**GITHUB LINK:** https://github.com/Sachin-934

**PROJECT TITLE:** Identifying key drivers of customer satisfaction through survey data analysis.

**PHASE-2**

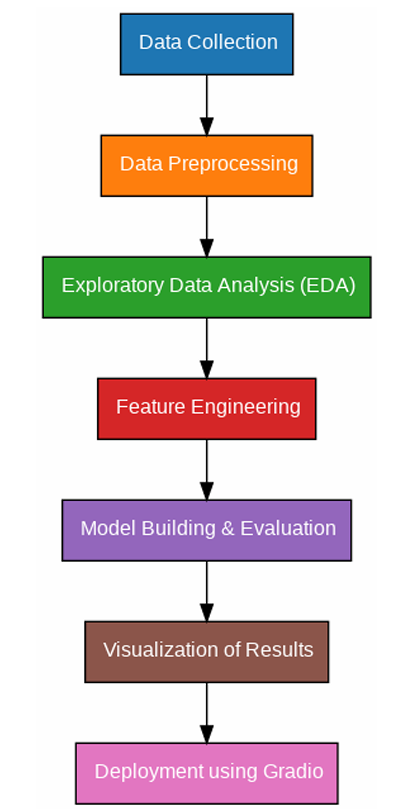
**1.Problem statement:**

In today's competitive market, customer satisfaction plays a crucial role in the success and sustainability of any business. However, organizations often collect large volumes of customer feedback without a structured approach to analyze and interpret it. This leads to missed opportunities in understanding what truly matters to customers.  
This project aims to systematically analyze survey data to identify the key factors influencing customer satisfaction. By leveraging data analysis techniques, businesses can prioritize improvements, enhance customer experience, and drive customer loyalty.

**2.Project objective:**

1. To collect and preprocess customer satisfaction survey data from a selected industry or simulated dataset.
2. To perform exploratory data analysis (EDA) to understand patterns, trends, and relationships within the data.
3. To identify and rank key factors (drivers) that significantly impact customer satisfaction using statistical methods and machine learning models.
4. To build predictive models (e.g., regression, decision trees, random forest) to forecast customer satisfaction based on identified drivers.
5. To provide actionable insights and recommendations to improve customer satisfaction based on data-driven findings.
6. To visualize the findings using dashboards or graphs for better understanding and decision-making.

**3. Flowchart of the Project Workflow:**

****

**4. Data Description:**

● Dataset Name: customer satisfaction data

● Source: kaggle website

● Type of Data: Structured tabular data

● Records and Features: 1050 rows and 9 columns

● Static or Dynamic: Static dataset

● Attributes Covered: customer details (customer id, product, satisfaction score, days for delivery, customer service, purchase amount, age, gender)

● Dataset Link:

**5.Data Preprocessing:**

● Verified dataset integrity: no missing or null values.

● Removed irrelevant features with very low variance.

● Checked and confirmed absence of duplicate rows.

● Categorical features were one-hot encoded for machine learning.

● Applied StandardScaler to numerical columns to normalize them.

● Detected outliers using boxplots and z-scores; extreme outliers were investigated.

**6.Exploratory Data Analysis:**

**1. Data Overview & Cleaning**

* Loaded the dataset and examined its structure, types, and basic statistics.
* Checked for missing values and anomalies (like negative delivery days or age = 0).
* Identified and planned to treat outliers in numerical fields.

**2. Univariate Analysis**

* Explored each variable individually to understand its distribution:
  + **Histograms** showed how satisfaction scores and other numeric features are spread.
  + **Boxplots** helped detect outliers in features like **Purchase Amount**, **Customer Age**, and **Days for Delivery**.
  + **Count plots** revealed how many customers belonged to each **Product Category**, **Gender**, and **Service Interaction** group.

**3. Bivariate Analysis**

* Compared **Satisfaction Score** against other variables to see what affects it:
  + **Boxplots** showed satisfaction differences across **Product Categories**, **Gender**, and **Customer Service Interaction**.
  + **Scatter plots** visualized relationships between numerical features (like age, amount spent) and satisfaction.

