1. Mobile Price Data – Multiclass Classification

Bob wants to find out some relation between features of a mobile phone(eg:- RAM,Internal Memory etc) and its selling price. But he is not so good at Machine Learning. So he needs your help to solve this problem. In this problem you do not have to predict actual price but a price range indicating how high the price is

The target variable indicates as below:

* 0 (low cost)
* 1 (medium cost)
* 2 (high cost)
* 3 (very high cost

1. **Predict Diabetes – Binary Classification**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney

Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes,

based on certain diagnostic measurements included in the dataset. Several constraints were placed

on the selection of these instances from a larger database. In particular, all patients here are females

at least 21 years old of Pima Indian heritage.2

From the data set in the (.csv) File We can find several variables, some of them are independent

(several medical predictor variables) and only one target dependent variable (Outcome).

Outcome- To express the final result 1 is Yes and 0 is No.

1. **Car Insurance Claim Prediction – Binary Classification**

The Dataset contains information on policyholders having the attributes like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the car, power, engine type, etc, and the target variable indicating whether the policyholder files a claim in the next 6 months (1) or not (0).

1. **Heart Attack Analysis & Prediction Dataset : Binary Classification**

* Age : Age of the patient
* Sex : Sex of the patient
* exang: exercise induced angina (1 = yes; 0 = no)
* ca: number of major vessels (0-3)
* cp : Chest Pain type chest pain type
  + Value 1: typical angina
  + Value 2: atypical angina
  + Value 3: non-anginal pain
  + Value 4: asymptomatic
* trtbps : resting blood pressure (in mm Hg)
* chol : cholestoral in mg/dl fetched via BMI sensor
* fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
* rest\_ecg : resting electrocardiographic results
  + Value 0: normal
  + Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  + Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
* thalach : maximum heart rate achieved
* target : 0= less chance of heart attack 1= more chance of heart attack

1. Credit Card Fraud Detection:

According to the Data Breach Index, more than 5 million records are being stolen on a daily basis, a concerning statistic that shows - fraud is still very common both for Card-Present and Card-not Present type of payments. In today’s digital world where trillions of Card transaction happens per day, detection of fraud is challenging.

Features: 7 feature columns, including Distance from Home, Online fraud, Card Fraud etc.

Outcome variable: fraudulent transaction = 1, Non-fraudulent Transaction= 0 .

1. **Possum Regression**

Can you use your regression skills to predict the age of a possum, its head length, whether it is male or female? The possum data frame consists of nine morphometric measurements on each of 104 mountain brushtail possums, trapped at seven sites from Southern Victoria to central Queensland."

13 Features with 3 categorical Data (sex, pop and site) and 10 numerical data

* Can we use total length to predict a possum's head length? – (Regression)
* Which possum body dimensions are most correlated with age and sex?
* Can we classify a possum's sex by its body dimensions and location? – (Classification problem using LogisticRegression, DecisionTree, RandomForest or K Nearest Neighbors)
* Can we predict a possum's trapping location from its body dimensions? - (Classification problem using LogisticRegression, DecisionTree, RandomForest or K Nearest Neighbors)

<https://www.kaggle.com/datasets/abrambeyer/openintro-possum>

1. **Swiss banknote conterfeit detection : Outlier Detection**

Will you be able to identify genuine and conterfeit banknotes, even if half of the data is conterfeit? Perfect for testing different outlier detection algorithms.

The dataset includes information about the shape of the bill, as well as the label. It is made up of 200 banknotes in total, 100 for genuine/conterfeit each.

While it might be pretty easy for a classifier to decide wether the banknotes are conterfeit or not, what about methods using outlier detection?  
Classical methods of outlier detection won't work, since half of the data consist of outliers (conterfeit bills), so more robust methods will be needed.

Attributes:  
-conterfeit: Wether a banknote is conterfeit (1) or genuine (0)  
-Length: Length of bill (mm)  
-Left: Width of left edge (mm)  
-Right: Width of right edge (mm)  
-Bottom: Bottom margin width (mm)  
-Top: Top margin width (mm)  
-Diagonal: Length of diagonal (mm)

<https://www.kaggle.com/datasets/chrizzles/swiss-banknote-conterfeit-detection>

1. **Customer Clustering : Clustering with K-Means**

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. Using the above data companies can then outperform the competition by developing uniquely appealing products and services.

You are owing a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. You want to understand the customers like who are the target customers so that the sense can be given to marketing team and plan the strategy accordingly.

6 Features with both Categorical and Numerical Data.

<https://www.kaggle.com/datasets/dev0914sharma/customer-clustering>

1. **Wine Dataset for Clustering**

This dataset is adapted from the Wine Data Set from https://archive.ics.uci.edu/ml/datasets/wine by removing the information about the types of wine for unsupervised learning.

The following descriptions are adapted from the UCI webpage:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The attributes are:

* Alcohol
* Malic acid
* Ash
* Alcalinity of ash
* Magnesium
* Total phenols
* Flavanoids
* Nonflavanoid phenols
* Proanthocyanins
* Color intensity
* Hue
* OD280/OD315 of diluted wines
* Proline

<https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering>

1. **Food Demand Forecasting**

It is a meal delivery company which operates in multiple cities. They have various fulfillment centers in these cities for dispatching meal orders to their customers. The client wants you to help these centers with demand forecasting for upcoming weeks so that these centers will plan the stock of raw materials accordingly.

The replenishment of majority of raw materials is done on weekly basis and since the raw material is perishable, the procurement planning is of utmost importance. Secondly, staffing of the centers is also one area wherein accurate demand forecasts are really helpful.

OBJECTIVE: The objective is to predict the demand for the next 10 weeks (Weeks: 146-155) for the center-meal combinations in the test set.

Given : 145 weeks of data for more than 3000+ Items and their demand each week.

https://www.kaggle.com/code/anoopashware/food-demand-forecasting-predict-orders