Project Ideas

Pattern Recognition Course

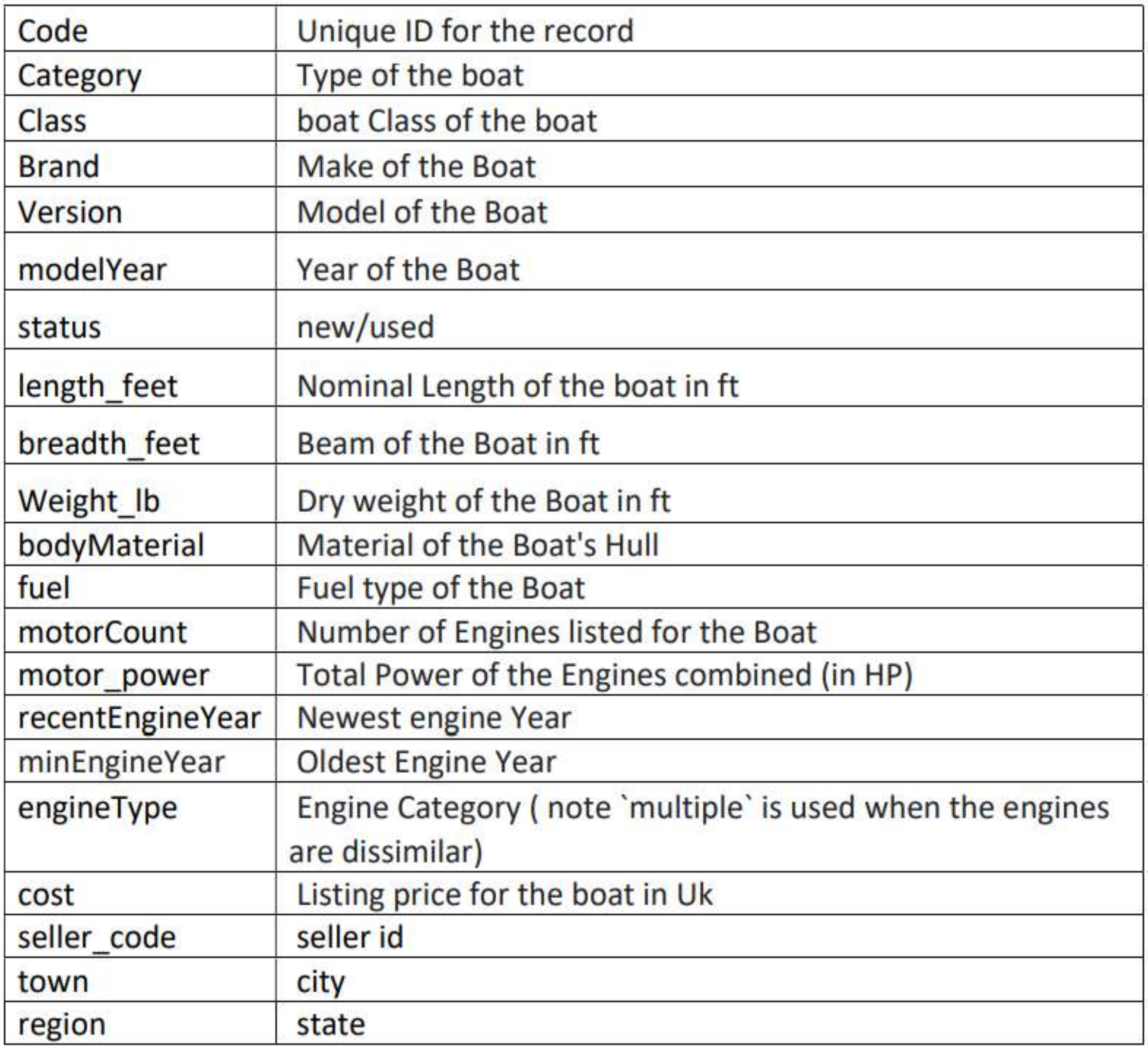
Project (Spring 2023-2024)

Dr. Dina Khattab

1. Anticipating Boat Costs.

In this project, you must build a model that forecasts the future value of boats based on historical data and market trends.

