DATASET SURVEY

- ## Diabetes

1) Pima Indians Diabetes Database:

Source: Kaggle

Link:
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Size:

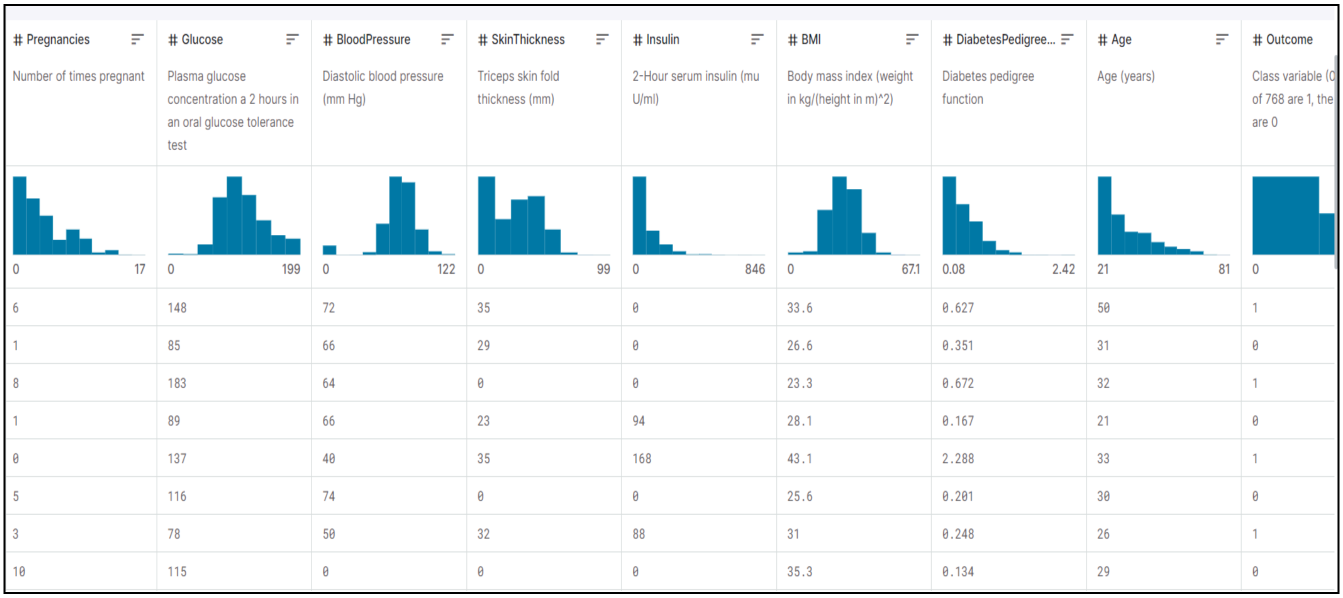
This dataset is consist of **9** columns and **769** rows

Predictor variables:

Pregnancies - Number of times pregnant
Glucose - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure - Diastolic blood pressure (mm Hg)
SkinThickness - Triceps skinfold thickness (mm)
Insulin - 2-Hour serum insulin (mu U/ml)
BMI - Body mass index (weight in kg/(height in m)^2)
Diabetes pedigree - Diabetes pedigree function
Age - Age (years)

Target (Dependent) variable:

Outcome - Class variable (0 or 1) 268 of 768 are 1, the others are 0



# Pregnancies	# Glucose	# BloodPressure	# SkinThickness	# Insulin	# BMI	# DiabetesPedigree...	# Age	# Outcome
Number of times pregnant	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Diastolic blood pressure (mm Hg)	Triceps skin fold thickness (mm)	2-Hour serum insulin (mu U/ml)	Body mass index (weight in kg/(height in m)^2)	Diabetes pedigree function	Age (years)	Class variable (0 of 768 are 1, the are 0
0 17	0 199	0 122	0 99	0 846	0 67.1	0.08 2.42	21 81	0
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0

2) Diabetes Health Indicators Dataset:

Source: Kaggle

Link:

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/

Size:

