### Asst.Professor

**Department of Emerging Technologies**

MRCET (Autonomous Institution, UGC Govt. of India)



MRCET CAMPUS

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**(EMERGING TECHNOLOGIES)**

**MALLA REDDY COLLEGE OF ENGINEERING AND TECHNOLOGY**

**(Autonomous Institution – UGC, Govt. of India)**

**(Affiliated to JNTU, Hyderabad, Approved by AICTE, Accredited by NBA & NAAC – ‘A’ Grade, ISO 9001:2015 Certified)**

Maisammaguda (v), Near Dullapally, Via: Kompally, Hyderabad – 500 100, Telangana State, India

**2023-2024**



CERTIFICATE

This is to certify that this is the Bonafide record of the project titled “**Titanic Classification-Predicting Passenger Survival**”, submitted by **MANOJKUMAR** (21N31A6713) **THARUN (**21N31A6714) and **SANJAY (**21N31A6715) of **B.Tech III YEAR – I Semester** in the partial fulfillment of the requirements for the degree of **Bachelor of Technology** in **Computer Science and Engineering (Data Science)**, Dept. of CSE (Emerging Technologies) during the year 2023-2024. The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

**Dr. M V Kamal Dr. P Dileep**

Head Of the Department (ET) Project Coordinator

**Department of CSE (ET) Department of CSE (ET)**

**EXTERNAL EXAMINER A.NAVEEN KUMAR**

Asst.Professor&

**Department of CSE (ET)**

# DECLARATION

We hereby declare that the project entitled “**Titanic Classification-Predicting Passenger Survival**” submitted to **Malla Reddy College of Engineering and Technology,** affiliated to Jawaharlal Nehru Technological University Hyderabad (JNTUH) as part of III Year B.Tech – I Semester and for the partial fulfillment of the requirement for the award of **Bachelor of Technology** in **Computer Science and Engineering (DataScience)** is a result of original research work done by me.

It is further declared that the project report or any part thereof has not been previously submitted to any University or Institute for the award of degree or diploma.

MANOJKUMAR (21N31A6713) ` THARUN (21N31A6714)

**CHOWDARY MEDHA - 20N31A6712**

**KUDIKALA SRI VAKULA - 20N31A6736**

**MAHESHWARA HARDHIK - 20N31A6738**

SANJAY (21N31A6715)

**ACKNOWLEDGEMENTS**

We feel ourself honored and privileged to place our warm salutation to our college “Malla Reddy College of Engineering and Technology (Autonomous Institution – UGC Govt. of India) and our Principal **Dr. S Srinivasa Rao,** Professor who gave us the opportunity to do the Application Development -1 (Project) during our III Year B.Tech and profound the technical skills.

We express our heartiest thanks to our Director **Dr. V S K Reddy,** Professor for encouraging us in every aspect of our project and helping us realize our full potential.

We also thankful to our Head of the Department **Dr. M V Kamal,**Professor for providing training and guidance, excellent infrastructure and a niceatmosphere for completing this project successfully.

We would like to express our sincere gratitude and indebtedness to our project supervisor **A.NAVEEN KUMAR(Asst.professor)** for his valuable suggestions and interest throughout the course of this project.

We convey our heartfelt thanks to our Project Coordinator

**Dr. P DILEEP,** Professor for allowing for their regular guidance and constant encouragement during our dissertation work.

We would like to thank all our supporting **staff** of the Department of CSE (Emerging Technologies) and even all other department who have been helpful directly and in-directly in making our project a success.

Finally, we would like to take this opportunity to thank our **families** for their support and blessings for completion of our project that gave us the strength to do our project.

MANOJKUMAR (21N31A6713)

THARUN (21N31A6714)

SANJAY (21N31A6715)

**ABSTRACT**

The sinking of the RMS Titanic on April 15, 1912, remains one of the most iconic maritime disasters in history, resulting in the loss of over 1,500 lives. This dataset classification projec

-t aims to predict the survival outcome of Titanic passengers using various machine learning techniques.

The dataset consists of information about passengers, such as their age, gender, class, ticket fare, embarkation point, and whether they survived or not. The primary objective is to build predictive models that can accurately classify passengers into two categories: "Survived" or "Did Not Survive.

Results indicate that certain factors, such as passenger class and gender, significantly influence the likelihood of survival. The final model demonstrates commendable performance in predicting passenger survival, providing insights into the intricate relationships between variou s features and outcomes. The implications of these findings extend beyond historical interest, with potential applications in contemporary survival prediction scenarios

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter No.** | | **Contents** | **Page no** |
| 1 |  | Introduction | 1 |
|  | 1.1 | Problem Definition | 2 |
|  | 1.2 | Existing System | 3 |
|  | 1.3 | Proposed System | 4 |
| 2 |  | System Requirements | 6 |
|  | 2.1 | Software Requirements | 7 |
|  | 2.2 | Hardware Requirements | 8 |
| 3 |  | Flow chart ,Sequential diagram ,Dataflow Diagrams / UML Diagrams | 9 |
| 4 |  | Implementations | 13 |
| 5 |  | Output screen | 15 |
| 6 | 6.1 | Conclusion | 19 |
|  | 6.2 | Future Scope | 20 |
| 7 |  | References | 21 |

# CHAPTER 1 INTRODUCTION

* The sinking of the RMS Titanic is one of the most infamous maritime disasters in history. On the night of April 15, 1912, the Titanic struck an iceberg and sank, resulting in the loss of over 1,500 lives out of the 2,224 passengers and crew on board. This tragic event has captivated the public's imagination for over a century and has also served as a fascinating case study for data science and machine learning.
* In the field of data science, the Titanic dataset is often used as a beginner's introduction to classification problems. The challenge is to predict whether a passenger on the Titanic survived or not based on various attributes and features. This problem is not only an opportunity to learn and apply data science techniques but also a way to pay tribute to the individuals who were on board the ill-fated ship.
* In summary, this study delves into the application of machine learning to analyze the Titanic dataset, aiming to build a robust classification model for predicting passenger survival. By doing so, it contributes to the ongoing discourse on data-driven approaches in historical research and underscores the relevance of such methodologies in uncovering hidden patterns within complex and tragic events.

.

**1.1PROBLEM DEFINITION:**

The sinking of the RMS Titanic in 1912 stands as a poignant reminder of the human cost associated with maritime disasters. In the aftermath of this tragic event, there is a compelling need to delve into the dataset encompassing passenger information to develop a predictive model that can ascertain the factors influencing survival outcomes. The overarching problem at hand is to create an effective classification model that accurately predicts whether a passenger on the Titanic survived or not based on their individual attributes.