- **Features of data:**



Code	Unique ID for the record
Category	Type of the boat
Class	boat Class of the boat
Brand	Make of the Boat
Version	Model of the Boat
modelYear	Year of the Boat
status	new/used
length_feet	Nominal Length of the boat in ft
breadth_feet	Beam of the Boat in ft
Weight_lb	Dry weight of the Boat in ft
bodyMaterial	Material of the Boat's Hull
fuel	Fuel type of the Boat
motorCount	Number of Engines listed for the Boat
motor_power	Total Power of the Engines combined (in HP)
recentEngineYear	Newest engine Year
minEngineYear	Oldest Engine Year
engineType	Engine Category (note `multiple` is used when the engines are dissimilar)
cost	Listing price for the boat in Uk
seller_code	seller id
town	city
region	state

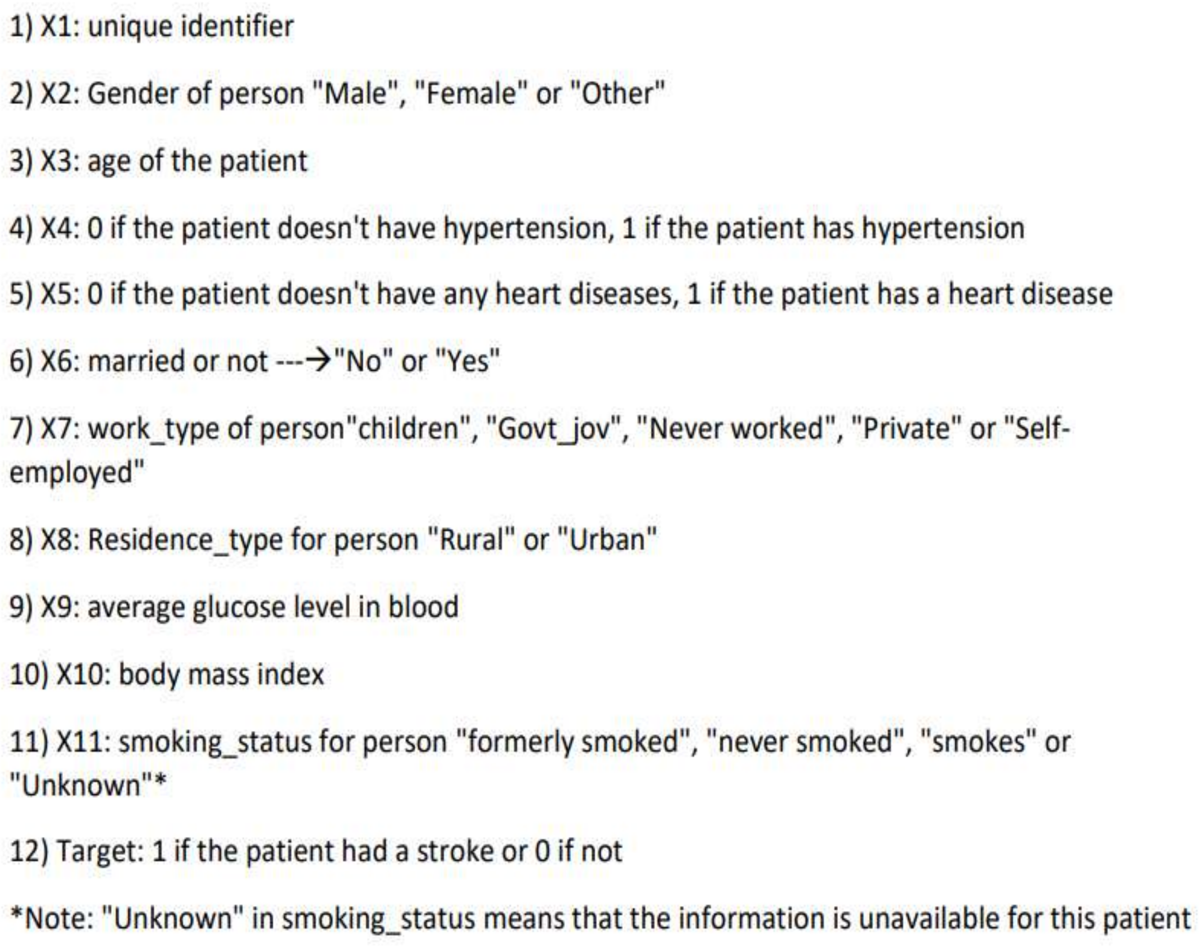
There are four remaining features:

- Postal_code
- Creation_date
- Creation_year
- Creation_month

2. Cause of Death.

- According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
- This dataset is used to predict whether a patient is likely to get stroke based on the input parameters. Each row in the data provides relevant information about the patient.
- Machine learning models can indeed play a significant role in predicting Stroke Prediction. By leveraging datasets containing relevant features, such as demographic information, medical history, and lifestyle factors, machine learning algorithms can learn patterns and make predictions about the likelihood of an individual developing a Stroke.

Attribute Information:



1) X1: unique identifier

2) X2: Gender of person "Male", "Female" or "Other"

3) X3: age of the patient

4) X4: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) X5: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) X6: married or not ---→"No" or "Yes"

7) X7: work_type of person"children", "Govt_jov", "Never worked", "Private" or "Self-employed"

8) X8: Residence_type for person "Rural" or "Urban"

9) X9: average glucose level in blood

10) X10: body mass index

11) X11: smoking_status for person "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) Target: 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

3. About Diabetes Dataset.

- The Diabetes Dataset contains information about individuals diagnosed with diabetes, including demographic attributes, medical history, and clinical measurements.
- This dataset serves as a valuable resource for studying diabetes management, risk factors, and predictive modeling for disease outcomes.

Dataset Features:

- Preg: To express the Number of pregnancies.
- Glucose: To express the Glucose level in blood.
- BPressure: To express the Blood pressure measurement.
- SThickness: To express the thickness of the skin.
- Insulin: To express the Insulin level in blood.
- BMI: To express the Body mass index.
- DiabetesPedigreeFunction: To express the Diabetes percentage.
- Age: To express the age.
- Outcome: To express the final result 1 is YES o is NO.

