AFAME TECHNOLOGIES

Technology Advertising Consulting

PROJECT SUBMITTED BY

SARVADAMAN DILIP MASKE

A REPORT SUBMITTED TO AFAME TECHNOLOGIES
TOWARDS FULFILLMENT OF THE

DATA ANALYST INTERNSHIP

SUBMITTED BY

SARVADAMAN DILIP MASKE

UNDER THE GUIDANCE OF

Project Team

Afame Technologies - Bengaluru, India

INDEX

PROJECT 1

SALES DATA ANALYSIS

Goal:

Use sales data analysis to find patterns, best-selling items, and revenue indicators to help in business decision-making.

This project will require you to delve into a sizable sales dataset in order to glean insightful information. In order

to successfully convey your findings, you will compute revenue measures like

1. total sales,

2. analyze sales trends over time,

3. determine the best-selling products, and build visualizations.

Description of data

description of each column:

1. **Row ID**: A unique identifier for each row in the dataset.
2. **Order ID**: A unique identifier for each order.
3. **Order Date**: The date when the order was placed.

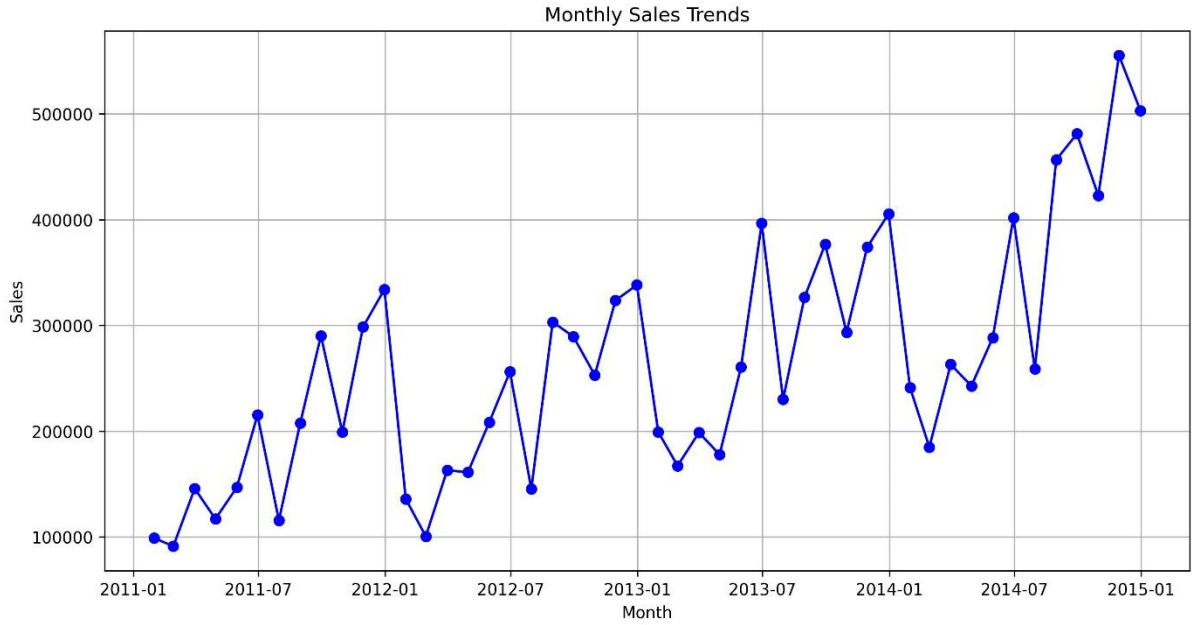
4. **Ship Date**: The date when the order was shipped.
5. **Ship Mode**: The mode of shipping (e.g., Same Day, Second Class).
6. **Customer ID**: A unique identifier for each customer.
7. **Customer Name**: The name of the customer.
8. **Segment**: The market segment to which the customer belongs (e.g., Consumer, Corporate).
9. **City**: The city where the customer is located.
10. **State**: The state where the customer is located.
11. **Country**: The country where the customer is located.
12. **Postal Code**: The postal code of the customer's location.
13. **Market**: The market region (e.g., US, APAC).
14. **Region**: The specific region within the market.
15. **Product ID**: A unique identifier for each product.
16. **Category**: The category of the product (e.g., Technology, Furniture).
17. **Sub-Category**: The sub-category of the product (e.g., Accessories, Chairs).
18. **Product Name**: The name of the product.
19. **Sales**: The sales amount for the order.
20. **Quantity**: The quantity of the product ordered.
21. **Discount**: The discount applied to the order.
22. **Profit**: The profit made from the order.
23. **Shipping Cost**: The cost of shipping the order.
24. **Order Priority**: The priority level of the order (e.g., Critical).