Key Components of the Problem:

1. **Dataset Complexity:** The Titanic dataset is multifaceted, comprising variables such as passenger class, age, gender, embarkation port, and family relationships. The challenge is to navigate through this complexity and identify the key features that significantly impact the likelihood of passenger survival.
2. **Missing Data and Data Preprocessing:** The dataset may contain missing values and inconsistencies that require careful handling. Addressing these issues through robust data preprocessing techniques is essential to ensure the reliability of the predictive model.
3. **Model Selection:** Choosing an appropriate machine learning algorithm is critical. The problem involves exploring a range of classification algorithms, including logistic regression, decision trees, random forests,

and support vector machines, to determine which model best captures the nuances of passenger survival.

1. **Feature Importance and Interpretability:** Understanding the relative importance

different features in predicting survival is crucial for gaining insights into the underlying dynamics of the disaster. The model should not only provide accurate predictions but also offer interpretability to discern the significance of various factors.

1. **Evaluation Metrics:** Establishing a comprehensive set of evaluation metrics, such as accuracy, precision, recall, and F1 score, is essential to quantify the performance of the classification model. Striking a balance between these metrics is necessary to ensure a well-rounded assessment.
   1. **EXISTING SYSTEM**

* The Titanic dataset is a classic example in machine learning for binary classification tasks. Many machine learning enthusiasts and practitioners use it as a beginner's project to predict passenger survival based on various features. The dataset typically includes information

about passengers such as age, sex, class, and whether they survived or not.

* There are numerous existing systems and models created for predicting passenger survival

on the Titanic. Some common machine learning algorithms used for this task include logistic regression, decision trees, random forests, and support vector machines.

* Many tutorials and blog posts walk through the process of creating a predictive model for the Titanic dataset using different machine learning frameworks like scikit-learn or TensorFlow.

## PROPOSED SYSTEM

1. **Data Preprocessing:** We clean and preprocess the dataset, handling missing values.

encoding categorical features, and normalizing numerical attributes

1. **Feature Selection:** We select the most relevant features that contribute to the prediction

of survival outcomes. This involves evaluating feature importance and correlation analysis

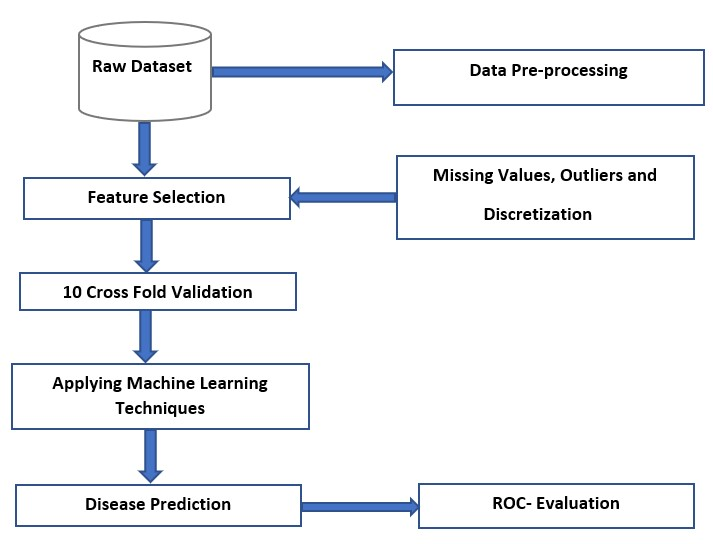
1. **Model Selection:** We experiment with various machine learning algorithms, including

Logistic regression, decision trees, random forests, support vector machines, and neural networks, to identify the most suitable model for the task.

1. **Model Evaluation:** We assess the performance of each model using evaluation metrics

such as accuracy, precision, recall, F1-score, and ROC-AUC.

**System architecture:**



### Fig 1. Architecture

**CHAPTER 2 REQUIRED SYSTEM**

Requirements analysis in systems engineering and software engineering encompasses those tasks that go into determining the needs or conditions to meet for a new or altered product, taking account of the possibly conflicting requirements of the various stakeholders, such as beneficiaries or users. A software requirements specification (SRS) is a document that is created when a detailed description of all aspects of the software to be built must be specified before the project is to commence. It is important to note that a formal SRS is not always written.

In fact, there are many instances in which effort expended on an SRS might be better spent in other software engineering activities. Requirements analysis is critical to the success of a development project. Requirements must be actionable, measurable, testable, related to identified business needs or opportunities, and defined to a level of detail sufficient for system design. By analyzing different hardware components and familiar software following are hardware & software used in our project:

### SOFTWARE REQUIRMENTS

1. **Operating system**: Windows 7 or newer. (32-bit or 64-bit)
2. **Browser**: A chromium-based web browser. Example: Google Chrome Or Mozilla Firefox
3. Code editor like **JUPYTER NOTEBOOK**
4. **Programming Language:**

**Python:** Widely used in machine learning and data science.

1. **Integrated Development Environment (IDE):**

**Jupyter Notebook or Spyder:** Popular for data analysis and machine learning tasks.

**Libraries:** Tkinter to implement photo image.

**Pandas:** For data manipulation and analysis.

**NumPy:** Essential for numerical operations in Python.

1. **Scikit-learn:** Includes tools for machine learning tasks such as classification.
2. **Matplotlib and Seaborn:** For data visualization.

**Data Collection:**

**Kaggle:** You can find the Titanic dataset on Kaggle. It includes information about passengers, such as age, class, and whether they survived.

1. **Data Preprocessing:**

**Handling Missing Data:** Decide on a strategy for dealing with missing values (e.g., imputation or removal).

**Encoding Categorical Variables:** Convert categorical variables into a format suitable for machine learning models.

1. **Model Evaluation:**

**Metrics:** Decide on evaluation metrics such as accuracy, precision, recall, or F1-score.

1. **Documentation:Jupyter Notebook, Markdown, or Readme:** Document your process, code, and results.

### HARDWARE REQUIRMENTS

The hardware requirements for developing a Titanic classification model are relatively

modest, as the development and training of machine learning models can often be done on standard personal computers. Here are the basic hardware requirements:

1. **Processor (CPU):**
   * A modern multi-core processor (e.g., Intel Core i5 or i7, AMD Ryzen 5 or 7) should

be sufficient for development and training small to medium-sized machine learning models.

1. **Memory (RAM):**
   * At least 8 GB of RAM is recommended for handling data manipulation and running

machine learning algorithms.