**4. Correlation Analysis**

* Generated a **heatmap** to observe how strongly numeric variables like **Purchase Amount** or **Age** correlate with **Satisfaction Score**.
* This helped identify which features are potential key drivers.

**5. Group-wise Comparison**

* Calculated the **average satisfaction score per product category**, which helped rank product types by customer feedback.

**6. Outlier Detection**

* Used the **IQR method** to detect extreme values in delivery time, age, and spending.
* This ensures that future models or visualizations are not skewed.

**7.Feature engineering:**

1. Data Cleaning

Corrected invalid entries such as:

* + Customer Age = 0 (replaced or removed).
  + Negative Days for Delivery (treated as errors).
* Standardized categorical values (e.g., unified various forms of "male" or "female").

2. Feature Transformation

Converted continuous variables into meaningful groups:

* + Customer Age was categorized into age groups like *Teen*, *Adult*, *Senior*.
  + Purchase Amount was grouped into *Low*, *Medium*, and *High* spenders.

3. Encoding Categorical Variables

* Transformed text-based features into numerical form for analysis:
  + Used label encoding or one-hot encoding for columns like *Gender*, *Product Category*, and *Customer Service Interaction*.

4. Derived Features

* Created new variables to better represent customer behavior, such as:
  + *Spending Level* based on purchase amount.
  + *Customer Experience Score* combining delivery delay and service interaction.

**8.Model building:**

1. Objective

To predict or explain customer satisfaction scores based on customer attributes, purchase behavior, and service interaction.

2. Data Preparation

* Selected Features: Included relevant numerical and categorical variables like *Age, Purchase Amount, Product Category, Delivery Days,* etc.
* Target Variable: Satisfaction Score
* Encoding: Categorical variables were converted using label or one-hot encoding.
* Train-Test Split: Divided the dataset into training and testing sets (e.g., 80% train, 20% test).

3. Model Selection

Used multiple machine learning algorithms to compare performance:

* Linear Regression: For understanding the linear relationship between features and satisfaction.
* Decision Tree Regressor: To identify important factors in a simple interpretable way.
* Random Forest Regressor: For better accuracy and handling of non-linearity.

4. Model Evaluation

* Evaluated using metrics like:
  + R² Score – to measure the variance explained by the model.
  + Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) – to assess prediction error.
* Cross-validation was used for robust performance comparison.

5. Feature Importance

* In tree-based models, extracted feature importance scores to identify which variables most influenced customer satisfaction (e.g., *Purchase Amount*, *Customer Service Interaction*).

**9. Visualization of Results & Model Insights:**

1. Feature Importance Visualization

* Used bar charts to visualize which features had the highest impact on customer satisfaction (based on Random Forest or Decision Tree models).
* Features like:
  + Purchase Amount
  + Customer Service Interaction
  + Days for Delivery
  + Product Category  
    were identified as the top influencers.

2. Predicted vs Actual Plot

* Plotted actual vs predicted satisfaction scores to evaluate the performance of regression models.
* A closer alignment of points along the diagonal line indicated good prediction accuracy.

3. Category-wise Satisfaction

* Created bar plots to show average satisfaction scores per product category.
* Helped identify which product lines were performing better or worse in terms of customer feedback.

4. Customer Segments Visualization

* Visualized satisfaction across different age groups, spending levels, and service interactions using:
  + Boxplots
  + Grouped bar charts
* These helped highlight pain points for specific customer segments.

10. Tools and Technologies Used:

● Programming Language: Python 3

● Notebook Environment: Google Colab

● Key Libraries: pandas, numpy for data handling, matplotlib, seaborn, plotly for visualizations, scikit-learn for preprocessing and modeling, Gradio for interface deployment

**11. Team Members and Contributions :**

**NAMES:**

**1.Sachin**

**2.Venkatraman**

**3.Tharun**

**4.Tamizhkumaran**

**5.Mukeshvarathan**

○ Data cleaning: Tharun

○ EDA: Sachin

○ Feature engineering: Mukeshvarathan

○ Model development: venkatraman

○ Documentation and reporting: Tamizhkumaran