METHODOLOGY:

1. To gain an initial understanding of the overall sales performance, the total sales amount is calculated. This involves summing up all the values in the 'Sales' column of the dataset.

we get total sales: 1,26,42,502.

2.



To understand the sales performance over time, we created a line graph titled "Monthly Sales Trends". This graph displays the sales data from January 2011 to January 2015.

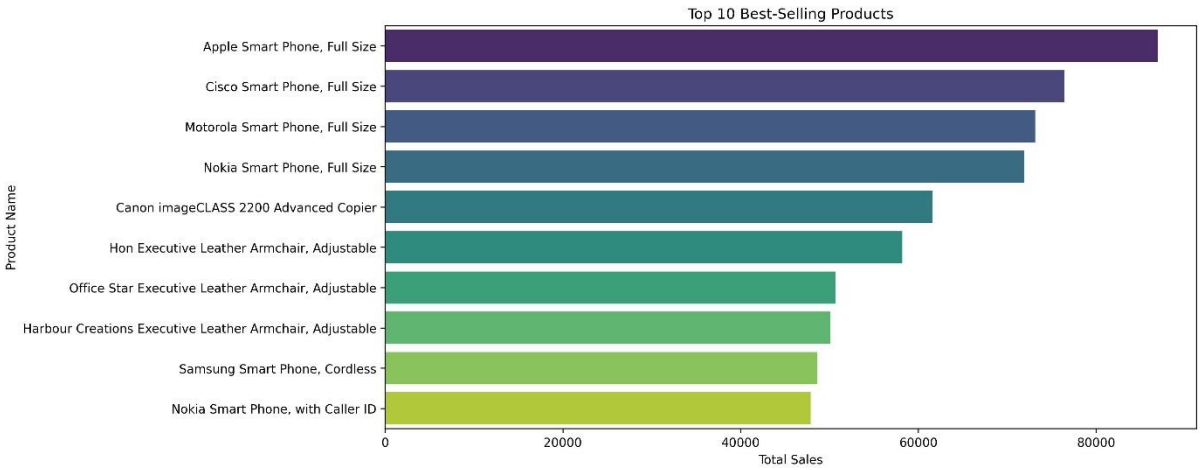
Horizontal Axis (X-axis): Represents the months from January 2011 to January 2015.

Vertical Axis (Y-axis): Represents the sales in dollars.

The blue line with markers on the graph shows the sales for each month. This visualization helps us identify patterns, peaks, and troughs in sales over the four-year period. By analyzing these trends, we can gain insights into the seasonal variations and overall performance of our sales.

This graph is essential for understanding how sales have fluctuated over time and can be used to make informed business decisions.

3.