1. **Storage:**
   * A solid-state drive (SSD) is preferable for faster data access, but a standard hard disk

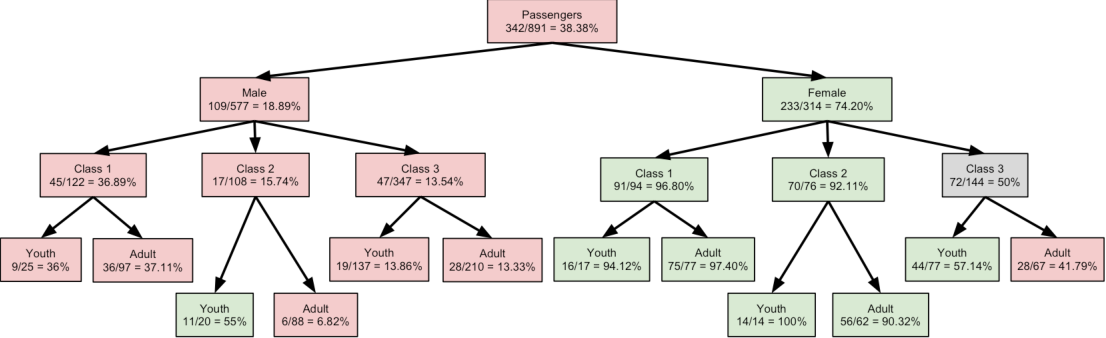
drive (HDD) with sufficient capacity (at least 256 GB) should also work.

4. **Internet Connection:**

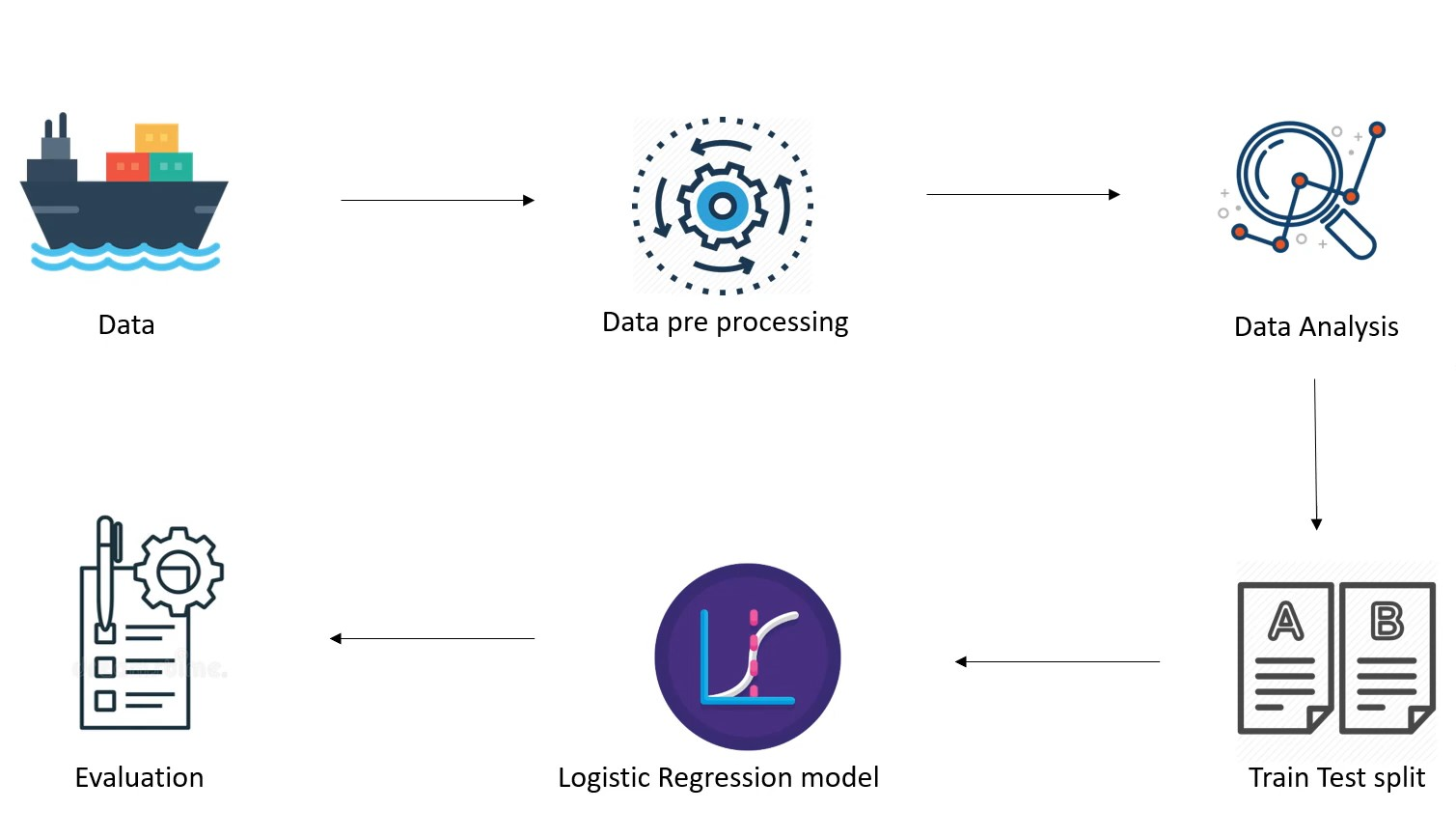
* A stable internet connection is required for data retrieval (if using online datasets),

software installations, and accessing documentation.