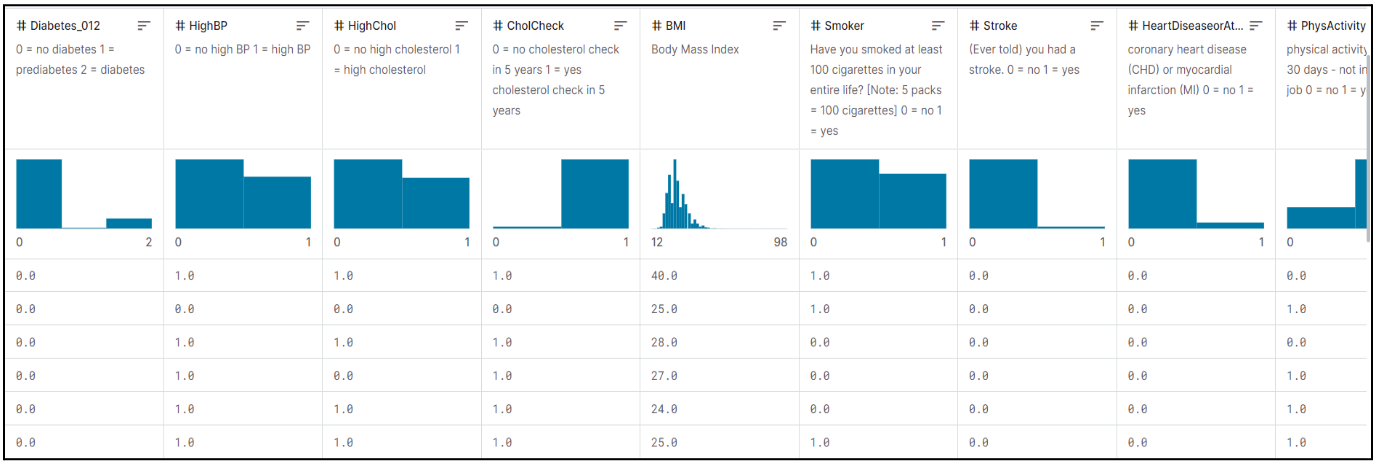
This dataset is consist of **21** columns and **253681** rows

Predictor variables:

HighBP - 0 = no high BP 1 = high BP
HighChol - 0 = no high cholesterol 1 = high cholesterol
CholCheck - 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
BMI - Body Mass Index
Smoker - Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
Stroke - (Ever told) you had a stroke. 0 = no 1 = yes
HeartDiseaseorAttack - coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
PhysActivity - physical activity in past 30 days - not including job 0 = no 1 = yes
HeavyAlcoholConsumption - Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no

AnyHealthcare - Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
GenHlth - Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
MentHlth - Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth - Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk - Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Sex - 0 = female 1 = male
Age - 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older

Target (Dependent) variable:
Diabetes_012 - 0 = no diabetes 1 = pre diabetes 2 = diabetes



# Diabetes_012	# HighBP	# HighChol	# CholCheck	# BMI	# Smoker	# Stroke	# HeartDiseaseorAt...	# PhysActivity
0 = no diabetes 1 = prediabetes 2 = diabetes	0 = no high BP 1 = high BP	0 = no high cholesterol 1 = high cholesterol	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years	Body Mass Index	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes	(Ever told) you had a stroke. 0 = no 1 = yes	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes	physical activity 30 days - not in job 0 = no 1 = y
0 2	0 1	0 1	0 1	12 98	0 1	0 1	0 1	0
0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0
0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0

- **Brain Tumor**

1) Brain Tumor:

Source: Kaggle

Link:
https://www.kaggle.com/datasets/jakeshbohaju/brain-tumor

Size:

This dataset is consist of **3764** image files

Predictor variables:

