Dataset 1

CAR PRICE PREDICTION

Problem Statement

A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:

- Which variables are significant in predicting the price of a car
- How well those variables describe the price of a car

Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.

You are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

Dataset 2

IDENTIFY CUSTOMER GROUPS FOR STRATEGIC MARKETING

Problem Statement

Golden Retail Inc. is a global retail chain that offers a wide range of products and services to millions of customers around the world. Over the years, they have accumulated a rich set of customer data through loyalty programs, purchase histories, and customer interactions. While the company has done well in terms of sales, the leadership team believes that more personalized customer engagement could unlock greater revenue potential and improve customer satisfaction.

With an ambitious goal of launching targeted marketing campaigns and personalized offers, the company wants to better understand its diverse customer base. However, with thousands of customers and a wide range of variables—such as income, spending behavior, age, tenure, and family size—it becomes increasingly difficult to manually segment or group customers in a meaningful way.

That's where you step in as a data analyst. Your mission is to help Golden Retail Inc. make sense of their customer data. You are expected to segment customers based on their characteristics and behaviors.

Your final goal is to present clearly defined customer groups that the marketing team can target with customized strategies—whether it's for premium loyalty programs, discount campaigns, or product recommendations. Your analysis will play a critical role in enabling data-driven marketing decisions for the company.

Dataset 3

MALL CUSTOMERS SEGMENTATION

Problem Statement

How can different clustering techniques be utilized to assist a supermarket in increasing their membership card conversion rate? By performing a customer segmentation analysis, you are expected to identify groups of customers with similar shopping preferences and purchasing histories. This approach allows companies to tailor their marketing strategies more effectively to each customer group.

Dataset 4

PREDICTING CUSTOMER SATISFACTION

Problem Statement

Customer satisfaction is a critical factor in the airline industry, directly influencing customer retention, brand reputation, and overall business performance. Given survey data from an airline, can we accurately predict whether a customer is satisfied or unsatisfied with their experience?

The problem will be approached as a binary classification task, where the goal is to classify customers into two categories: Satisfied and Unsatisfied based on various factors such as service quality, in-flight experience, and overall customer feedback.

By leveraging machine learning models, this analysis seeks to identify key determinants of customer satisfaction, enabling airlines to make data-driven decisions for improving their services and enhancing the passenger experience.

Dataset 5

HEART DISEASE PREDICTION

Problem Statement

Heart disease remains a leading cause of mortality worldwide, making early detection and accurate diagnosis critical for effective treatment and prevention. This study focuses on a multivariate dataset, which involves multiple statistical and mathematical variables, enabling comprehensive numerical data analysis.

The dataset consists of 14 key attributes related to patient health, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic

results, maximum heart rate achieved, exercise-induced angina, oldpeak (ST depression induced by exercise relative to rest), the slope of the peak exercise ST segment, number of major vessels, and Thalassemia. The primary objective of this study is to predict the presence of heart disease in patients based on these attributes.

Dataset 6

COFFEE QUALITY IDENTIFICATION

Problem Statement

Understanding the factors that influence coffee quality is essential for producers, consumers, and industry stakeholders. Given a dataset of Colombian coffee samples with various attributes such as origin, processing method, and sensory characteristics, clustering techniques can be applied to group similar coffee samples based on their quality metrics.

The objective is to explore these clusters to identify patterns and relationships that contribute to coffee quality. This analysis can provide valuable insights into how different factors impact coffee grading, helping stakeholders make informed decisions to enhance production and quality assessment.

Dataset 7

TAXI PRICE PREDICTION

Problem Statement

A global ride-hailing company is planning to launch operations in a new city and needs to set its dynamic pricing strategy for taxi rides. They wish to build a predictive framework that can estimate the ride-fare based on ride characteristics, enabling them to competitively price trips while maintaining profitability.

Specifically, they want to understand:

- Which variables significantly influence the fare (price) of a taxi ride.
- How well those variables explain and predict the ride fare in the target city's context.

To support this, a consulting team has compiled a large dataset of taxi trips—including features such as pickup and drop-off locations, timestamps, distance travelled, passenger count, surge indicator, and other ride attributes. You are required to build a regression model using the available independent variables to predict the ride fare. The resulting model will guide the company's pricing strategy, fare-setting decisions, operational planning and promotional offers in the new market.

Dataset 8

COUNTRY SEGMENTATION FOR FUND ALLOCATION

Problem Statement

HELP International is an international humanitarian NGO committed to fighting poverty and providing basic amenities and disaster relief to people in underdeveloped countries. The organization has successfully raised around $10 million, which now needs to be allocated strategically and effectively to the countries most in need.

To ensure that the funds are distributed based on objective and data-driven insights, HELP International aims to analyze various socio-economic and health indicators of different countries. The dataset includes several attributes such as child mortality rate (the number of deaths of children under five years of age per 1,000 live births), exports and imports (as a percentage of GDP per capita, reflecting a country's trade balance), health expenditure (the total health spending per capita as a percentage of GDP), income per person, inflation rate (the annual growth rate of total GDP), life expectancy (the average number of years a newborn is expected to live under current conditions), total fertility rate (the average number of children born to each woman), and GDP per capita (the total GDP divided by the population, indicating average economic output per person).

The objective is to categorize countries into distinct clusters based on these socio-economic and health factors to identify those that are in the most urgent need of assistance. By applying unsupervised learning techniques, particularly clustering algorithms, the NGO can group countries with similar conditions and strategically allocate funds to ensure maximum impact during times of disaster or economic hardship.

Dataset 9

ALMOND TYPES CLASSIFICATION

Problem Statement

A premium nut-processing company seeks to automate the classification of almond types to improve sorting accuracy, product consistency and quality control. The company has access to a dataset containing measurements of individual almond kernels — attributes such as the length of the kernel's major axis, the width of its minor axis (based on number of pixels), and other physical features extracted via image processing. They wish to determine which of these features most effectively distinguish almond varieties, and they want to build a predictive model that can reliably classify which variety a sample belongs to. You are required to apply Machine learning techniques to develop and evaluate a model. The resulting system will enable the company to streamline their processing line, reduce human sorting errors, ensure each product batch is correctly labelled by variety, and maintain consistent quality across production.