DATASETS INFORMATION MANUAL

1.BRAIN STROKE DATASET:-

Context:

A stroke is a medical condition in which poor blood flow to the brain causes cell death. There are two main types of stroke: ischemic, due to lack of blood flow, and hemorrhagic, due to bleeding. Both cause parts of the brain to stop functioning properly. Signs and symptoms of a stroke may include an inability to move or feel on one side of the body, problems understanding or speaking, dizziness, or loss of vision to one side. Signs and symptoms often appear soon after the stroke has occurred. If symptoms last less than one or two hours, the stroke is a transient ischemic attack (TIA), also called a mini-stroke. A hemorrhagic stroke may also be associated with a severe headache. The symptoms of a stroke can be permanent. Long-term complications may include pneumonia and loss of bladder control.

The main risk factor for stroke is high blood pressure. Other risk factors include high blood cholesterol, tobacco smoking, obesity, diabetes mellitus, a previous TIA, end-stage kidney disease, and atrial fibrillation. An ischemic stroke is typically caused by blockage of a blood vessel, though there are also less common causes. A hemorrhagic stroke is caused by either bleeding directly into the brain or into the space between the brain's membranes. Bleeding may occur due to a ruptured brain aneurysm. Diagnosis is typically based on a physical exam and is supported by medical imaging such as a CT scan or MRI scan. A CT scan can rule out bleeding, but may not necessarily rule out ischemia, which early on typically does not show up on a CT scan. Other tests such as an electrocardiogram (ECG) and blood tests are done to determine risk factors and rule out other possible causes. Low blood sugar may cause similar symptoms.

Prevention includes decreasing risk factors, surgery to open up the arteries to the brain in those with problematic carotid narrowing, and warfarin in people with atrial fibrillation. Aspirin or statins may be recommended by physicians for prevention. A stroke or TIA often requires emergency care. An ischemic stroke, if detected within three to four and half hours, may be treatable with a medication that can break down the clot. Some hemorrhagic strokes benefit from surgery. Treatment to attempt recovery of lost function is called stroke rehabilitation, and ideally takes place in a stroke unit; however, these are not available in much of the world.

Attribute Information

1) gender: "Male", "Female" or "Other"

2) age: age of the patient

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

4) heart disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease 5) Ever-married: "No" or "Yes"

6) work type: "children", "Govtjov", "Never worked", "Private" or "Self-employed" 7) Residencetype: "Rural" or "Urban"

8) avg glucose level: average glucose level in blood

9) BMI: body mass index

10) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

11) stroke: 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**

This dataset little preprocessed, I dropped outliers and very rare categorical values.

I dropped also the "id" columns. I suggest for this dataset, a drop in the "age" feature is little than 38 years old.

2. HEART DISEASE DATASET:-

Public Health Dataset

Context

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

Content

Attribute Information:

1. age
2. sex

3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

3. LUNG CANCER DATASET:-

Does Smoking cause Lung Cancer.

About Dataset

The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status. The data is collected from the website online lung cancer prediction system .

Total no. of attributes: 16   No .of instances: 284

Attribute information:

1. Gender: M(male), F(female)
2. Age: Age of the patient
3. Smoking: YES=2 , NO=1.
4. Yellow fingers: YES=2 , NO=1.
5. Anxiety: YES=2 , NO=1.
6. Peer_pressure: YES=2 , NO=1.
7. Chronic Disease: YES=2 , NO=1.
8. Fatigue: YES=2 , NO=1.
9. Allergy: YES=2 , NO=1.
10. Wheezing: YES=2 , NO=1.

11. Alcohol: YES=2 , NO=1.
12. Coughing: YES=2 , NO=1.
13. Shortness of Breath: YES=2 , NO=1.
14. Swallowing Difficulty: YES=2 , NO=1.
15. Chest pain: YES=2 , NO=1.
16. Lung Cancer: YES , NO.

4. THYROID DISEASE DATASET:-

Patient demographics and blood test results along Thyroid Disease diagnostic

Context

The datasets featured below were created by reconciling thyroid disease datasets provided by the UCI Machine Learning Repository.