Image - Image name

Mean - First order feature mean

Variance - First order feature variance

Standard Deviation - First order feature std deviation

Entropy - Second order feature entropy

Skewness - First order feature skewness

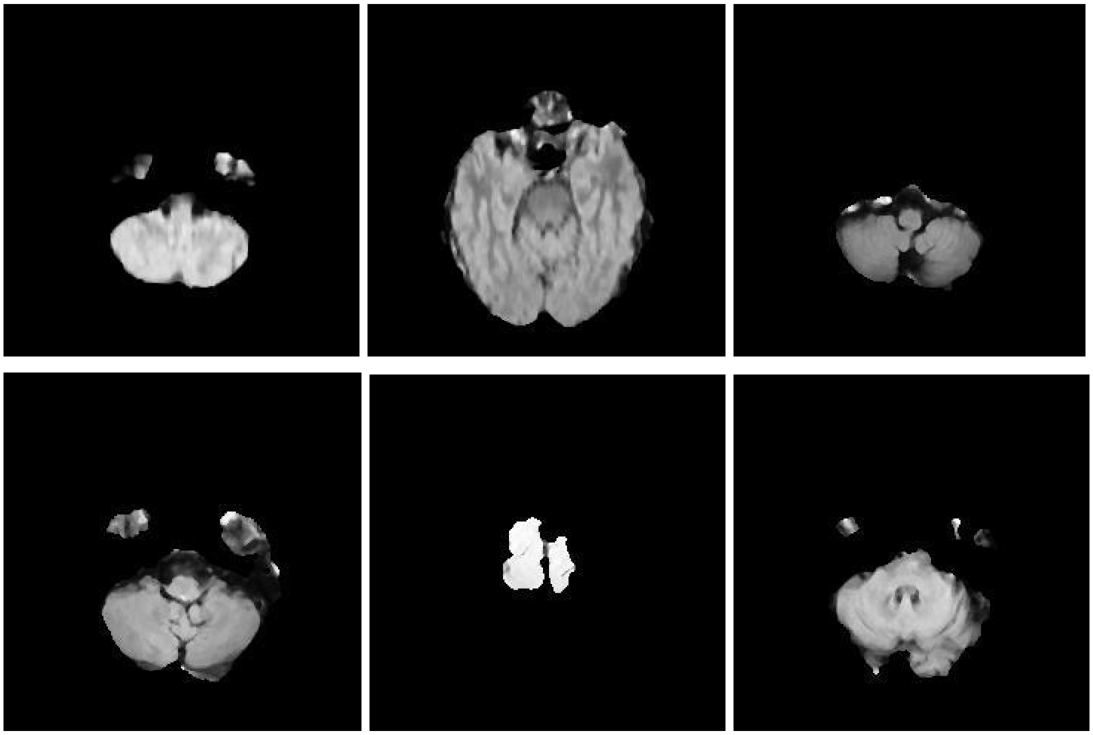
Kurtosis - First order feature kurtosis

Contrast - Second order feature contrast

Energy - Second order feature energy

Target (Dependent) variable:

Class - Target value Tumor = 1 Non tumor =0



2) Brain Tumor MRI Dataset

Source: Kaggle

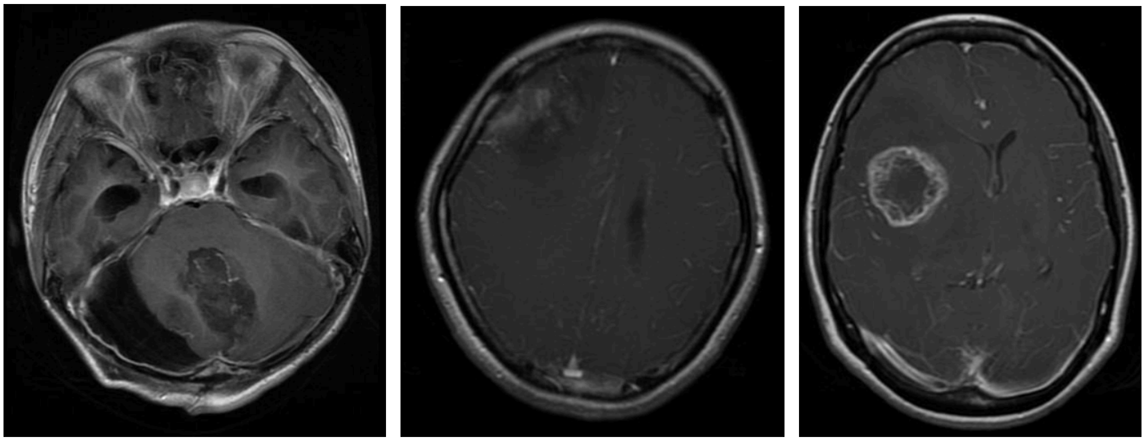
Link:
https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Size:
This dataset is consist of **7022** image files

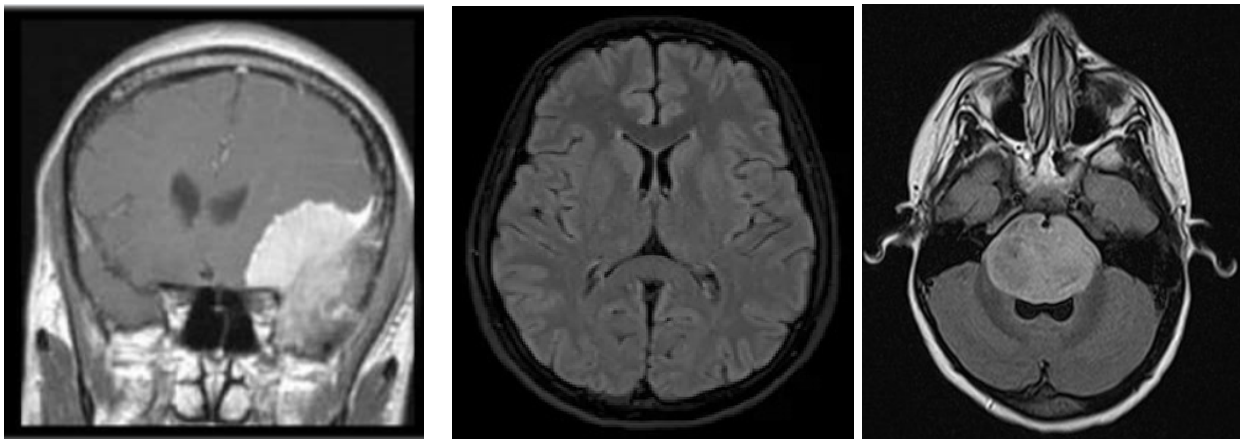
Human brain MRI images are classified into 4 classes:

1) **Glioma**
2) **Meningioma**
3) **No tumor**
4) **Pituitary**

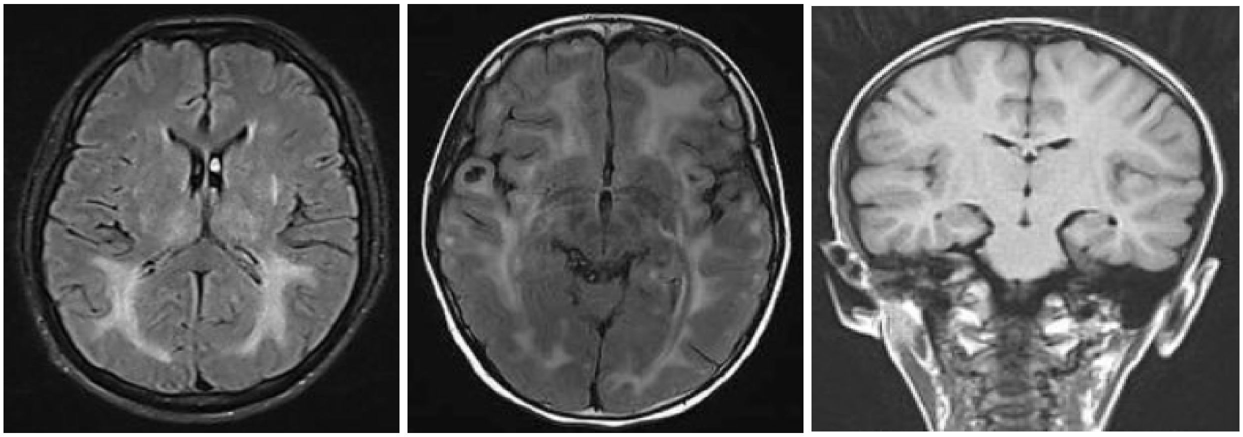
Glioma:



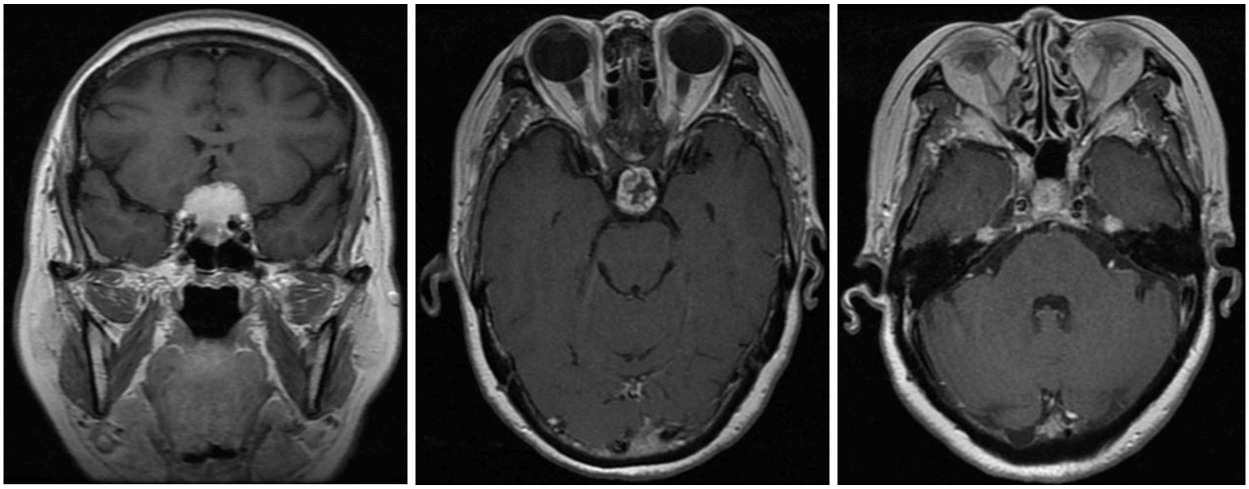
Meningioma:



Notumor :



Pituitary:



- **Lung Cancer**

1) Lung Cancer Prediction:

Source: Kaggle

Link:
https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

Size:

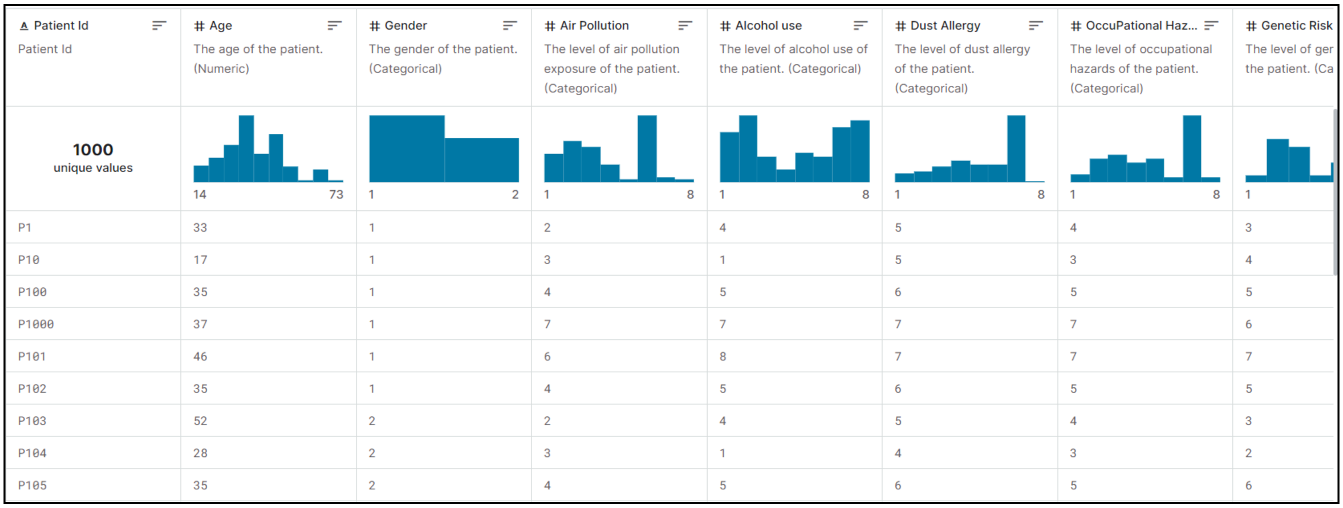
This dataset is consist of **26** columns and **1000** rows