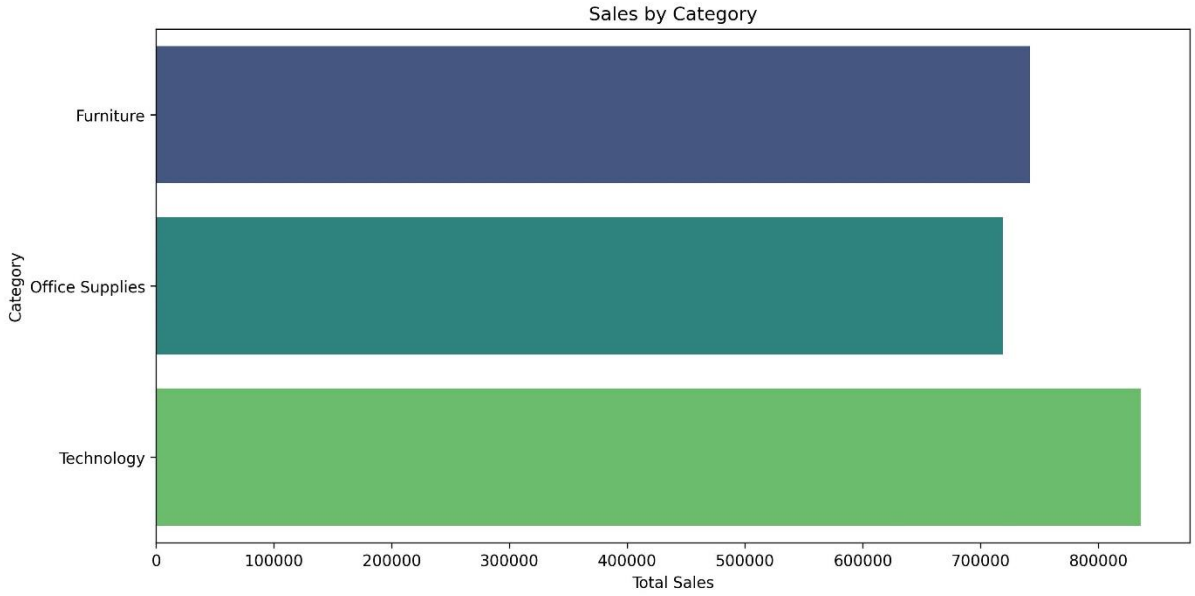


To analyze the popularity of different products, we created a horizontal bar chart titled “Top 10 Best-Selling Products”. This chart displays the top 10 products based on their total sales.

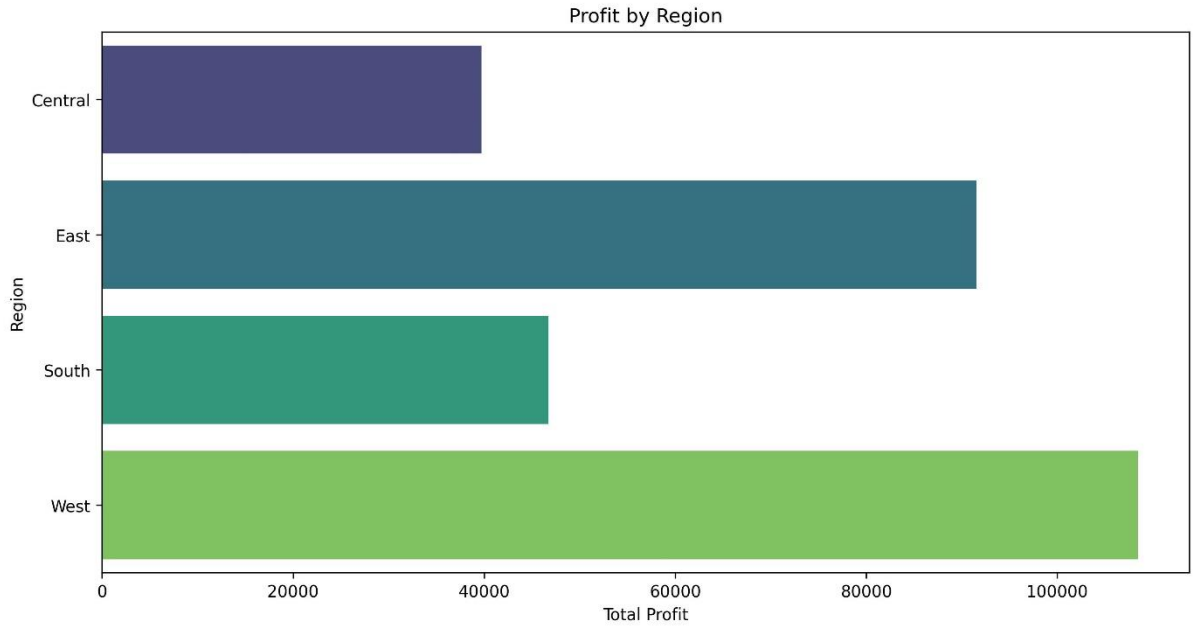
Apple Smart phone, Cisco smart phone and Motorola most sold products