Content

The size for the file featured within this Kaggle dataset is shown below — along with a list of attributes, and their description summaries:

thyroidDF.csv - 9172 observations x 31 attributes

Attribute information:

1. age - age of the patient (int)
2. sex - sex patient identifies (str)
3. on_thyroxine - whether patient is on thyroxine (bool)
4. query on thyroxine - *whether patient is on thyroxine (bool)
5. on antithyroid meds - whether patient is on antithyroid meds (bool)
6. sick - whether patient is sick (bool)
7. pregnant - whether patient is pregnant (bool)
8. thyroid_surgery - whether patient has undergone thyroid surgery (bool)
9. I131_treatment - whether patient is undergoing I131 treatment (bool)
10. query_hypothyroid - whether patient believes they have hypothyroid (bool)
11. query_hyperthyroid - whether patient believes they have hyperthyroid (bool)
12. lithium - whether patient * lithium (bool)
13. goitre - whether patient has goitre (bool)
14. tumor - whether patient has tumor (bool)

15. hypopituitary - whether patient * hyperpituitary gland (float)
16. psych - whether patient * psych (bool)
17. TSH_measured - whether TSH was measured in the blood (bool)
18. TSH - TSH level in blood from lab work (float)
19. T3_measured - whether T3 was measured in the blood (bool)
20. T3 - T3 level in blood from lab work (float)
21. TT4_measured - whether TT4 was measured in the blood (bool)
22. TT4 - TT4 level in blood from lab work (float)
23. T4U_measured - whether T4U was measured in the blood (bool)
24. T4U - T4U level in blood from lab work (float)
25. FTI_measured - whether FTI was measured in the blood (bool)
26. FTI - FTI level in blood from lab work (float)
27. TBG_measured - whether TBG was measured in the blood (bool)
28. TBG - TBG level in blood from lab work (float)
29. referral_source - (str)
30. target - hyperthyroidism medical diagnosis (str)
31. patient_id - unique id of the patient (str)

Target Metadata

The diagnosis consists of a string of letters indicating diagnosed conditions.

A diagnosis "-" indicates no condition requiring comment. A diagnosis of the form "X|Y" is interpreted as "consistent with X, but more likely Y". The conditions are divided into groups where each group corresponds to a class of comments.

hyperthyroid conditions:

1. hyperthyroid
2. T3 toxic
3. toxic goitre
4. secondary toxic
5. hypothyroid
6. primary hypothyroid
7. compensated hypothyroid
8. secondary hypothyroid

binding protein:

1. increased binding protein
2. decreased binding protein

general health:

1. concurrent non-thyroidal illness

replacement therapy:

2. consistent with replacement therapy
3. underreplaced
4. overreplaced

antithyroid treatment:

1. antithyroid drugs
2. I131 treatment
3. surgery

miscellaneous:

1. discordant assay results
2. elevated TBG
3. elevated thyroid hormones

Source

Thyroid Data - https://archive.ics.uci.edu/ml/datasets/thyroid+disease

5.DAIBETES DATASET:-

This dataset seems to be related to diabetes, containing information about various health-related features for different individuals. Each row in the dataset represents a person, and the columns represent different attributes or measurements associated with them.

Here's a breakdown of the columns in the dataset:

1. **Pregnancies:** The number of times the individual has been pregnant.

2. **Glucose**: The plasma glucose concentration, typically measured after fasting, which is an indicator of blood sugar levels.
3. **BloodPressure**: The diastolic blood pressure (mm Hg) of the individual.
4. **SkinThickness**: The skinfold thickness (mm) at the triceps, which may provide information about body fat.
5. **Insulin**: The 2-Hour serum insulin (mu U/ml) level, an important hormone that regulates blood sugar.
6. **BMI (Body Mass Index):** A numerical value of body composition calculated from weight and height. It gives an idea of whether an individual is underweight, normal weight, overweight, or obese.
7. **DiabetesPedigreeFunction**: A function that represents the diabetes genetic history of an individual in a family.
8. **Age**: The age of the individual in years.
9. **Outcome**: This seems to be the target variable or label. It is binary, with 1 indicating the presence of diabetes and 0 indicating the absence.

It looks like this dataset could be used for building a predictive model to determine whether an individual is likely to have diabetes based on these health-related features. The attributes such as glucose levels, BMI, and age are commonly associated with diabetes risk assessment.

Before using this dataset for any analysis or modeling, you might want to perform some preprocessing steps, such as handling missing values (e.g., replacing zeroes with appropriate values), scaling the features, and splitting the data into training and testing sets.

Please note that this description is based solely on the provided column names and their general interpretations. Additional domain knowledge or context might be required for a more accurate understanding of the dataset.