### CHAPTER 3

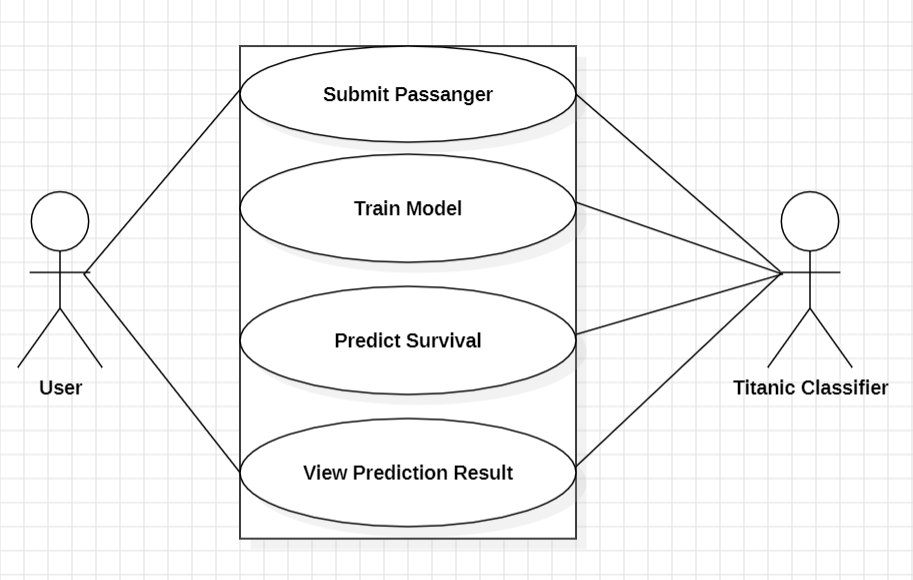


**Fig 2. Flowchart**



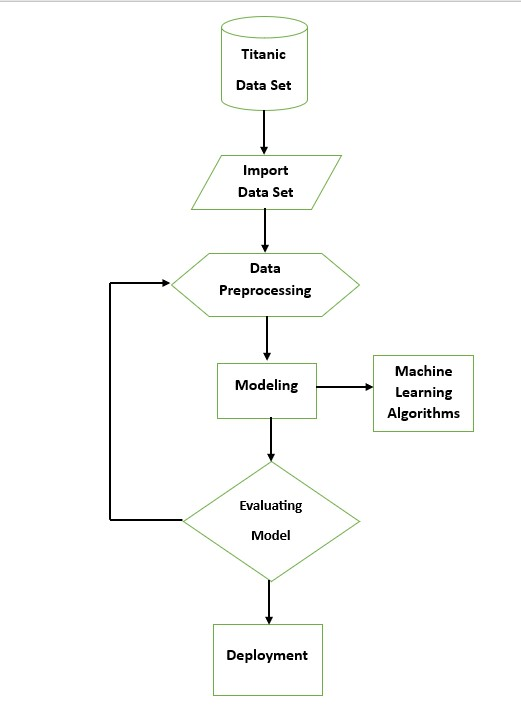
### Fig3.Sequential Diagram

**3.1 UML DIAGRAM**

****

**Fig 4.USE CASE DIAGRAM**

### DATAFLOW



**Fig 5. Data Flow Diagram**

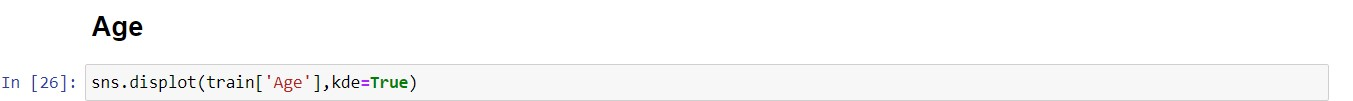
### CHAPTER 4

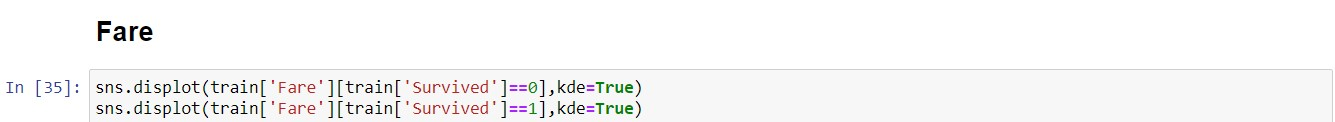
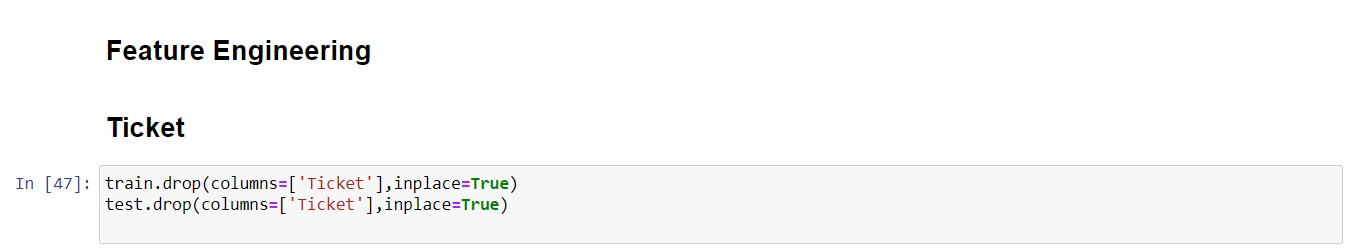
**IMPLEMENTATION :**