Each bar represents a product, with the length of the bar indicating the sales amount. The longest bar shows the product with the highest sales, allowing for a quick comparison of product popularity. This visualization helps us understand which products are the most popular and generate the most revenue.

This chart is essential for identifying key products that drive sales and can inform inventory and marketing strategies.



From above Horizontal bar chart we can see that, Technology has highest sales followed by Furniture.



From above Horizontal bar chart we can see that, highest profit is getting to west region followed by east region while central region has lowest profit.

PROJECT 2

HR DATA ANALYSIS

GOAL:

Activities to complete:

Data cleansing involves removing unnecessary columns.

Giving the columns new names.

Eliminating redundant entries.

sanitizing specific columns.

Eliminate the dataset's NaN values.

DATA DESCRIPTION:

Age: The age of the employee.

Attrition: Indicates whether the employee has left the company (Yes/No).

Business Travel: Frequency of business travel (e.g., Travel_Rarely, Travel Frequently, Non-Travel).

Daily Rate: The daily rate of the employee.

Department: The department where the employee works (e.g., Sales, Research & Development, Human Resources).

Distance From Home: The distance between the employee's home and workplace.

Education: The education level of the employee (e.g., 1 = 'Below College', 2 = 'College', 3 = 'Bachelor', 4 = 'Master', 5 = 'Doctor').

Education Field: The field of education of the employee (e.g., Life Sciences, Medical, Marketing).

Employee Count: The count of employees (usually 1 for each row).

Employee Number: A unique identifier for each employee.

Environment Satisfaction: Satisfaction with the work environment (1 = Low, 2 = Medium, 3 = High, 4 = Very High).

Gender: The gender of the employee (e.g., Male, Female).

Hourly Rate: The hourly wage of the employee.

Job Involvement: The level of involvement in the job (1 = Low, 2 = Medium, 3 = High, 4 = Very High).

Job Level: The level of the job within the organization.

Job Role: The role of the employee (e.g., Sales Executive, Research Scientist).

Job Satisfaction: Satisfaction with the job (1 = Low, 2 = Medium, 3 = High, 4 = Very High).

Marital Status: The marital status of the employee (e.g., Single, Married, Divorced).

Monthly Income: The monthly income of the employee.

Monthly Rate: The monthly rate of the employee.

Num Companies Worked: The number of companies the employee has worked for.

Over18: Indicates if the employee is over 18 years old (Y/N).

Over Time: Indicates if the employee works overtime (Yes/No).

Percent Salary Hike: The percentage increase in salary.

Performance Rating: The performance rating of the employee (1 = Low, 2 = Good, 3 = Excellent, 4 = Outstanding).

Relationship Satisfaction: Satisfaction with relationships at work (1 = Low, 2 = Medium, 3 = High, 4 = Very High).

Standard Hours: The standard number of working hours (usually 80).

Stock Option Level: The level of stock options granted to the employee.

Total Working Years: The total number of years the employee has worked.

Training Times Last Year: The number of training sessions attended by the employee last year.