Predictor variables:

Patient Id - Patient Id
Age - The age of the patient. (Numeric)
Gender - The gender of the patient. (Categorical)
Air Pollution - The level of air pollution exposure of the patient. (Categorical)
Alcohol use - The level of alcohol use of the patient. (Categorical)
Dust Allergy - The level of dust allergy of the patient. (Categorical)
OccuPational Hazards - The level of occupational hazards of the patient. (Categorical)
Genetic Risk - The level of genetic risk of the patient. (Categorical)

Target (Dependent) variable:

Chronic Lung Disease - The level of chronic lung disease of the patient. (Categorical)



Patient Id	# Age	# Gender	# Air Pollution	# Alcohol use	# Dust Allergy	# OccuPational Haz...	# Genetic Risk
Patient Id	The age of the patient. (Numeric)	The gender of the patient. (Categorical)	The level of air pollution exposure of the patient. (Categorical)	The level of alcohol use of the patient. (Categorical)	The level of dust allergy of the patient. (Categorical)	The level of occupational hazards of the patient. (Categorical)	The level of ger the patient. (Ca
1000 unique values	14 73	1 2	1 8	1 8	1 8	1 8	1
P1	33	1	2	4	5	4	3
P10	17	1	3	1	5	3	4
P100	35	1	4	5	6	5	5
P1000	37	1	7	7	7	7	6
P101	46	1	6	8	7	7	7
P102	35	1	4	5	6	5	5
P103	52	2	2	4	5	4	3
P104	28	2	3	1	4	3	2
P105	35	2	4	5	6	5	6

2) Survey Lung Cancer:

Source: Kaggle

Link:
https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction/input

Size:

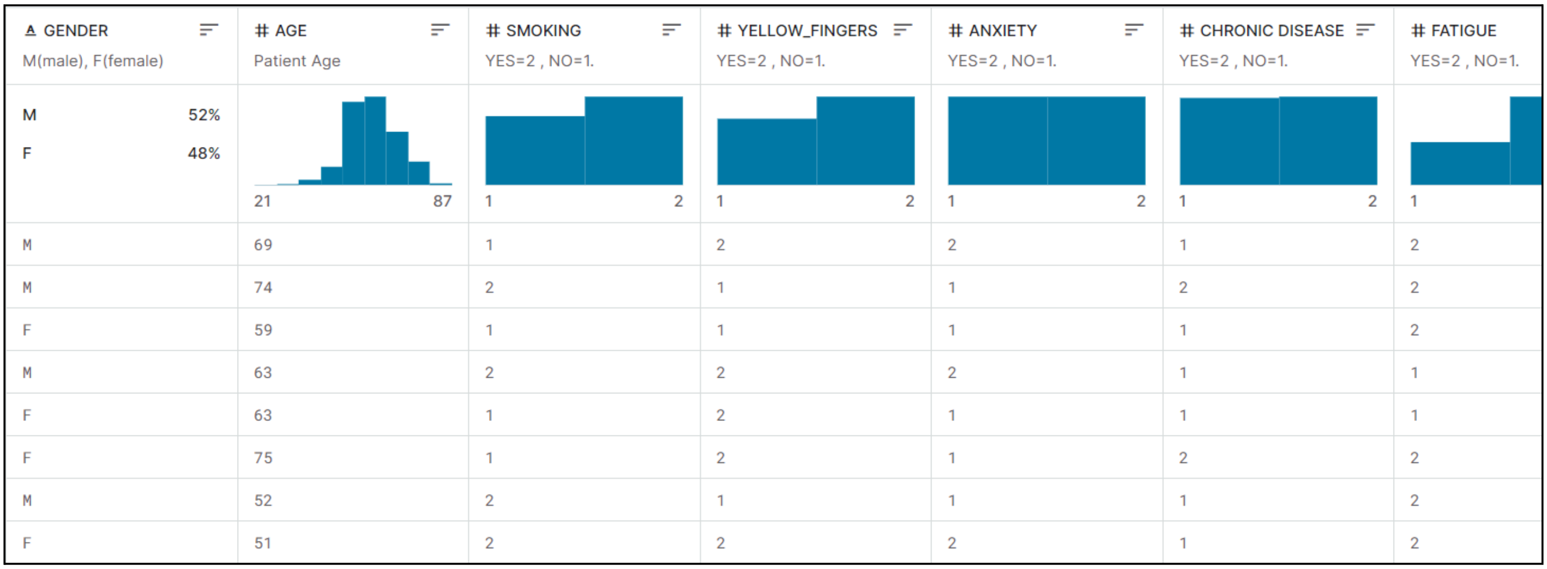
This dataset is consist of **16** columns and **284** rows

