stroke

A stroke occurs when the blood supply to part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes. A stroke is a medical emergency, and prompt treatment is crucial. Early action can reduce brain damage and other complications.

objective is to understand what are the reasons that cause stroke to peoeple and see if we can succefully detect stroke on some features using ML technics.

Who is of people at risk for a stroke?

Import libraries

Double-click (or enter) to edit

[ ]

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
% matplotlib inline
```

Load Dataset

[ ]

```
#i will load data from google drive
from google.colab import drive
drive.mount ('/content/gdrive')
```

Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).

[ ]

```
#Data Loading,read the DS
dataset = pd.read_csv('/content/healthcare-dataset-stroke-data.csv')
```

[ ]

```
#show head of dataset
dataset.head()
```

[ ]

```
#show tail of dataset
dataset.tail()
```

Explore Data Analysis

Double-click (or enter) to edit

[ ]

```
# get some info about data
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

from info i get more information about my data ,sutch as the name, number of columns,,data type, and null values.at stroke dataset we have null values in bmi feature .

[ ]

```
#describe DS
dataset.describe()
```

[ ]

```
# detect how many rows and columns
dataset.shape
```

```
(5110, 12)
```

[ ]

```
#check the null value
dataset.isnull().sum() # bmi feature has 201 null value
```

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

[ ]

```
# Check if we have duplicate values by using 'id' feature
dataset[dataset.duplicated(['id'])]
```

---

the bmi feature has 201 null value .

[ ]

```
#get % null value from dataset
dataset.isna().sum()/dataset.shape[0]
```

```
id                   0.000000
gender               0.000000
age                  0.000000
hypertension         0.000000
heart_disease        0.000000
ever_married         0.000000
work_type            0.000000
Residence_type       0.000000
avg_glucose_level    0.000000
bmi                  0.039335
smoking_status       0.000000
stroke               0.000000
dtype: float64
```

[ ]

```
#sace copy from data set and work on it
df = dataset.copy()
```

[ ]

```
df
```

Data Cleaning

[ ]

```
#show bmi value as hist vesualisation
df.bmi.hist()
```

[ ]

```
#i will handle this missing values by using median
df.bmi.fillna(df.bmi.median(),inplace=True)
```

[ ]

```
#check the missing value after handling
df.isnull().sum()
#filled with the median of the same column. For feature extraction
```

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

[ ]

```
#check stroke values preprosse
df.stroke.value_counts()
##at this DS it have imbalanced data ,where the value of patient doesn't have Stroke =4861,patient have stoke only =249.
```

```
0    4861
1     249
Name: stroke, dtype: int64
```

[ ]

```
#i will drop id column becuse it dosen't help me on analysis DS
df=df.drop('id',axis=1)
df.head(2)
```