Work Life Balance: The work-life balance rating (1 = Bad, 2 = Good, 3 = Better, 4 = Best).

Years At Company: The number of years the employee has been with the company.

Years In Current Role: The number of years the employee has been in the current role.

Years Since Last Promotion: The number of years since the employee's last promotion.

Years With Curr Manager: The number of years the employee has worked with the current manager.

METHODOLOGY:

Removing Unnecessary Columns:

Identify the columns you want to drop. These might be columns that don't contribute significantly to your analysis or have too many missing values. Use the drop method to remove those columns.

Renaming Columns:

If you want to rename specific columns, use the rename method.

Eliminating Redundant Entries:

To remove duplicate rows, use the drop_duplicates method.

Sanitizing Specific Columns:

Depending on what you mean by "sanitizing," you might need to handle outliers, incorrect data, or other issues. For outlier removal, consider using z-scores or interquartile range (IQR) method.

Handling Missing Values (NaNs):

To drop rows with any NaN values, use the dropna method Alternatively, you can impute missing values using methods like mean, median, or forward/backward fill.

PROJECT 3

TITANIC SURVIVAL PREDICTION

GOAL

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank

after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone on board,

resulting in the death of 1502 out of 2224 passengers and crew.

Use the Titanic dataset to build a model that predicts whether a passenger on the Titanic survived or

not. This is a classic beginner project with readily available data.

The dataset typically used for this project contains information about individual passengers, such as

their age, gender, ticket class, fare, cabin, and whether or not they survived.

DATA DESCRIPTION:

Passenger Id: Unique ID for each passenger.

Survived: Survival status (0 = No, 1 = Yes).

P class: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd).

Name: Name of the passenger.

Sex: Gender of the passenger.

Age: Age of the passenger.

Sib Sp: Number of siblings/spouses aboard the Titanic.

Parch: Number of parents/children aboard the Titanic.

Ticket: Ticket number.

Fare: Passenger fare.

Cabin: Cabin number.

Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

METHODOLOGY:

The data we collected likely has errors, missing numbers, and corrupted values because it's still in its raw form. Before we can draw any conclusions, we need to clean and organize the data, a process known as data wrangling. This makes large, complex data sets easier to access and analyze. Feature engineering is another important step, where we create more relevant features from the raw data to improve the predictive power of our learning algorithms.

Our process to solve the problem starts with collecting the raw data. Next, we import the dataset into our working environment and perform data preprocessing, which includes data wrangling and feature engineering. After that, we explore the data and build a model using machine learning algorithms. We then assess the model and repeat the process until it performs well. Finally, we compare the results within the algorithm and select the model that best fits the problem.

- **Feature Engineering** is the most critical stage of data analytics. It involves creating predictions and selecting the features to use in training. When developing a machine learning model, domain expertise is used to identify useful features in the dataset. This helps in understanding the dataset and improving the model's accuracy. Poor feature selection can lead to a poor prediction model, so choosing the right features is crucial for accuracy and predictive power. Unnecessary or irrelevant features are filtered out. Based on exploratory analysis, the following features are used: age, sex, cabin, title, class, family size (parch plus sibsp columns), fare, and embarked. The target column is determined by the survival column. These features were chosen because they have values that affect the survival rate. In bar plots, these features will be on the x-axis. Even a smart algorithm can give inaccurate predictions if the wrong features are chosen. Therefore, feature engineering is the foundation for creating an accurate predictive model.

- **Logistic regression** is a technique that works best when the dependent variable is dichotomous (binary or categorical). It describes the data and explains the relationship between a single binary dependent variable and one or more independent variables, which can be nominal, ordinal, interval, or ratio-level.
- A **decision tree** is a type of supervised learning algorithm commonly used for classification tasks. It can handle both continuous and categorical input and output variables. Here's a simple breakdown:
  Root Node: Represents a split point on a single input variable (x) and the variable itself.
  Leaf Nodes: Contain the dependent factor (y), which is the outcome or prediction.