Predictor variables:

Gender - M(male), F(female)
Age - Patient Age
Smoking - YES=2 , NO=1
Yellow_Fingers - YES=2 , NO=1
Anxiety - YES=2 , NO=1
Chronic Disease - YES=2 , NO=1
Fatigue - YES=2 , NO=1
Allergy - YES=2 , NO=1
Wheezing - YES=2 , NO=1
Alcohol Consuming - YES=2 , NO=1
Shortness Of Breath - YES=2 , NO=1
Chest Pain - YES=2 , NO=1

Target (Dependent) variable:

Lung_Cancer - YES=2 , NO=1



GENDER	# AGE	# SMOKING	# YELLOW_FINGERS	# ANXIETY	# CHRONIC DISEASE	# FATIGUE
M(male), F(female)	Patient Age	YES=2 , NO=1.	YES=2 , NO=1.	YES=2 , NO=1.	YES=2 , NO=1.	YES=2 , NO=1.
M 52% F 48%	21 – 87	1 – 2	1 – 2	1 – 2	1 – 2	1
M	69	1	2	2	1	2
M	74	2	1	1	2	2
F	59	1	1	1	1	2
M	63	2	2	2	1	1
F	63	1	2	1	1	1
F	75	1	2	1	2	2
M	52	2	1	1	1	2
F	51	2	2	2	1	2

- **Alzheimer**

1) Alzheimer_s Dataset:

Source: Kaggle

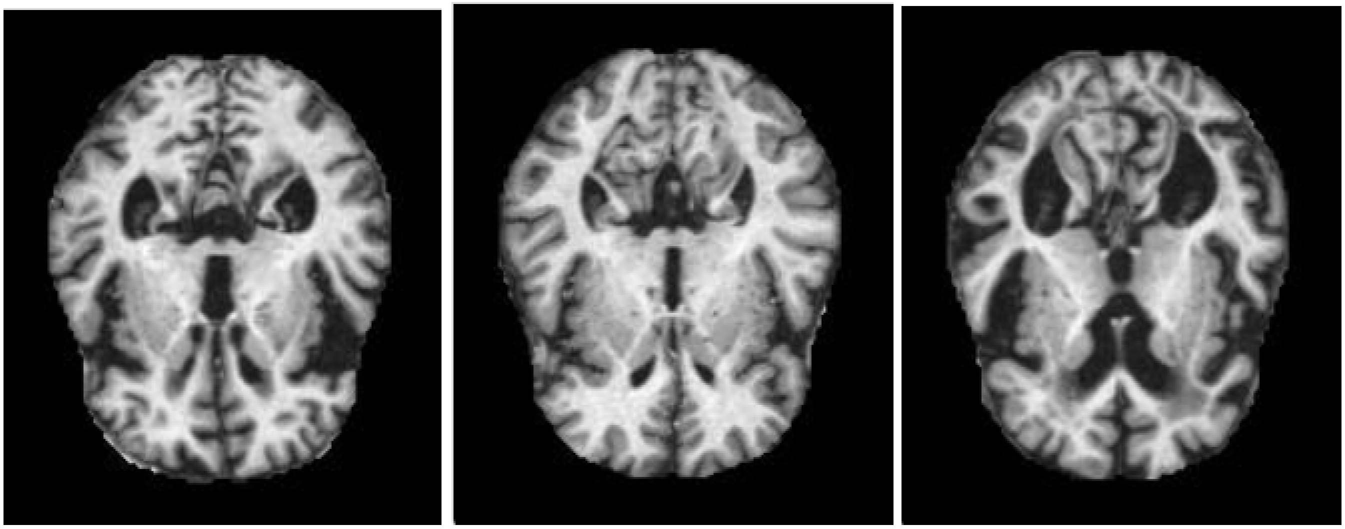
Link:
https://www.kaggle.com/code/amyjang/alzheimer-mri-model-tensorflow-2-3-data-loading/input

