
# Titanic Analysis

## Data Mining

**Business Analytics**

**Ziad Mohamed Ahmed**
**20221373065**

**Ramy Ahmed**
**20221445844**

**Kareem Amr**
**20221372918**

**Muaz Sulaiman**
**20221310296**

**Mohamed Youssef**
**20221441395**

**Menna Metwally**
**20221310226**

# Introduction:

"**Data is the new oil**" that was said by **Clive Humby**, the famous British mathematician and entrepreneur in the field of data science and customer-centric business strategies. In a world where data is taking over the universe. It is highly important to think in an analytical way. Curiosity, critical thinking, questions, and passion all to find answers.

We will present our deep analysis in a fun scenario. Today, I'm **Jack Dawson**, representative for a big insurance company covering boat accidents costs. I met **Thomas Andrews**, Irish shipbuilder who was best known for designing the luxury liners Olympic and Titanic. Who then have asked "**How much could I pay to cover any loss through my Titanic ship journey**?". Now that is a hard question to answer, but with science it's not that hard. We used Data Science and Statistics methodology for finding the right answers.


# Ask Phase:

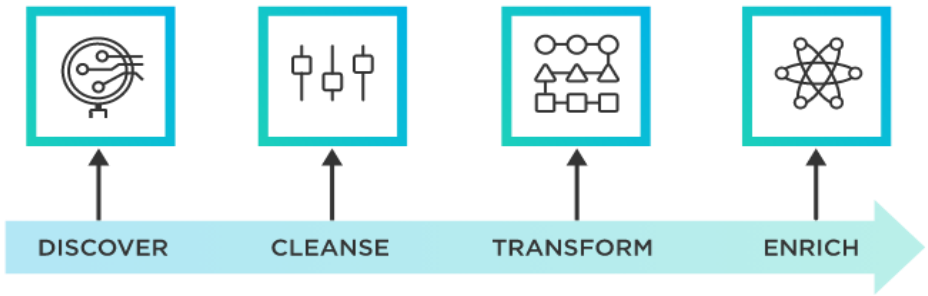
- What are the factors that could affect the survival rates?
- Can your companion affect your survival rate?
- Does gender affect your survival rate?
- What is the most age category that did not stand out in the Titanic tragedy?
- Can we build a model to predict the survival rate to help insurance companies take the right amount of money to cover costs?

**Hypothetical claims to check:**

- Is it true that women and children were given higher priority?
- Is it true that priority was given to first and second-class passengers over third-class passengers?

# Prepare Phase:

"*One could only learn from the past and move on through the present to make a better future.*"




We also had to use some tools to help us arrive to our destination:

- Python, to help manipulate the data, predict, and apply statistics.
- Power BI, to view an interactive dashboard and make sure the Average Model works.

### 1) Importing needed packages

- Pandas: Provides fast and flexible data structures for easy manipulation and analysis of data.
- NumPy: Provides powerful arrays and matrices for numerical computations and data manipulation.
- Matplotlib: A comprehensive library for creating visualizations, plots, and graphs.
- SKlearn: A machine learning library with various tools and algorithms for classification, regression, clustering, and more

### 2) Uploading Dataset

- We uploaded the dataset using the read_xl function.

# Process Phase:

"***No data is clean, but most is useful***" Dean Abbott. Do people die more by sharks or coconut? The typical answer would be "sharks." Truth is people do die more from coconuts falling on their head. *"Falling coconuts kill 150 people worldwide each year, 15 times the number of fatalities attributable to sharks," said George Burgess, Director of the University of Florida's International Shark Attack File and a noted shark researcher. Reference:* *https://www.snopes.com/fact-check/coconuts-kill-more-sharks/*. That's what bias is, a disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. That's why without clean data, you are going to get biased answers.

### 1) Exploring the Dataset

- There are 12 attributes [PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked]

| Quantitative | Qualitative |
|---|---|
| Survived | Pclass |
| Sibsp | Parch |
| Fare | Pass ID |
| Name | Sex |
| Ticket | Cabin |

- There are 891 records.
- No duplicated records
- There are 3 data types [**float64(2), int64(5), object (5)**]
- Age contains 177 null values, Cabin contains 687, while Embarked only 2.

### 2) Filling the nulls

- Age: We got the average age per each gender and class, so each gender got 3 categories with 3 means getting total of 6 means. Female [**34.61 28.72 21.75**] Male [**41.28 30.74 26.50**]
- Embarked: We got the first mode which was '**S'**

### 3) Removing Irrelevant Attributes

- Due to the huge number of nulls we decided to remove the attribute as it is not necessary for our analysis.
- Ticket ID wasn't relevant to any analysis

### 4) Checking Noisy

- We used boxplots to identify the outliers.
- Fare attribute had 3 outlier value of 512 while the max is 263

## Analyze Phase:

*"What is not defined, cannot be measured. What is not measured, cannot be improved. What is not improved, will always degrade*." William Thomson Kelvin. After cleaning and exploring our data, it's time to define our numbers, so we can measure them, in order to improve them, and to not degrade it.