**CLASSIFIER VISUALIZATION CODE:**

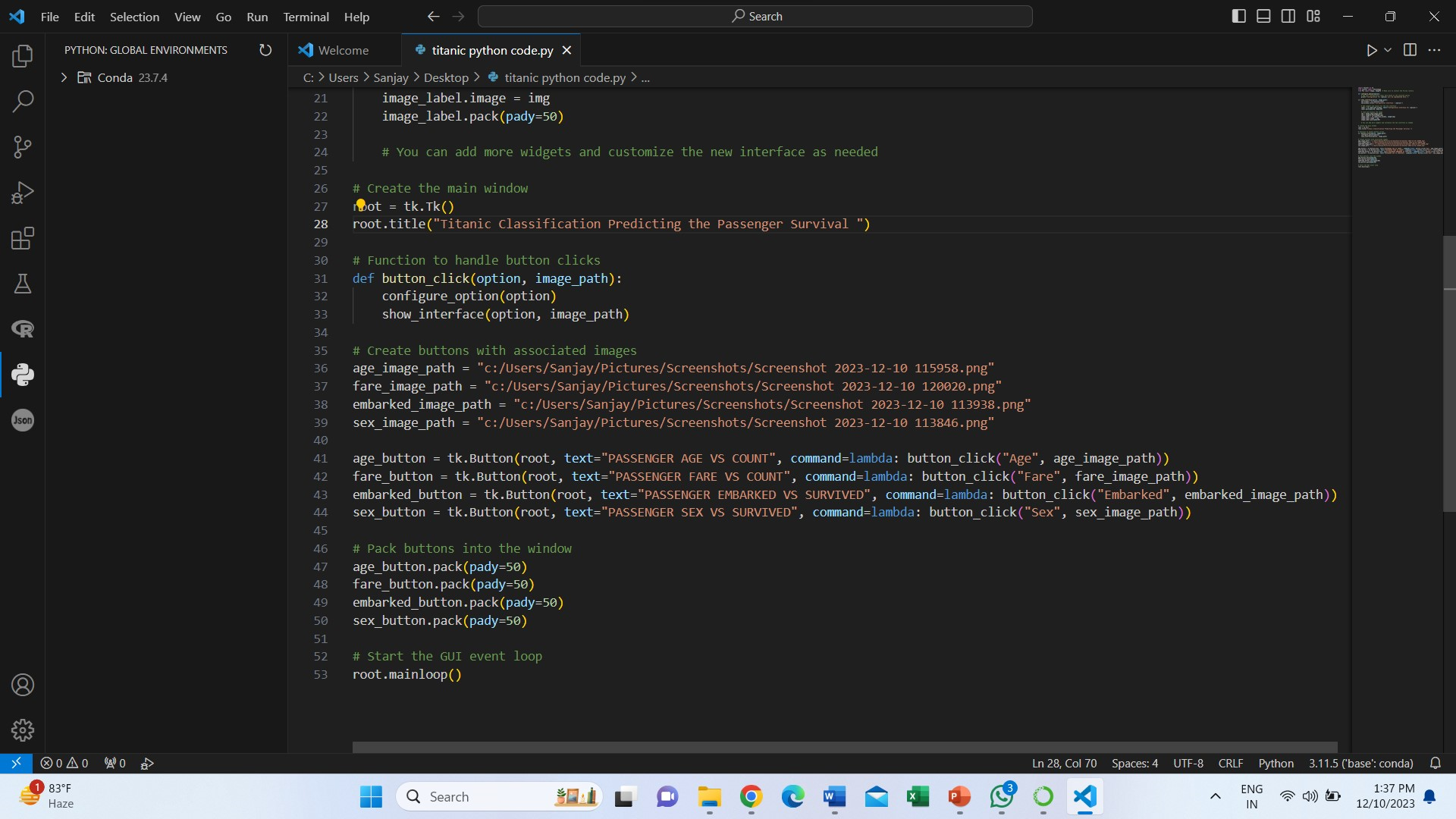
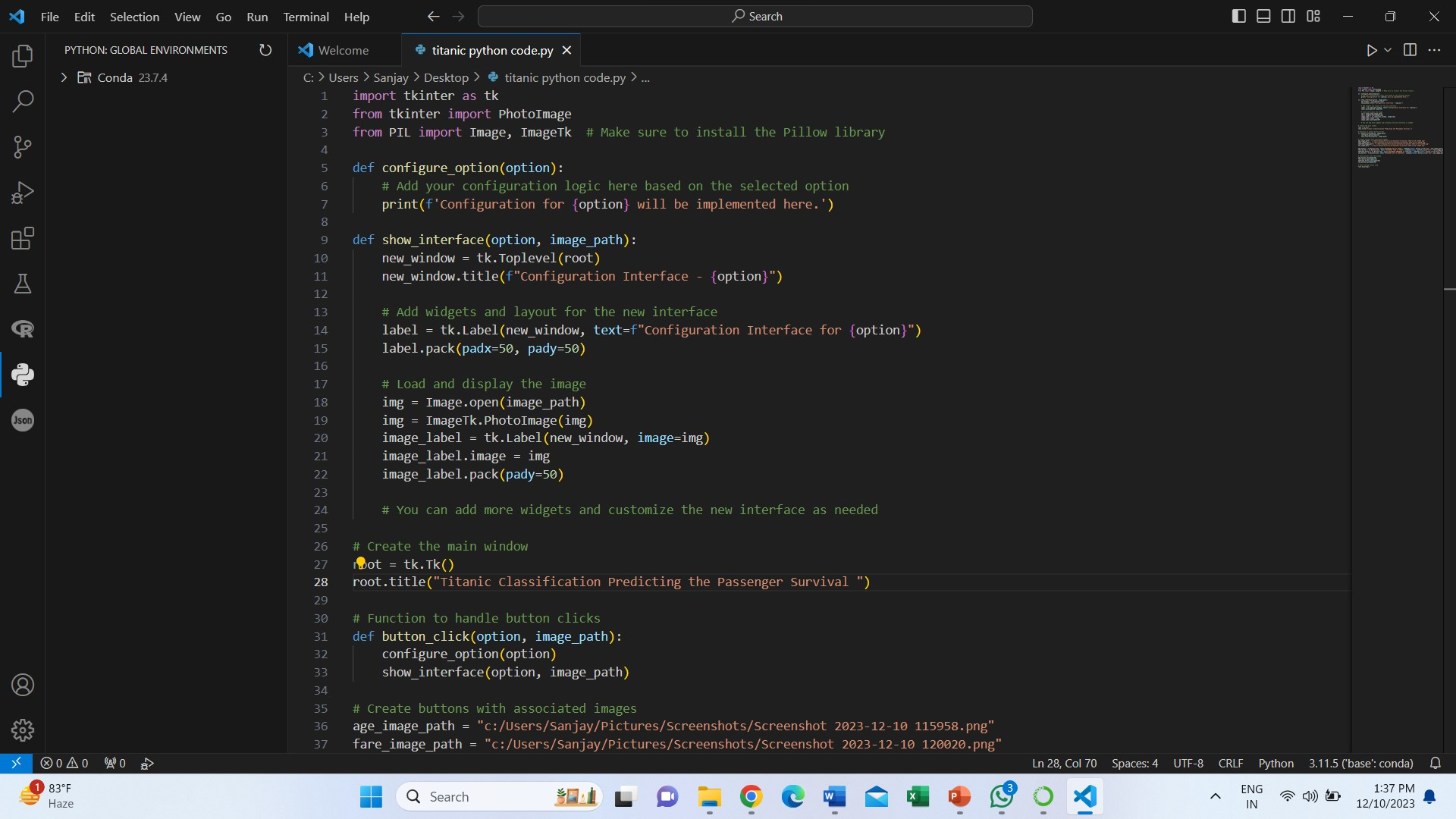








**INTERFACE CODE:**

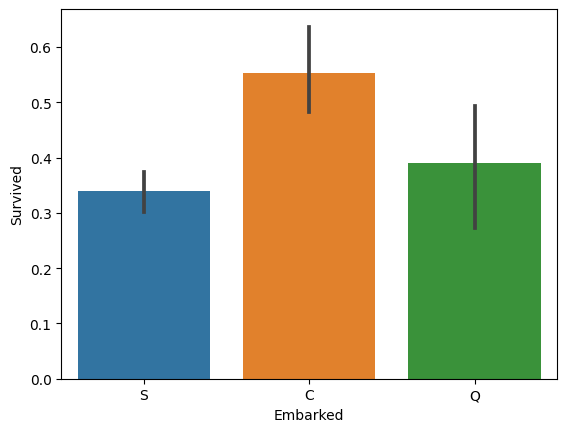


**5. OUTPUT SCREEN**

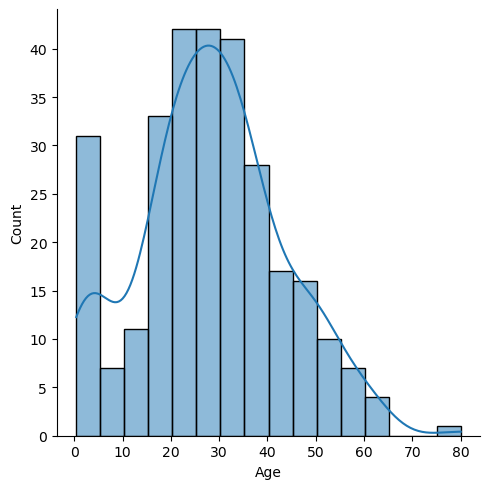
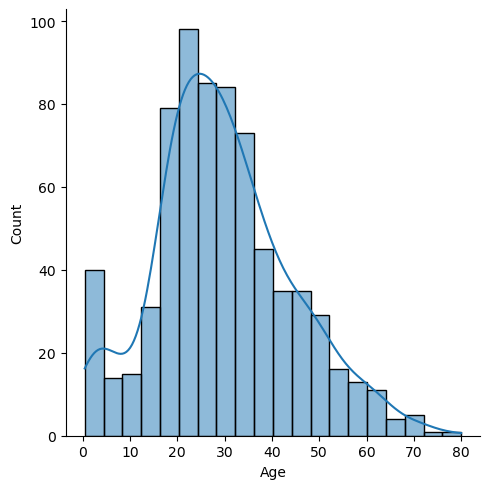
**SEX:**