visualization

[ ]

```
!pip install dataprep

# Restart the runtime
```

```
Collecting dataprep
  Downloading dataprep-0.4.1-py3-none-any.whl (3.5 MB)
     |████████████████████████████████| 3.5 MB 4.9 MB/s
Requirement already satisfied: bokeh<3,>=2 in /usr/local/lib/python3.7/dist-packages (from dataprep) (2.3.3)
Collecting usaddress<0.6.0,>=0.5.10
  Downloading usaddress-0.5.10-py2.py3-none-any.whl (63 kB)
     |████████████████████████████████| 63 kB 2.1 MB/s
Collecting flask_cors<4.0.0,>=3.0.10
  Downloading Flask_Cors-3.0.10-py2.py3-none-any.whl (14 kB)
Collecting varname<0.9.0,>=0.8.1
  Downloading varname-0.8.1-py3-none-any.whl (20 kB)
Requirement already satisfied: jinja2<3.0,>=2.11 in /usr/local/lib/python3.7/dist-packages (from dataprep) (2.11.3)
Collecting dask[array,dataframe,delayed]<3.0,>=2.25
  Downloading dask-2.30.0-py3-none-any.whl (848 kB)
     |████████████████████████████████| 848 kB 58.8 MB/s
Collecting aiohttp<4.0,>=3.6
  Downloading aiohttp-3.8.1-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (1.1 MB)
     |████████████████████████████████| 1.1 MB 44.9 MB/s
Requirement already satisfied: bottleneck<2.0,>=1.3 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.3.2)
Requirement already satisfied: scipy<2,>=1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.4.1)
Requirement already satisfied: flask<2.0.0,>=1.1.4 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.1.4)
Requirement already satisfied: numpy<2,>=1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.19.5)
Requirement already satisfied: pandas<2.0,>=1.1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.1.5)
Collecting levenshtein<0.13.0,>=0.12.0
  Downloading levenshtein-0.12.0-cp37-cp37m-manylinux1_x86_64.whl (158 kB)
     |████████████████████████████████| 158 kB 56.9 MB/s
Collecting wordcloud<2.0,>=1.8
  Downloading wordcloud-1.8.1-cp37-cp37m-manylinux1_x86_64.whl (366 kB)
     |████████████████████████████████| 366 kB 60.3 MB/s
Collecting metaphone<0.7,>=0.6
  Downloading Metaphone-0.6.tar.gz (14 kB)
Requirement already satisfied: ipywidgets<8.0,>=7.5 in /usr/local/lib/python3.7/dist-packages (from dataprep) (7.6.5)
Requirement already satisfied: tqdm<5.0,>=4.48 in /usr/local/lib/python3.7/dist-packages (from dataprep) (4.62.3)
Collecting regex<2021.0.0,>=2020.10.15
  Downloading regex-2020.11.13-cp37-cp37m-manylinux2014_x86_64.whl (719 kB)
     |████████████████████████████████| 719 kB 34.5 MB/s
Collecting pydantic<2.0,>=1.6
  Downloading pydantic-1.9.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (10.9 MB)
     |████████████████████████████████| 10.9 MB 56.3 MB/s
Collecting jsonpath-ng<2.0,>=1.5
```

```
  Downloading jsonpath_ng-1.5.3-py3-none-any.whl (29 kB)
Collecting python-stdnum<2.0,>=1.16
  Downloading python_stdnum-1.17-py2.py3-none-any.whl (943 kB)
     |████████████████████████████████| 943 kB 43.0 MB/s
Collecting nltk<4.0,>=3.5
  Downloading nltk-3.6.7-py3-none-any.whl (1.5 MB)
     |████████████████████████████████| 1.5 MB 38.6 MB/s
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (21.4.0)
Collecting multidict<7.0,>=4.5
  Downloading multidict-5.2.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (160 kB)
     |████████████████████████████████| 160 kB 70.3 MB/s
Collecting frozenlist>=1.1.1
  Downloading frozenlist-1.2.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (192 kB)
     |████████████████████████████████| 192 kB 73.8 MB/s
Collecting yarl<2.0,>=1.0
  Downloading yarl-1.7.2-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (271 kB)
     |████████████████████████████████| 271 kB 75.2 MB/s
Collecting aiosignal>=1.1.2
  Downloading aiosignal-1.2.0-py3-none-any.whl (8.2 kB)
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (2.0.10)
Collecting asynctest==0.13.0
  Downloading asynctest-0.13.0-py3-none-any.whl (26 kB)
Requirement already satisfied: typing-extensions>=3.7.4 in /usr/local/lib/python3.7/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (3.10.0.2)
Collecting async-timeout<5.0,>=4.0.0a3
  Downloading async_timeout-4.0.2-py3-none-any.whl (5.8 kB)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (3.13)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (5.1.1)
Requirement already satisfied: packaging>=16.8 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (21.3)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (2.8.2)
Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (7.1.2)
Requirement already satisfied: toolz>=0.8.2 in /usr/local/lib/python3.7/dist-packages (from dask[array,dataframe,delayed]<3.0,>=2.25->dataprep) (0.11.2)
Collecting partd>=0.3.10
  Downloading partd-1.2.0-py3-none-any.whl (19 kB)
Collecting fsspec>=0.6.0
  Downloading fsspec-2022.1.0-py3-none-any.whl (133 kB)
     |████████████████████████████████| 133 kB 44.0 MB/s
Requirement already satisfied: cloudpickle>=0.2.2 in /usr/local/lib/python3.7/dist-packages (from dask[array,dataframe,delayed]<3.0,>=2.25->dataprep) (1.3.0)
Requirement already satisfied: click<8.0,>=5.1 in /usr/local/lib/python3.7/dist-packages (from flask<2.0.0,>=1.1.4->dataprep) (7.1.2)
Requirement already satisfied: Werkzeug<2.0,>=0.15 in /usr/local/lib/python3.7/dist-packages (from flask<2.0.0,>=1.1.4->dataprep) (1.0.1)
Requirement already satisfied: itsdangerous<2.0,>=0.24 in /usr/local/lib/python3.7/dist-packages (from flask<2.0.0,>=1.1.4->dataprep) (1.1.0)
```

