**FINAL PROJECT**

**Problem 1: Happiness prediction**

The dataset contains the Happiness Score for over 150 countries for the year of 2020. The data gathered from the Gallup World Poll gives a national average of Happiness scores for countries all over the world. It is an annual landmark survey of the state of global happiness.

**Dataset Features**

1. Country Name
2. Regional Indicator
3. Happiness score (Ladder Score)
4. Healthy Life Expectancy (HLE)
5. Social support
6. Freedom to make life choices
7. Generosity
8. Perceptions of Corruption

**EDA Questions**

1. Which country has the highest score? Why?
2. How many observations are there?
3. Are there any null values? How does this dataset need to be cleaned?
4. Is there any correlation between the features?
5. Are the minimum and maximum happiness scores reasonable? Are there any outliers?
6. What is the mean happiness score?

**Analysis**

1. Develop a base model with useful features.
2. Develop other models for regression and try to tune hyperparameters.

**Problem 2: Cancer detection**

This problem aims to predict whether a breast cancer diagnosis is benign or malignant based on various observations or features. The dataset contains:

* **30 Features**, including:
  + Radius (mean of distances from the center to points on the perimeter)
  + Texture (standard deviation of gray-scale values)
  + Perimeter
  + Area
  + Smoothness (local variation in radius lengths)
  + Compactness
  + Concavity (severity of concave portions of the contour)
  + Concave points (number of concave portions of the contour)
  + Symmetry
  + Fractal dimension ("coastline approximation" - 1)
* **569 Instances** with a class distribution of:
  + 212 Malignant
  + 357 Benign

The objective is to classify the tumors into one of the two classes:

1. Malignant
2. Benign

**Questions:**

1. Perform EDA if necessary

2. Develop different algorithms to best classify cancer patients.

**Problem 3: Airline clustering**

Data Description: The file EastWestAirlines contains information on passengers who belong to an airline’s frequent flier program. For each passenger the data include information on their mileage history and on different ways they accrued or spent miles in the last year.

**Dataset Features**

ID# --Unique ID

Balance--Number of miles eligible for award travel

Qual\_mile--Number of miles counted as qualifying for Topflight status

cc1\_miles -- Number of miles earned with freq. flyer credit card in the past 12 months:

cc2\_miles -- Number of miles earned with Rewards credit card in the past 12 months:

cc3\_miles -- Number of miles earned with Small Business credit card in the past 12 months:

1 = under 5,000

2 = 5,000 - 10,000

3 = 10,001 - 25,000

4 = 25,001 - 50,000

5 = over 50,000

Bonus\_miles--Number of miles earned from non-flight bonus transactions in the past 12 months

Bonus\_trans--Number of non-flight bonus transactions in the past 12 months

Flight\_miles\_12mo--Number of flight miles in the past 12 months

Flight\_trans\_12--Number of flight transactions in the past 12 months

Days\_since\_enrolled--Number of days since enrolled in flier program

Award?--whether that person had award flight (free flight) or not

**Questions**

1. Perform EDA (especially renaming some of the columns).

2. Identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers.

3. Visualize the results, where necessary.

**Problem 4: Association rule**

We’ll be using the Online Retail Dataset. This is a dataset containing transnational transactions made on a UK-based online retail store between 01/12/2010 and 09/12/2011. It comprises 541909 rows, with 8 attributes:

**Dataset Features**

InvoiceNo: a unique 6-digit number assigned to each transaction. If this code starts with letter ‘C’, then the order was cancelled.

StockCode: a unique 5-digit number assigned to each distinct product.

Description: the product name.

Quantity: the number of each product (item) purchased per transaction.

InvoiceDate: the date and time each transaction was completed.

UnitPrice: the product price per unit in pounds sterling.

CustomerID: a unique 5-digit number assigned to each customer.

Country: the name of the country from where the purchase was made.

**Questions:**

1. Remove null and all cancelled transactions.

2. Encode 0-1

3. Find some useful visualizations

4. Use appropriate metrics to find useful insights from the transaction dataset. (top 5 relationships).

**Problem 5: Dimensionality reduction**

Dataset Features

There are 8 variables in the data:

* mpg: miles per gallon
* cyl: number of cylinders
* disp: engine displacement (cu. inches) or engine size
* hp: horsepower
* wt: vehicle weight (lbs.)
* acc: time taken to accelerate from O to 60 mph (sec.)
* yr: model year
* car name: car model name

Reduce the number of features by using dimensionality reduction techniques like PCA and TSNE and generate meaningful insights.