from os.path import dirname

from os.path import join

import pandas as pd

def load\_data(data\_file\_name):

"""Loads data from module\_path/data/data\_file\_name

Args:

data\_file\_name (str) : Name of csv file to be loaded from

module\_path/data/data\_file\_name. Example: 'diabetes.csv'

Returns:

Pandas.core.frame.DataFrame: A pandas dataframe containing the loaded data.

Examples:

>>> load\_data('diabetes.csv')

"""

file\_path = join(dirname(\_\_file\_\_), 'data', data\_file\_name)

return pd.read\_csv(file\_path, na\_values=['None'])

def load\_acute\_inflammations():

"""

Loads the Acute Inflammations dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations

Note: The dataset contains the following columns:

`PatientID`: Patient Identifier

`Temperature`: Temperature of patient { 35C-42C }

`Nausea`: Occurrence of nausea { 1, 0 }

`LumbarPain`: Lumbar pain { 1, 0 }

`UrinePushing`: Urine pushing (continuous need for urination) { 1, 0 }

`MicturitionPain`: Micturition pains { 1, 0 }

`UrethralBurning`: Burning of urethra, itch, swelling of urethra outlet { 1, 0 }

`Inflammation`: Inflammation of urinary bladder { 1, 0 }

`Nephritis`: Nephritis of renal pelvis origin { 1, 0 }

"""

return load\_data('acute\_inflammations.csv')

def load\_cervical\_cancer():

"""

Loads the Cervical Cancer (Risk Factors) dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

Note: The dataset contains the following columns:

`PatientID`: Patient Identifier

Age

Number of sexual partners

First sexual intercourse (age)

Num of pregnancies

Smokes

Smokes (years)

Smokes (packs/year)

Hormonal Contraceptives

Hormonal Contraceptives (years)

IUD

IUD (years)

STDs

STDs (number)

STDs:condylomatosis

STDs:cervical condylomatosis

STDs:vaginal condylomatosis

STDs:vulvo-perineal condylomatosis

STDs:syphilis

STDs:pelvic inflammatory disease

STDs:genital herpes

STDs:molluscum contagiosum

STDs:AIDS

STDs:HIV

STDs:Hepatitis B

STDs:HPV

STDs: Number of diagnosis

STDs: Time since first diagnosis

STDs: Time since last diagnosis

Dx:Cancer

Dx:CIN

Dx:HPV

Dx

Hinselmann: target variable

Schiller: target variable

Cytology: target variable

Biopsy: target variable

"""

return load\_data('cervical\_cancer.csv')

def load\_diabetes():

"""

Loads the healthcare.ai sample diabetes dataset

Note: The dataset contains the following columns:

PatientEncounterID

PatientID

SystolicBPNBR

LDLNBR

A1CNBR

GenderFLG

ThirtyDayReadmitFLG

"""

return load\_data('diabetes.csv')

def load\_diagnostic\_breast\_cancer():

"""

Loads the Wisconsin Diagnostic Breast Cancer dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Note: The dataset contains the following columns:

`ID`: ID number

`Diagnosis`: Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

`Radius`: radius (mean of distances from center to points on the perimeter)

`Texture`: texture (standard deviation of gray-scale values)

`Perimeter`: perimeter

`Area`: area

`Smoothness`: smoothness (local variation in radius lengths)

`Compactness`: compactness (perimeter^2 / area - 1.0)

`Concavity`: concavity (severity of concave portions of the contour)

`ConcavePoints`: concave points (number of concave portions of the contour)

`Symmetry`: symmetry

`FractalDimension`: fractal dimension ("coastline approximation" - 1)

For each of these ten features, the mean, standard error, and "worst"

or largest (mean of the three largest values) of these features were

computed for each image, resulting in 30 features. Features ending with

"M" indicate Mean Radius. Features ending with "S" indicate Standard

Error. Features ending with "W" indicate Worst Radius.

"""

return load\_data('diagnostic\_breast\_cancer.csv')

def load\_fertility():

"""

Loads the Fertility dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Fertility

Note: The dataset contains the following columns:

`PatientID`: Patient Identifier

`Season`: Season in which the analysis was performed. 1) winter,

2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)

`Age`: Age at the time of analysis. 18-36 (0, 1)

`ChildishDiseases`: Childish diseases (ie , chicken pox, measles, mumps,

polio) 1) yes, 2) no. (0, 1)

`Trauma`: Accident or serious trauma 1) yes, 2) no. (0, 1)

`SurgicalIntervention`: Surgical intervention 1) yes, 2) no. (0, 1)

`HighFevers`: High fevers in the last year 1) less than three months

ago, 2) more than three months ago, 3) no. (-1, 0, 1)

`AlcoholConsumption`: Frequency of alcohol consumption 1) several times

a day, 2) every day, 3) several times a week, 4) once a week,

5) hardly ever or never (0, 1)

`SmokingHabit`: Smoking habit 1) never, 2) occasional 3) daily.