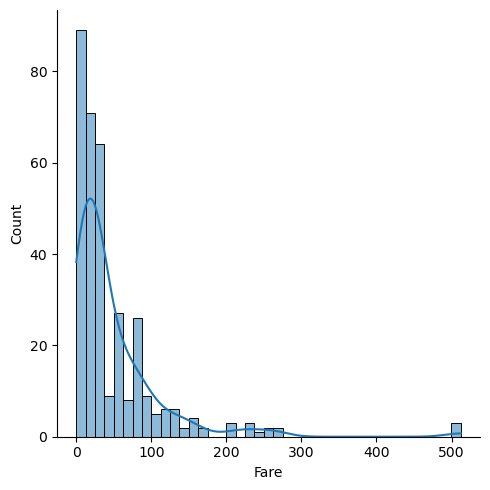
**EMBARKED**:

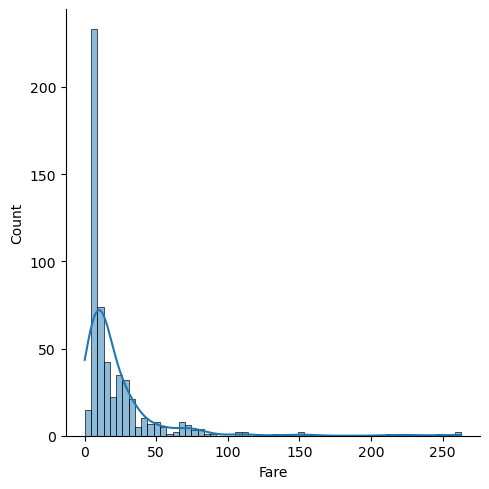


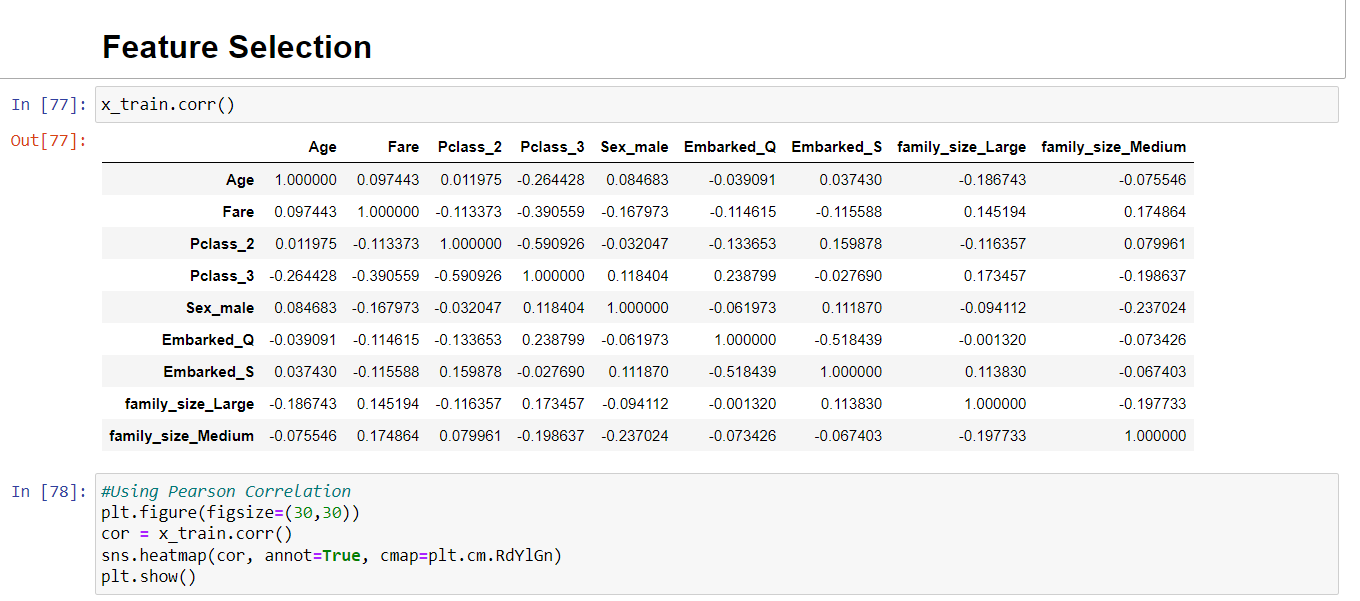
**AGE:**

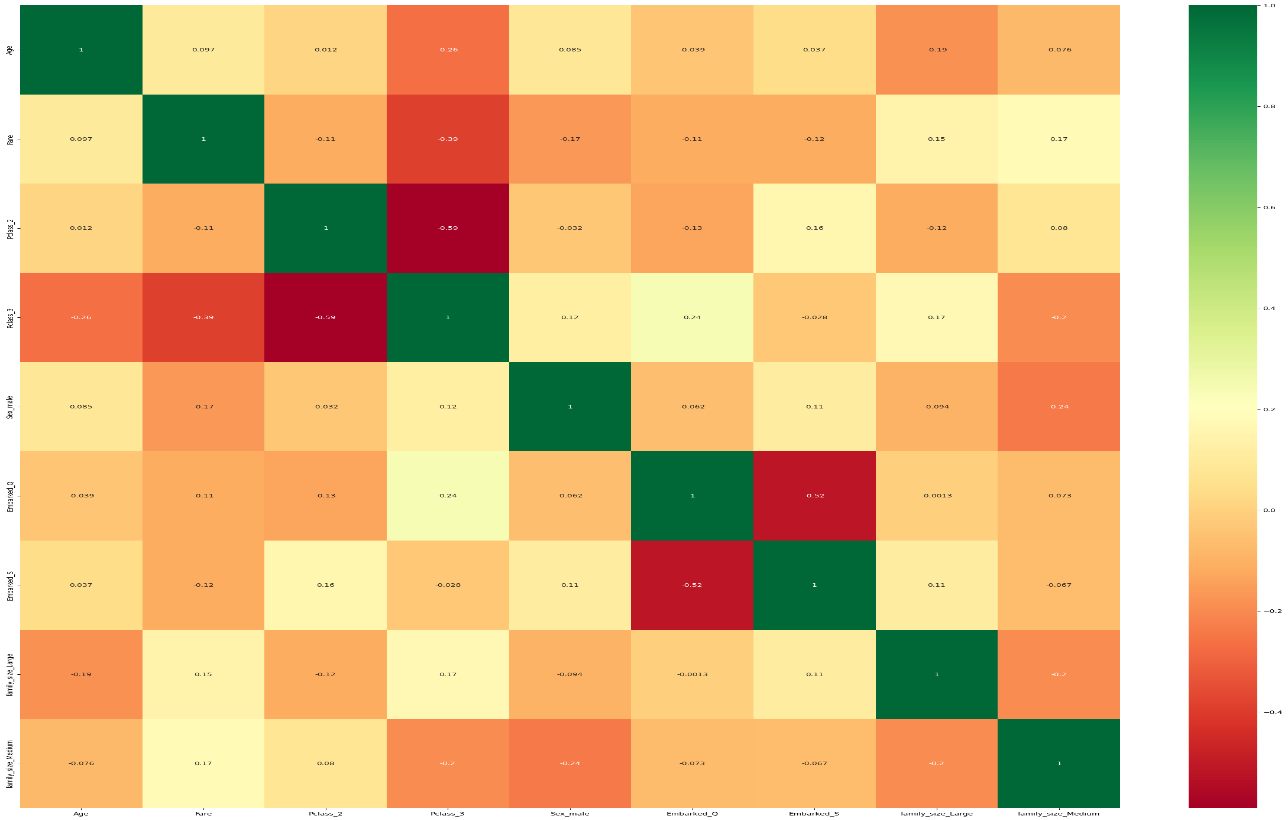


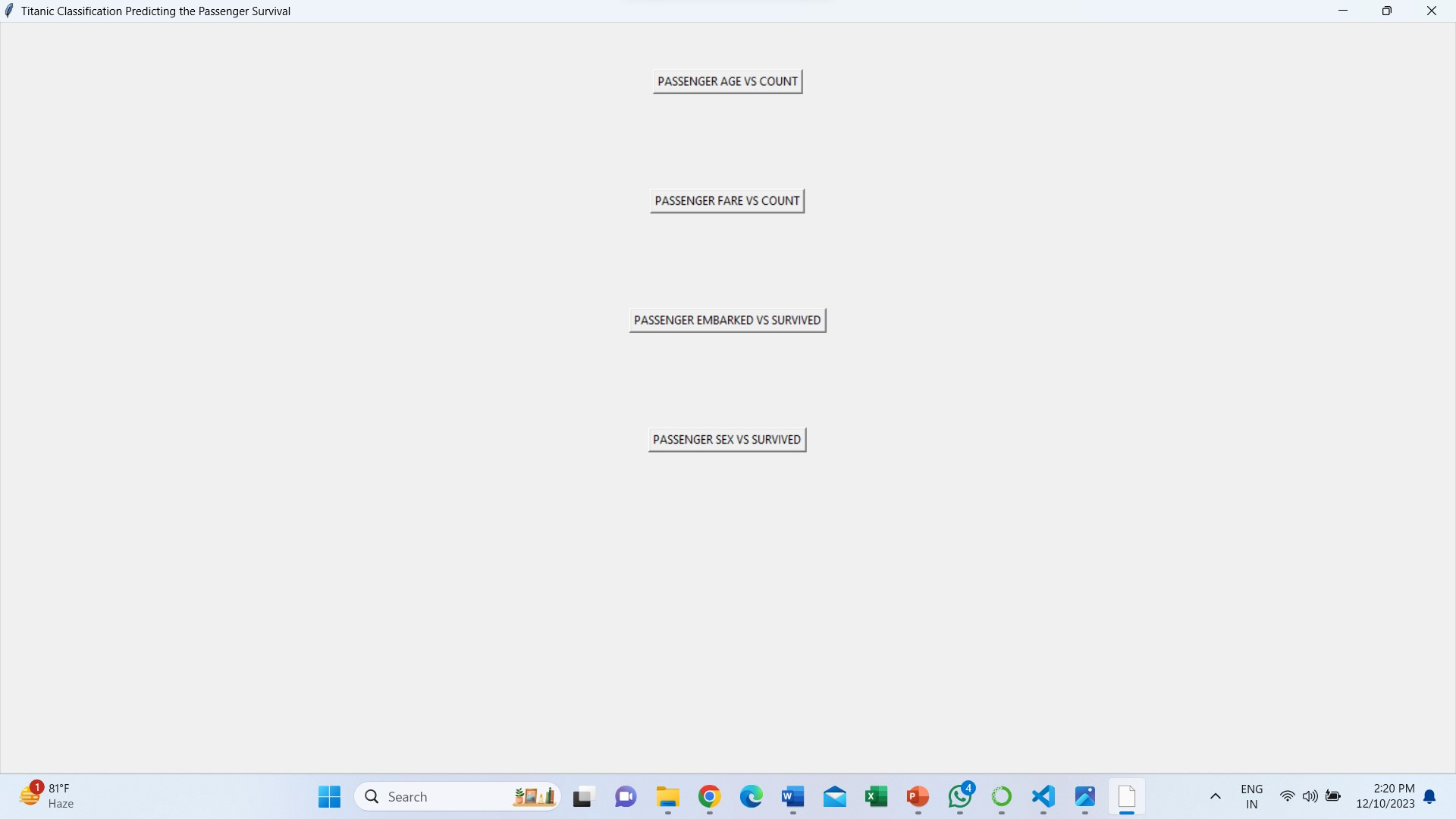
**TICKET FARE:**

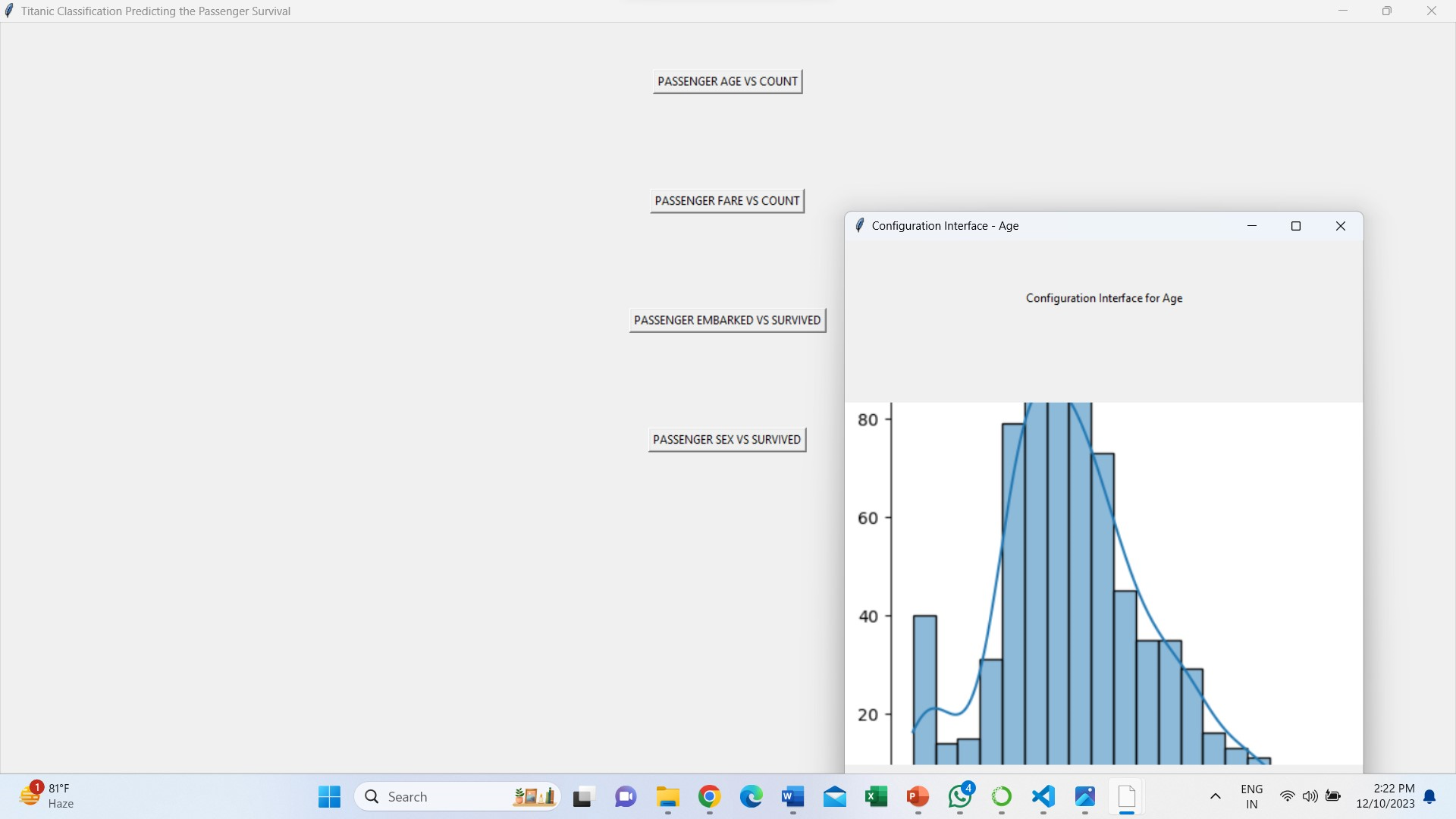




**FEATURE SELECTION(correlation)**:



**INTERFACE**:



**6.1Conclusion:**

### CHAPTER-6 CONCLUSION AND FUTURE SCOPE

The project entitled as **TITANIC CLASSIFICATION-PREDICTING PASSENGER SURVIVAL**is the system that deals with the issues related to the feedback process.

* + - This project is successfully implemented with all the features mentioned in system requirements specification.
    - The classifier provides appropriate information to users according to the chosen

dataset.

In this Titanic classification project, we aimed to predict the survival status of passengers based on various features such as class, gender, age.

**Conclusion:** The project provides a solid foundation for predicting Titanic passenger survival. Continuous improvement and exploration of advanced techniques offer opportunities for further enhancements in predictive capabilities.

### Future Scope

1. **Advanced Techniques:**
   1. Try advanced machine learning methods.
   2. Test ensemble models for improved predictions.
2. **Feature Enhancement:**
   1. Experiment with more features.
   2. Explore interactions between existing features.