Size:
This dataset is consist of **5000** image files

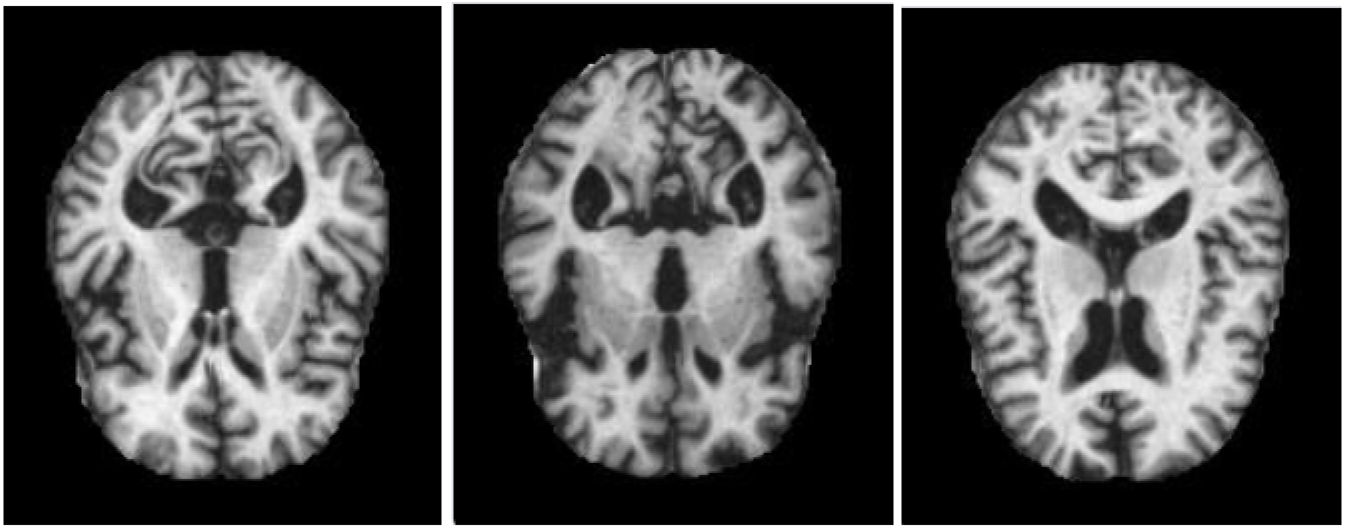
Classes:

1) MildDemented
2) VeryMildDemented
3) NonDemented
4) ModerateDemented

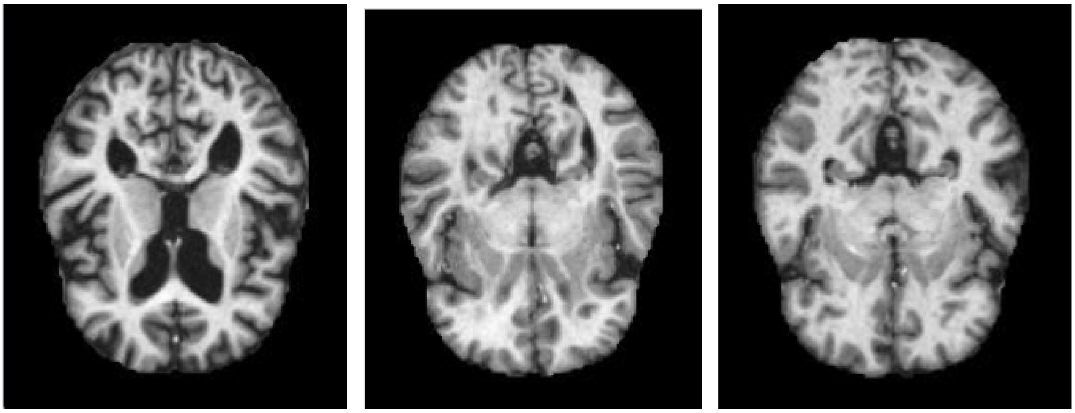
MildDemented



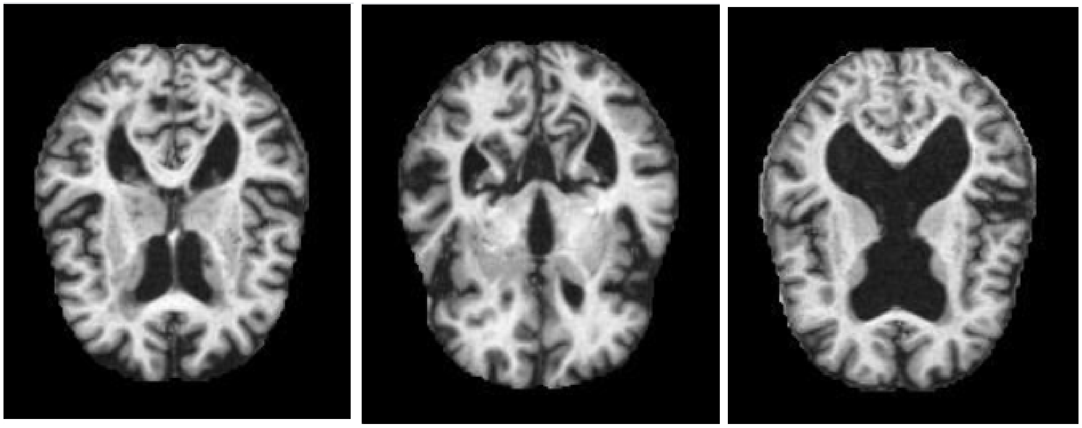
ModerateDemented



NonDemented



VeryMildDemented



2) Augmented Alzheimer MRI Dataset:

Source: Kaggle

Link:
https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset

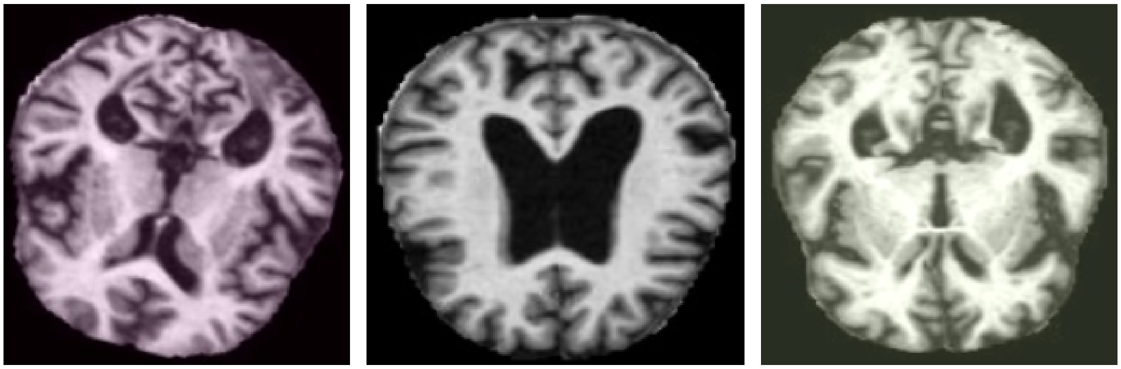
Size:

This dataset is consist of **40.4K** image files

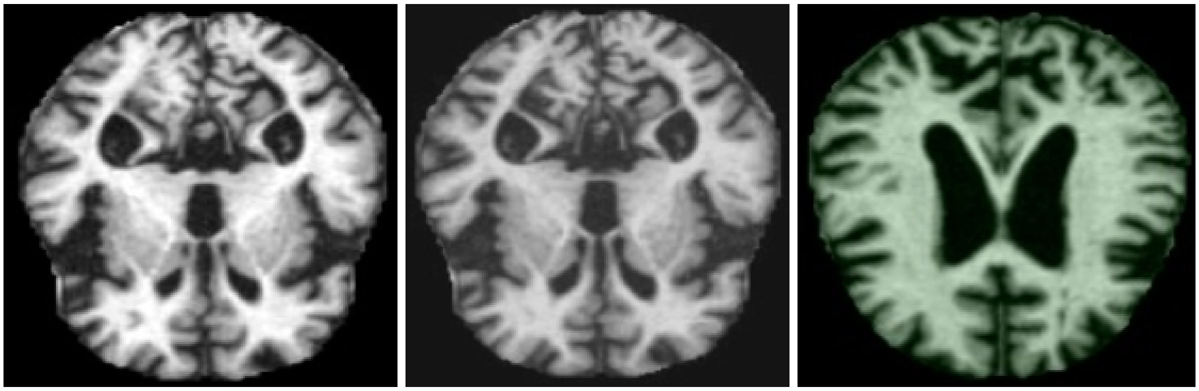
Classes:

1) MildDemented
2) VeryMildDemented
3) NonDemented
4) ModerateDemented

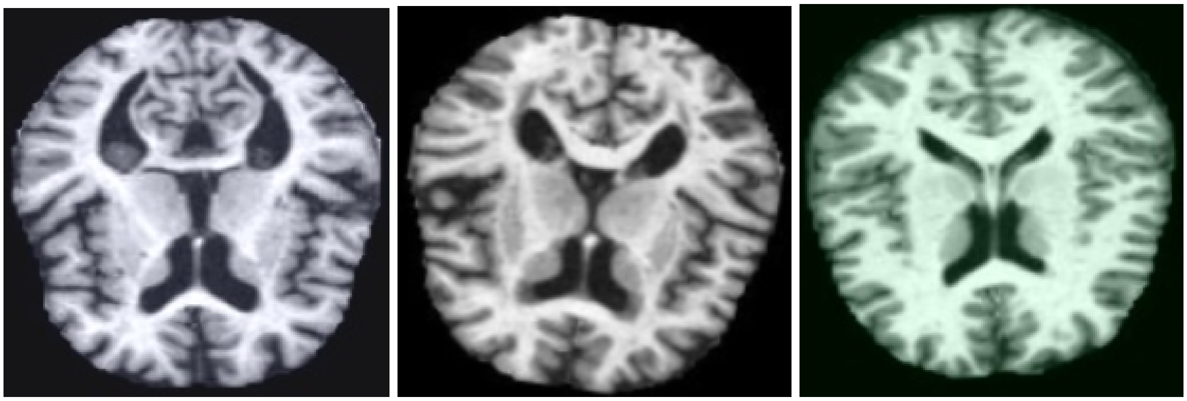
MildDemented



ModerateDemented



NonDemented



VeryMildDemented