(-1, 0, 1)

`SittingHours`: Number of hours spent sitting per day ene-16 (0, 1)

`Diagnosis`: Diagnosis normal (N), altered (O)

"""

return load\_data('fertility.csv')

def load\_heart\_disease():

"""

Loads the Stratlog (Heart) dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

Note: The dataset contains the following columns:

`PatientID`: Patient Identifier

`Age`: age

`Sex`: sex

`ChestPainType`: chest pain type (4 values)

`BloodPressure`: resting blood pressure

`Cholesterol`: serum cholesterol in mg/dl

`BloodSugar`: fasting blood sugar > 120 mg/dl

`EC\_Results`: resting electrocardiographic results (values 0,1,2)

`MaxHeartRate`: maximum heart rate achieved

`Angina`: exercise induced angina

`OldPeak`: oldpeak = ST depression induced by exercise relative to rest

`PeakSlope`: the slope of the peak exercise ST segment

`MajorVessels`: number of major vessels (0-3) colored by flourosopy

`Thal`: thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

`Outcome`: Absence (1) or presence (2) of heart disease

"""

return load\_data('heart\_disease.csv')

def load\_mammographic\_masses():

"""

Loads the Mammographic Mass dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass

Note: The dataset contains the following columns:

`PatientID`: Patient Identifier

`BiRadsAssessment`: BI-RADS assessment: 1 to 5 (ordinal,

non-predictive)

`Age`: patient's age in years (integer)

`Shape`: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)

`Margin`: mass margin: circumscribed=1 microlobulated=2 obscured=3

ill-defined=4 spiculated=5 (nominal)

`Density`: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)

`Severity`: benign=0 or malignant=1 (binominal, goal field!)

"""

return load\_data('mammographic\_masses.csv')

def load\_pima\_indians\_diabetes():

"""

Loads the PIMA Indians Diabetes dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Note: The dataset contains the following columns:

`PatientID`: Patient Identifier

`Pregnancies`: Number of times pregnant

`PlasmaGlucose`: Plasma glucose concentration a 2 hours in an oral

glucose tolerance test

`DiastolicBP`: Diastolic blood pressure (mm Hg)

`TricepSkinfoldThickness`: Triceps skin fold thickness (mm)

`Insulin`: 2-Hour serum insulin (mu U/ml)

`BMI`: Body mass index (weight in kg/(height in m)^2)

`DiabetesPedigreeFunction`: Diabetes pedigree function

`Age`: Age (years)

`Diabetes`: Class variable (Y or N)

"""

return load\_data('pima\_indians\_diabetes.csv')

def load\_prognostic\_breast\_cancer():

"""

Loads the Wisconsin Prognostic Breast Cancer dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29

Note: The dataset contains the following columns:

`ID`: ID number

`Outcome`: Outcome (R = recur, N = nonrecur)

`Time`: Time (recurrence time if field 2 = R, disease-free time if field 2 = N)

`TumorSize`: diameter of the excised tumor in centimeters

`LymphNodeStatus`: number of positive axillary lymph nodes observed at time of surgery

Ten real-valued features are computed for each cell nucleus:

`Radius`: radius (mean of distances from center to points on the perimeter)

`Texture`: texture (standard deviation of gray-scale values)

`Perimeter`: perimeter

`Area`: area

`Smoothness`: smoothness (local variation in radius lengths)

`Compactness`: compactness (perimeter^2 / area - 1.0)

`Concavity`: concavity (severity of concave portions of the contour)

`ConcavePoints`: concave points (number of concave portions of the contour)

`Symmetry`: symmetry

`FractalDimension`: fractal dimension ("coastline approximation" - 1)

For each of these ten features, the mean, standard error, and "worst"

or largest (mean of the three largest values) of these features were

computed for each image, resulting in 30 features. Features ending with

"M" indicate Mean Radius. Features ending with "S" indicate Standard

Error. Features ending with "W" indicate Worst Radius.

"""

return load\_data('prognostic\_breast\_cancer.csv')

def load\_thoracic\_surgery():

"""

Loads the Thoracic Surgery dataset from the UCI ML Library

URL: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

Note: The dataset contains the following columns:

`PatientID`: Patient Identifier

`DGN`: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well

multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)

`PRE4`: Forced vital capacity - FVC (numeric)

`PRE5`: Volume that has been exhaled at the end of the first second of forced

expiration - FEV1 (numeric)

`PRE6`: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)

`PRE7`: Pain before surgery (T,F)

`PRE8`: Haemoptysis before surgery (T,F)

`PRE9`: Dyspnoea before surgery (T,F)

`PRE10`: Cough before surgery (T,F)

`PRE11`: Weakness before surgery (T,F)

`PRE14`: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14

(largest) (OC11,OC14,OC12,OC13)

`PRE17`: Type 2 DM - diabetes mellitus (T,F)

`PRE19`: MI up to 6 months (T,F)

`PRE25`: PAD - peripheral arterial diseases (T,F)

`PRE30`: Smoking (T,F)

`PRE32`: Asthma (T,F)

`AGE`: Age at surgery (numeric)

`Risk1Y`: 1 year survival period - (T)rue value if died (T,F)

"""

return load\_data('thoracic\_surgery.csv')