First 2 minutes (approx.) of your presentation -- Your Motivation

You MUST mention which dataset(s) you used, and what EXACTLY is your problem definition

This is your chance to explain your MOTIVATION. Don't get carried away; be precise and clear

**Hello everyone, my name is Mao hen and my team consists of me, Yixiang and Zhi Xuan.Today, we are honored to share with you the journey of our project that delves into one of history's most tragic events—the sinking of the Titanic.**

**Our journey begins with a problem statement: to find out what is the best prediction model and the best predictor to predict the survival rate of the passengers aboard this ill-fated vessel.**

**The dataset we've utilized for this project is from**

**kaggle and it contains invaluable information about the passengers, including crucial details like age, gender, socio-economic status, and fare.**

**But why focus on the Titanic?**

**Beyond its historical significance, the Titanic tragedy serves as a reminder of the inherent vulnerabilities within our society and transportation systems.**

**Imagine being aboard that massive ship on that fateful night.**

**The lack of lifeboats left many passengers and crew members with no recourse but to face the icy waters of the North Atlantic**

**. However, amidst the chaos and uncertainty, certain groups of individuals**

**—women, children, and those from the upper class—stood a greater chance of survival.**

**Our motivation lies in uncovering these patterns and leveraging them to develop a predictive model that could potentially save lives in future maritime disasters.**

**The lessons learned from the Titanic tragedy serve as a sobering reminder of our collective responsibility to prioritize safety and preparedness in all aspects of life.**

**In conclusion, our project is not just about predicting survival rates; it's about ensuring that their sacrifices serve as a reminder for positive change.**

Next 3 minutes (approx.) of your presentation -- Set The Stage

Present your Exploratory Data Analysis and some initial data-driven Insights from the dataset.You MAY also mention how you are planning to set up the Analysis / ML problem for this case. You MUST mention how you collected / curated / cleaned / prepared the data for this problem. Did you only use tools and techniques learned in this course? What ELSE did you learn / try?

**To begin our exploratory data analysis, we first had to handle our missing values. In our case, Cabin and Age had the most missing values with cabin having almost 600+ missing values. For cabin, we simply removed it from the list of columns due to the lack of data. However, for age we decided to use the mean age to sub into the missing values as it would affect our prediction outcome the least.**

**Next would be the converting of string values to categorical values. For these, we had to convert them to numerical integers in order to actually compare with our survived data, as such we converted males and females to 1 and 0 in order to allow for data comparisons. A similar conversion was also done on “embarked”**

**Following that would be categorizing Age and Fare, we had to convert our data into categorical data in order for it to be compared to a categorical data ‘of survived’ To do this, we had to program it such that price ranges were separated into groups, thus providing us with the categorical data needed. The same was done to Fare as well.**

**Lastly would be Data Cleaning. For this, we had to take out variables that were evidently not going to predict the survival rate due to the lack of relation of the variable For example. Names and PassengerID would clearly not play an impact on the survival of a person for the majority of the cases. Along with that, we decided to remove ticket due to its close relation to Fare and the variance in data collected as many of the ticket numbers were random letters and numbers.**

**With that, we were able to begin our data analysis. With all our predictor variables being converted to categorical data, we were able to use bar plots to visualize all the possible relations that each variable might have to our prediction variable ‘survived’. Along with that we also used a confusion matrix to see the correlation of each individual predictor to the ‘Survived’ variable.**

**Our analysis allowed us to conclude that Sex, Fare and Pclass had the most impact on determining our prediction variable. However, there were some outliers that were spotted, such as Age, where only those in the lowest age range, had a higher survival rate than death rate. Whereas other age ranges had a significantly higher chance of death than survival.**

Next 3 minutes (approx.) of your presentation -- Core Analysis

If you used ML (regression, classification, or something else); mention mainly WHICH one(s).

You may now briefly CLARIFY why and how the ML problem(s) aim(s) to solve your objective.

How did you apply ML technique(s) to SOLVE your problem? Which model(s), how and why?

Did you only use tools and techniques learned in this course? What ELSE did you learn / try?

**Before we dive into implementing the algorithms, we first split the train dataset to test which algorithm will give us the best accuracy with the most number of correct predictions. We first created a function that uses the model input to predict the train and test data and produce the accuracy score for each of the data with the confusion matrix for visualization.**

**We have researched the Machine Learning algorithms that can be used for this dataset and problem. We found out that just using the algorithms that we learn in the course is not enough or does not provide the best accuracy. Hence, we picked out 7 algorithms which include logistic regression, k-nearest neighbors, support vector machines, naive bayes classifier, decision tree, random forest and perceptron, and compared the accuracy of these algorithms. We used the confusion matrix that we learnt in this course on each of these algorithms to see how many predictions were correct. In this slide, these are all the confusion matrices of each algorithm.**

**In this slide, we consolidated all the accuracy score for each algorithm and we can see that the random forest has the highest accuracy for the test data.**

**We then implemented the random forest classifier on the three best predictors, Pclass,sex and fare. After comparing each of these predictors using the best algorithm, we can tell from the confusion matrix that sex is the best predictor to predict the survival rate of passengers in titanic.**

Last 2 minutes (approx.) of your presentation -- Finish Strong

What is the OUTCOME of your project? Did it solve your original problem? Anything interesting?

What are your data-driven INSIGHTS and recommendations / views towards the target problem?

**Answering the problem statement mentioned at the start, the best prediction model and best predictor to predict the survival rate of the passengers is using the random forest classifier algorithm on predictor variable sex that gives the highest accuracy.**

**After finishing the project, we realized that data cleaning was the most important section of the project as it greatly helped with data visualization. For example, changing the age and fare to bands is crucial in helping visualize the data.**

**The last step of our project was to implement the model of random forest classifier using the sex predictor variable to predict the actual test data.**

**Based on our findings, we can conclude that in such disastrous situations, the safety and well-being of females and children are prioritized. And with that, we have come to the end of our presentation. Thank you.**