3. **Handling Imbalanced Data:**
   1. Address uneven class distribution.
   2. Use oversampling or undersampling techniques.
4. **Cross-Validation Strategies:**
   1. Improve model assessment methods.
   2. Consider more sophisticated cross-validation techniques.

**LINKS:**

**GITHUB:**

[**https://github.com/Gajjelasanjay/Gajjelasanjay/blob/main/vertopal.com\_titanic%20classification%20(1).ipynb**](https://github.com/Gajjelasanjay/Gajjelasanjay/blob/main/vertopal.com_titanic%20classification%20(1).ipynb)

**REFERENCE VIDEOS LINK:**

[**https://www.linkedin.com/in/gajjela-sanjay-664763283/**](https://www.linkedin.com/in/gajjela-sanjay-664763283/)

**REFERENCES**

[1] Analyzing Titanic disaster using machine learning

algorithms-Computing, Communication and Automation

(ICCCA), 2017 International Conference on 21

December 2017, IEEE.

[2] Eric Lam, Chongxuan Tang, "Titanic Machine Learning

From Disaster", LamTang-Titanic Machine Learning

From Disaster, 2012.

[3] S. Cicoria, J. Sherlock, M. Muniswamaiah, L. Clarke,

"Classification of Titanic Passenger Data and Chances of

Surviving the Disaster", Proceedings of Student-Faculty

Research Day CSIS, pp. 1-6, May 2014.

[4] Corinna Cortes, Vlasdimir Vapnik, “Support-vector

networks”, Machine Learning, Volume 20, Issue 3,pp

273-297.

[5] L Breman- “random forests”, Machine Learning, 2001

Ng. CS229 Notes, Standford University, 2012.