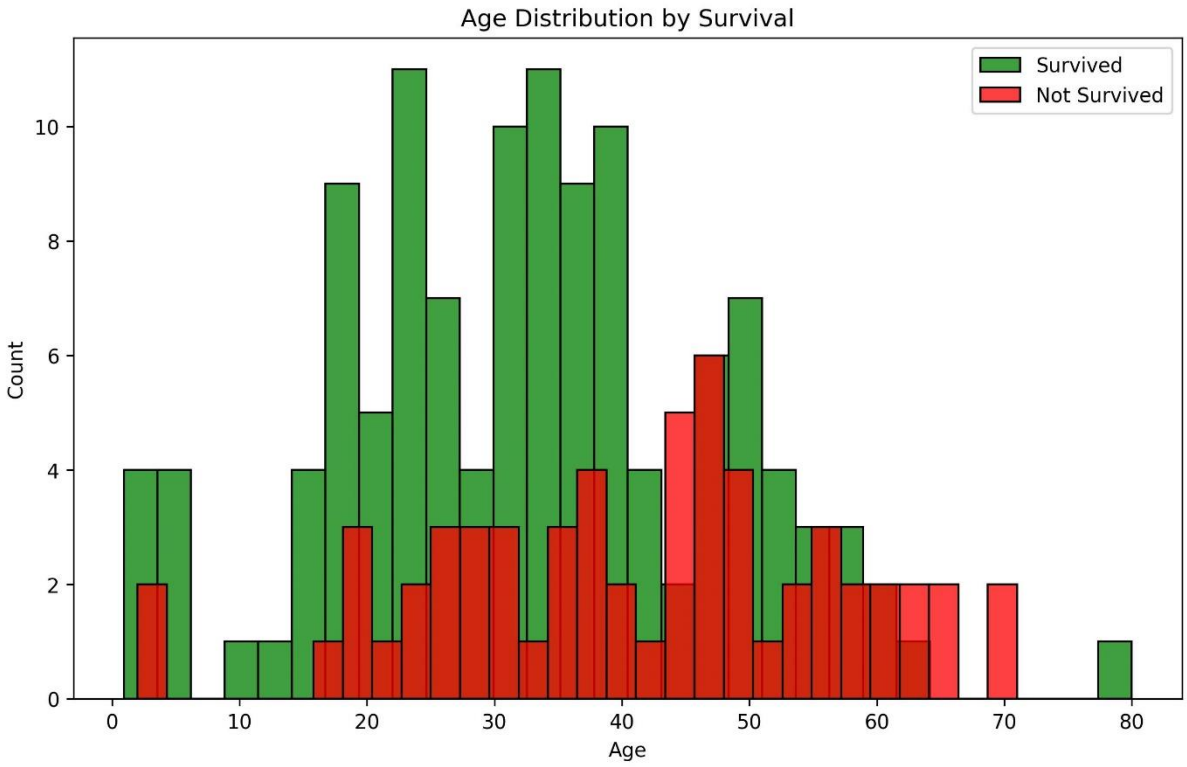
Each decision in the tree splits the data into subsets based on the value of an input variable, making it easier to classify or predict outcomes.

- **Random forest** is a supervised classification algorithm that creates a "forest" with many decision trees. The more trees in the forest, the more accurate the results. This approach can solve both classification and regression problems. The final prediction is made by averaging the predictions from all the trees (for regression) or by taking the majority vote (for classification).
- **Support Vector Machine (SVM)** is a supervised machine learning algorithm that can handle both classification and regression problems.