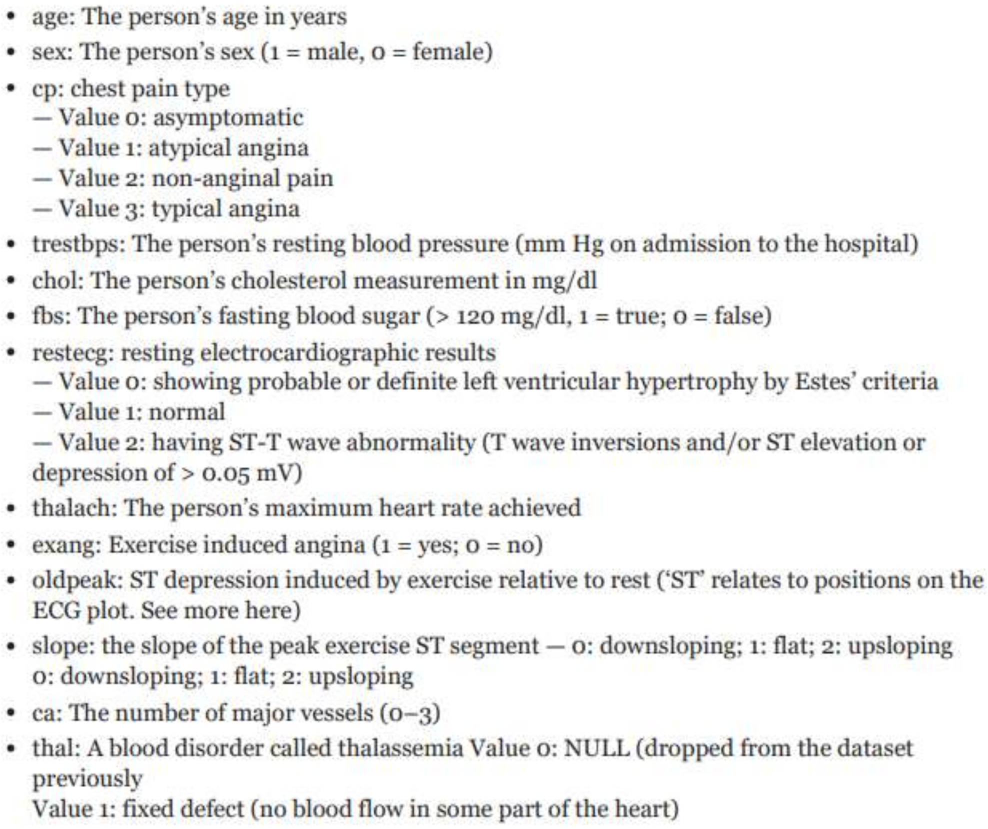
4. Heart disease dataset.

Dataset description.

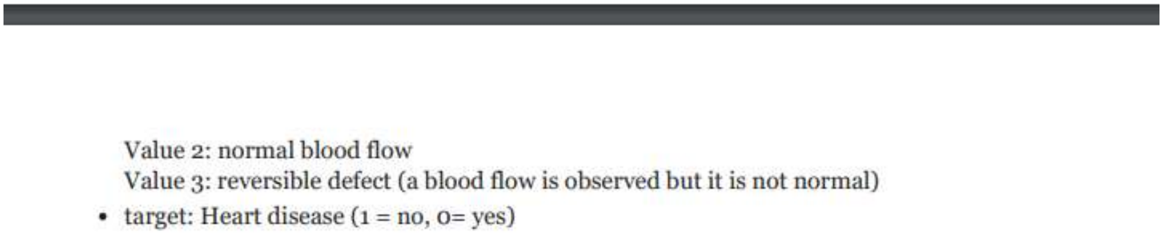
- Every day, the average human heart beats around 100,000 times, pumping 2,000 gallons of blood through the body. Inside your body there are 60,000 miles of blood vessels.
- The signs of a woman having a heart attack are much less noticeable than the signs of a male. In women, heart attacks may feel uncomfortable squeezing, pressure, fullness, or pain in the center of the chest. It may also cause pain in one or both arms, the back, neck, jaw or stomach, shortness of breath, nausea and other symptoms. Men experience typical symptoms of heart attack, such as chest pain, discomfort, and stress. They may also experience pain in other areas, such as arms, neck, back, and jaw, and shortness of breath, sweating, and discomfort that mimics heartburn.

- It’s a lot of work for an organ which is just like a large fist and weighs between 8 and 12 ounces.

Dataset columns.



- age: The person’s age in years
- sex: The person’s sex (1 = male, 0 = female)
- cp: chest pain type
  — Value 0: asymptomatic
  — Value 1: atypical angina
  — Value 2: non-anginal pain
  — Value 3: typical angina
- trestbps: The person’s resting blood pressure (mm Hg on admission to the hospital)
- chol: The person’s cholesterol measurement in mg/dl
- fbs: The person’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- restecg: resting electrocardiographic results
  — Value 0: showing probable or definite left ventricular hypertrophy by Estes’ criteria
  — Value 1: normal
  — Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- thalach: The person’s maximum heart rate achieved
- exang: Exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot. See more here)
- slope: the slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2: upsloping
  0: downsloping; 1: flat; 2: upsloping
- ca: The number of major vessels (0–3)
- thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously
  Value 1: fixed defect (no blood flow in some part of the heart)