[6] SJ Russsel P Norvig-“Artificial intelligence: A modern

approach”-2016.

[7] Lonnie Stevans, David L. Gleicher, ”Who Survived the

Titanic? A logistic regression analysis”-Article in

International Journal of Maritime History, December

2004. International Journal of Computer Applications (0975 – 8887)

Volume 179 – No.44, May 2018

[8] MICHAEL AARON WHITLEY, Using statistical

learning to predict survival of passengers on the RMS

Titanic by Michael Aaron Whitley, 2015.

[9] Kunal Vyas, Zeshi Zheng, Lin Li, Titanic- Machine

Learning From Disaster- 2015.

[10] EECS 349 Titanic- Machine Learning From Disaster,

Xiaodong Yang, Northwestern University.

[11] Prediction of Survivors in Titanic Dataset: A

Comparitive Study using Machine Learning Algorithms,

Tryambak Chatterlee, IJERMT-2017.

[12] An Introduction to Logistic Regression Analysis and

Reporting by Chao-Yig Joanne Peng, Kuk Lida Lee &

Gary M. Ingersoll, April 2010.

[13] Zhenyan Liu, Yifei Zeng, Yida Yan, Pengfei Zhang and

Yong Wang, Machine Learning for Analyzing Malware,

Journal of Cyber Security and Mobility, Vol: 6 Issue:

3, July 2017.

[14] Andy Liaw and Metthew Wiener, Classification and

Regression by Random Forest, vol. 2/3, December 2002.

[15] Galit Shmueli and Otto R. Koppius MIS Quarterly,

Predictive Analytics in Information System Research, ,

Vol. 35, No. 3(September 2011), pp. 553-572.

THE END..