### 1) Correlation

- We didn't find any correlations more than >0.8 or less than -0.8<

### 2) Enrichment

- We categorized Age to be 8 categories ['00-04', '05-14', '15-24', '25-34', '35-44', '45-54', '55-64', '65+'] this will help in encoding.
- We created an attribute called FamilyCount consistent of SibSp + Parch + 1
- We created an attribute called Gathering that has 4 values [Alone, SibSp & ParCh , SibSp , ParCh]
- We created an attribute called FareCategory that has 8 values ['00-05', '05-10', '10-20', '20-30', '30-40', '40-50', '50-60','60+'] this will help in encoding.

### 3) Assumptions

- Female survival rate is higher than men.
- Children survival rate are higher than adults.
- Class 1 survival rates are higher than the others.
- Alone travelers' survival rate is lower than the others.

These values will be proved in the Share Phase later

### 4) Organizing the Dataset for the 'Average Model'

- We made a copy from the clean dataset so we can start preparing that model to be then viewed on power BI.
- After some analysis, we found it's useful to predict survival rates based on these four parameters: Age, Pclass, Sex, Gathering categories.
- We then gave a score to each category based on its average survival rate from 0 to 1.
- We calculated the Total Score for each row and then started to categorize them to Low, Mid, and High.

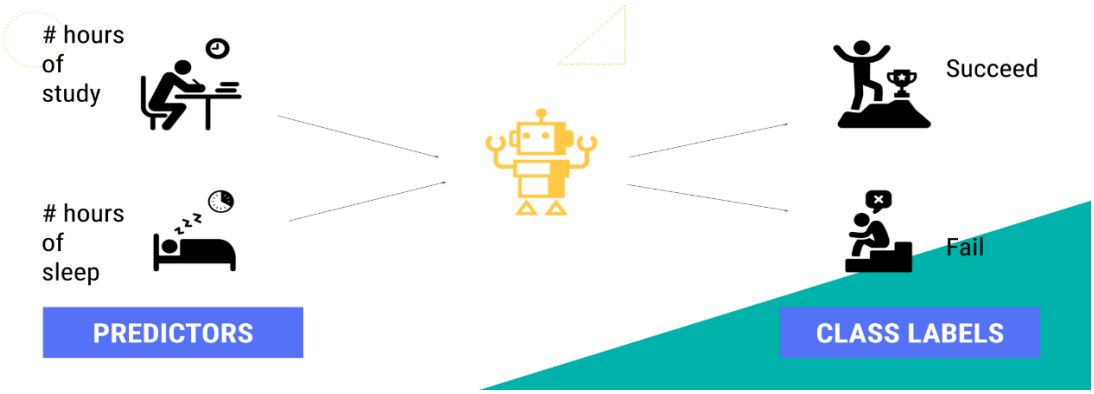
### 5) Organizing the Dataset for the Models

- We got a copy from this data to calculate the K-means, because K-means won't work with encoded data.
- All the models need encoding so we encoded [Sex, Embarked, AgeCategory, Gathering, FareCategory] and we removed the original attributes.

### 6) Preparing for Models

- We need to separate the data into Features Variable and Target Variable
- We separated the data into training and testing with 80% 20% division.

### 7) Models




- KNN: A classification algorithm that uses the closest labeled data points to make predictions.
- K-means: A clustering algorithm that separates data into distinct groups based on their similarity to each other.
- Decision Tree: A predictive modeling approach that builds a tree-like model of decisions and their possible consequences.
- Naive Bayes: A probabilistic algorithm that uses Bayes' theorem to make predictions by assuming independence between the features.

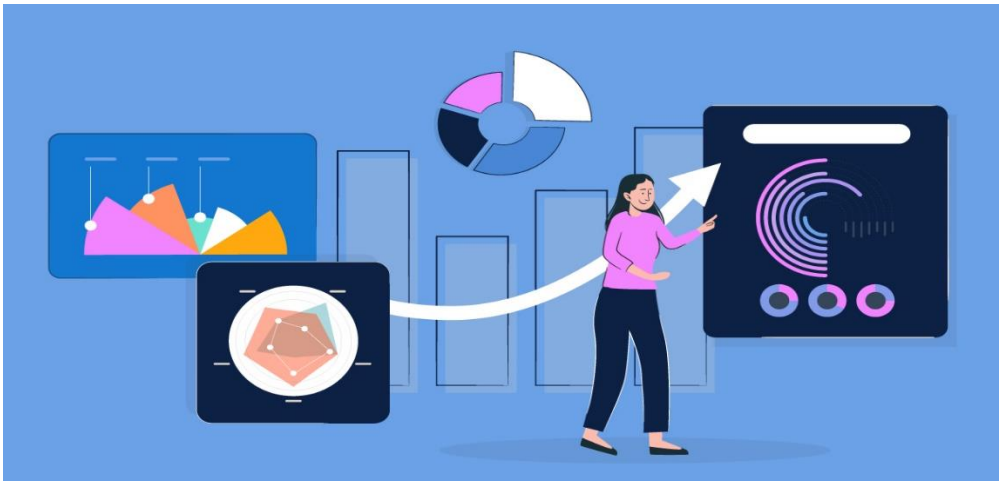
| KNN | Decision Tree | Naïve Bayes |
|---|---|---|
| 0.78 | 0.76 | 0.80 |

**K-means**

- We took 'Age','Fare','Pclass' attributes as our features.
- From elbow graph we concluded that clusters of 3 is the ideal value
- We added a new column called cluster to add the cluster of each record.

# Share Phase:

*The greatest value of a picture is when it forces us to notice what we never expected to see*



*To provide the best of the best, this section will be represented in power BI*