  Value 2: normal blood flow
  Value 3: reversible defect (a blood flow is observed but it is not normal)
- target: Heart disease (1 = no, 0= yes)

5. Cocktail Juice Quality Dataset.

- Cocktail Juice quality dataset is also popular and interesting for all the machine learning and deep learning enthusiasts.
- This dataset is also beginner friendly, and you can easily apply machine learning algorithm in this data. With the help of this dataset, you can train your model to predict the Juice quality.
- This dataset has Juice's physicochemical properties. Regression and classification both approach of machine learning can be used by using Cocktail Juice quality dataset.
- In the dataset, the classes are ordered and not balanced (e.g. there are much more normal Juices than excellent or poor ones).

Information about input variables based on physicochemical tests:

- Fruit quality.
- Fixed acidity.
- Citric acid.
- Percentage of sweeteners.
- Chlorides.
- Calories.
- Density.
- Ph.
- Sulphates.
- Output variables.
  Quality (score between 0 and 10)

Features.

- Two types of variables are there in the dataset, i.e., input and output variables.
- Input variables are Fruit quality, fixed acidity, citric acid and so forth.
- The output variable is quality.
- Attributes are present and the attribute characteristics are real.

6. News-Detection Datasets.

Dataset Description.

The dataset contains two types of articles unreal and real News. This dataset was collected from real world sources; the truthful articles were obtained by crawling articles from Reuters.com (News website).

As for the fake news articles, they were collected from different sources. The fake news articles were collected from unreliable websites that were flagged by PolitiFact (a fact-checking organization in the USA) and Wikipedia.

Project goal.

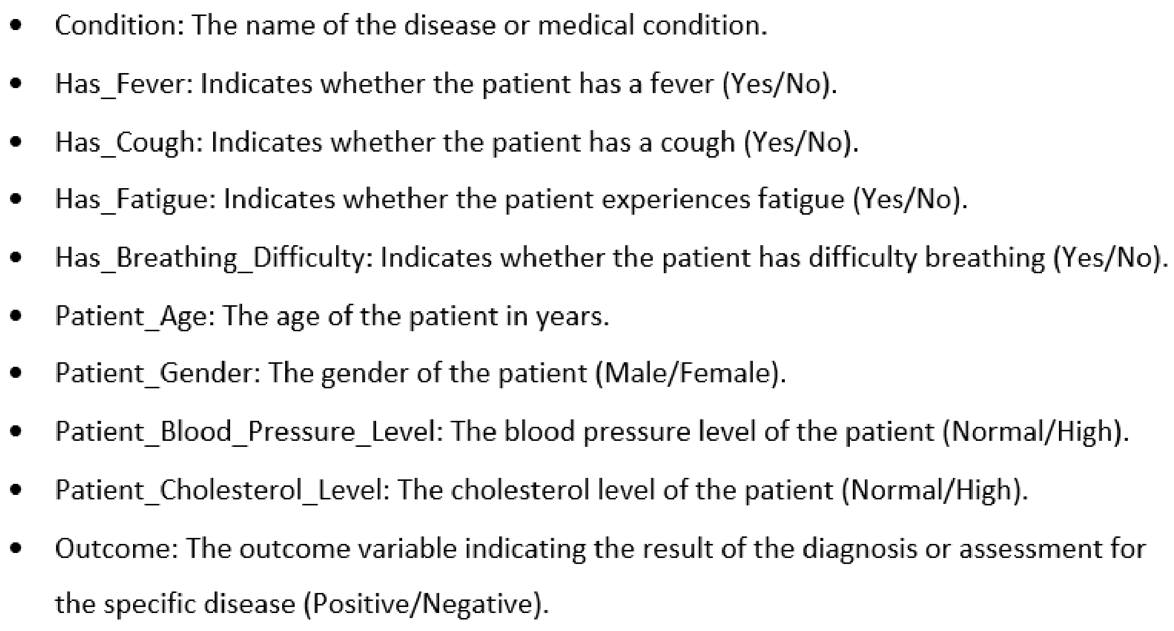
Develop a machine learning model to predict whether which article is reliable, and which is unreliable using a provided dataset.