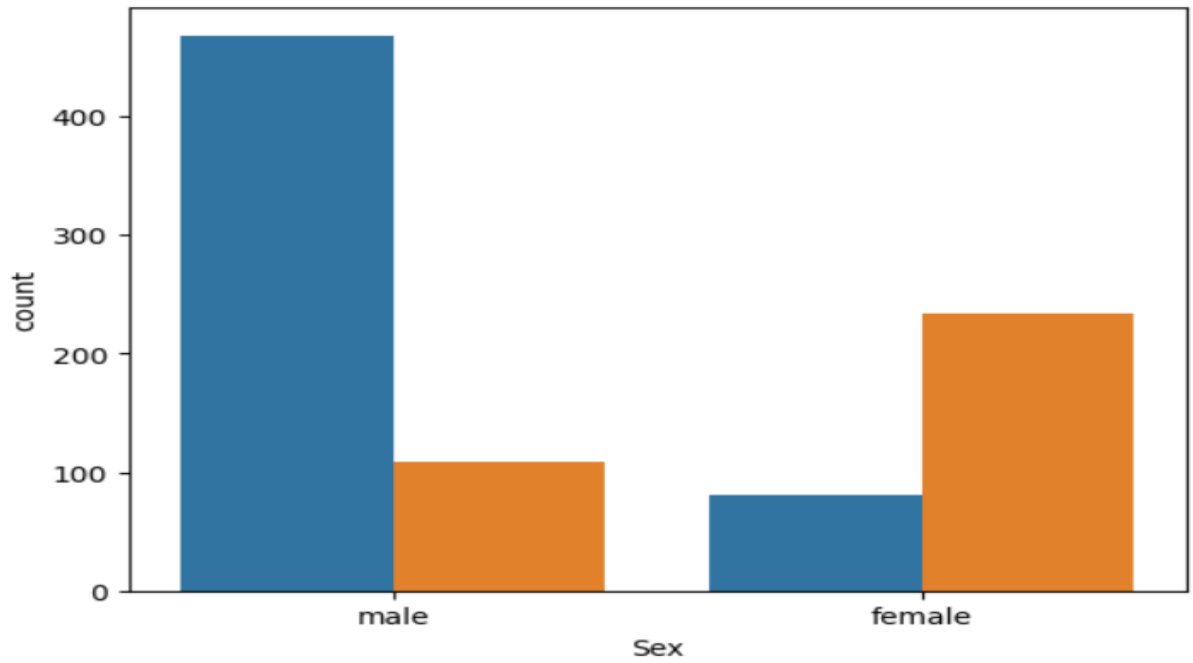
MODEL EVALUATION:

The model's accuracy is assessed using a "confusion matrix." A confusion matrix is a table design that makes it possible to see
how well an algorithm performs and is right.
It provides a measurement of the model's or algorithm's percentage of correctly predicted outcomes. "1" is the best value, while
"0" is the worst.

EXPLORATORY DATA ANALYSIS:



Most people survived who is middle aged



Brown color represent survived people. Most female survived compared to male.

RESULT:

We tested our trained algorithms against a test data set after training them, and we evaluated their performance by comparing their goodness of fit to a confusion matrix. 30% of the data and 70% of the data are used as training data sets. When compared to logistic regression (76.00%) for the given data set, the decision tree & random forest approach has a accuracy in prediction of the survival rate (73.00%). While SVM has less accuracy 62.00%.

REFERENCES:

1. www.ijcrt.org
2. www.jetir.org

3. SURVIVAL PREDICTION FOR TITANIC DATA USING MACHINE LEARING ALGORITHMS 1P. Ravindra, 2Dr. P. Vijayapal Reddy, 1Assistant Professor, 2Professor 1,2Computer Science Engineering, Matrusri Engineering College, Hyderabad, India
4. TITANIC SURVIVAL PREDICTION Kavya N C, Mr. Srinivasulu M University B.D.T College of Engineering, Davanagere, Karnataka, India

PYTHON CODES GITHUB LINK:

https://github.com/Sarvadaman-Maske/AFAME-TECHNOLOGIES-INTERNSHIP