The dataset contains various features that may be relevant for this classification task. Your goal is to explore the dataset, preprocess the data as needed, select appropriate features, choose, and train machine learning algorithms, evaluate their performance of the model.

7. Disease Classifier.

- Unlock the secrets of various diseases with our comprehensive disease symptom and patient profile dataset.
- This intriguing dataset offers a wealth of information, unveiling the intricate connections between symptoms, demographic factors, and health indicators.
- Dive deep into the realm of fever, cough, fatigue, and breathing difficulties, intertwined with age, gender, blood pressure, and cholesterol levels. Whether you're a medical researcher, healthcare practitioner, or data enthusiast, this dataset holds the key to discovering profound insights.
- Explore concealed patterns, uncover distinctive symptom profiles, and embark on a captivating journey through the landscape of medical conditions. Prepare to revolutionize healthcare comprehension with our dataset.

Data Description and Usage:



- Condition: The name of the disease or medical condition.
- Has_Fever: Indicates whether the patient has a fever (Yes/No).
- Has_Cough: Indicates whether the patient has a cough (Yes/No).
- Has_Fatigue: Indicates whether the patient experiences fatigue (Yes/No).
- Has_Breathing_Difficulty: Indicates whether the patient has difficulty breathing (Yes/No).
- Patient_Age: The age of the patient in years.
- Patient_Gender: The gender of the patient (Male/Female).
- Patient_Blood_Pressure_Level: The blood pressure level of the patient (Normal/High).
- Patient_Cholesterol_Level: The cholesterol level of the patient (Normal/High).
- Outcome: The outcome variable indicating the result of the diagnosis or assessment for the specific disease (Positive/Negative).

Target:

- By using machine learning techniques. Develop a model to predict the outcome variable (Positive/Negative) based on the given symptoms, demographic information, and health indicators.
- Evaluate the performance of your model using appropriate metrics and provide insights into the predictive power of the features included in the dataset. Additionally, discuss any challenges you encountered during the modeling process and propose potential solutions for improvement.

8. XYZ Dataset.

Dataset Description.

- Early detection and management of cardiovascular diseases (CVDs) are crucial for reducing mortality rates associated with heart attacks, strokes, and other related conditions. Machine learning models can indeed play a significant role in predicting and identifying individuals at high risk of developing CVDs.
- By leveraging datasets containing relevant features, such as demographic information, medical history, and lifestyle factors, machine learning algorithms can learn patterns and make predictions about the likelihood of an individual developing a heart disease.

Project goal.

- Develop a machine learning model to predict whether a patient has heart failure or not using a provided dataset.
- The dataset contains various features that may be relevant for this classification task. Your goal is to explore the dataset, preprocess the data as needed, select appropriate features, choose, and train machine learning algorithms, evaluate their performance, and finally, deploy the best-performing model for predicting heart failure.
- Be prepared to explain your model's choices and discuss its implications for early detection and management of cardiovascular diseases.

Attribute Information.

- X1: age of the patient [years]

- X2: sex of the patient [M: Male, F: Female]

- X3: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

- X4: resting blood pressure [mm Hg]

- X5: serum cholesterol [mm/dl]

- X6: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

- X7: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

- X8: maximum heart rate achieved [Numeric value between 60 and 202]

- X9: exercise-induced angina [Y: Yes, N: No]

- X10: oldpeak = ST [Numeric value measured in depression]

- X11: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

- Target: output class [1: heart disease, 0: Normal]