Requirement already satisfied: Six in /usr/local/lib/python3.7/dist-packages (from flask_cors<4.0.0,>=3.0.10->dataprep) (1.15.0)
Requirement already satisfied: widgetsnbextension~=3.5.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (3.5.2)
Requirement already satisfied: nbformat>=4.2.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (5.1.3)
Requirement already satisfied: ipython-genutils~=0.2.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (0.2.0)
Requirement already satisfied: traitlets>=4.3.1 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (5.1.1)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (1.0.2)
Requirement already satisfied: ipython>=4.0.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (5.5.0)
Requirement already satisfied: ipykernel>=4.5.1 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (4.10.1)
Requirement already satisfied: jupyter-client in /usr/local/lib/python3.7/dist-packages (from ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (5.3.5)
Requirement already satisfied: prompt-toolkit<2.0.0,>=1.0.4 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (1.0.18)
Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (57.4.0)
Requirement already satisfied: simplegeneric>0.8 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.8.1)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.5)
Requirement already satisfied: pygments in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (2.6.1)
Requirement already satisfied: pexpect in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (4.8.0)
Requirement already satisfied: decorator in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (4.4.2)
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/dist-packages (from jinja2<3.0,>=2.11->dataprep) (2.0.1)
Collecting ply
Downloading ply-3.11-py2.py3-none-any.whl (49 kB)
|████████████████████████████████| 49 kB 5.3 MB/s
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (4.9.1)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (4.3.3)
Requirement already satisfied: importlib-resources>=1.4.0 in /usr/local/lib/python3.7/dist-packages (from jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (5.4.0)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (4.10.0)
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in /usr/local/lib/python3.7/dist-packages (from jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (0.18.0)
Requirement already satisfied: zipp>=3.1.0 in /usr/local/lib/python3.7/dist-packages (from importlib-resources>=1.4.0->jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (3.7.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from nltk<4.0,>=3.5->dataprep) (1.1.0)
Collecting nltk<4.0,>=3.5
Downloading nltk-3.6.6-py3-none-any.whl (1.5 MB)
|████████████████████████████████| 1.5 MB 43.5 MB/s
Downloading nltk-3.6.5-py3-none-any.whl (1.5 MB)
|████████████████████████████████| 1.5 MB 27.5 MB/s
Downloading nltk-3.6.3-py3-none-any.whl (1.5 MB)
|████████████████████████████████| 1.5 MB 54.6 MB/s

```
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging>=16.8->bokeh<3,>=2->dataprep) (3.0.6)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (from pandas<2.0,>=1.1->dataprep) (2018.9)
Collecting locket
  Downloading locket-0.2.1-py2.py3-none-any.whl (4.1 kB)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.7/dist-packages (from prompt-toolkit<2.0.0,>=1.0.4->ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.5)
Collecting python-crfsuite>=0.7
  Downloading python_crfsuite-0.9.7-cp37-cp37m-manylinux1_x86_64.whl (743 kB)
     |████████████████████████████████| 743 kB 48.3 MB/s
Requirement already satisfied: future>=0.14 in /usr/local/lib/python3.7/dist-packages (from usaddress<0.6.0,>=0.5.10->dataprep) (0.16.0)
Collecting probableparsing
  Downloading probableparsing-0.0.1-py2.py3-none-any.whl (3.1 kB)
Collecting asttokens<3.0.0,>=2.0.0
  Downloading asttokens-2.0.5-py2.py3-none-any.whl (20 kB)
Collecting executing
  Downloading executing-0.8.2-py2.py3-none-any.whl (16 kB)
Collecting pure_eval<1.0.0
  Downloading pure_eval-0.2.1-py3-none-any.whl (11 kB)
Requirement already satisfied: notebook>=4.4.1 in /usr/local/lib/python3.7/dist-packages (from widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (5.3.1)
Requirement already satisfied: Send2Trash in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.8.0)
Requirement already satisfied: terminado>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.12.1)
Requirement already satisfied: nbconvert in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (5.6.1)
Requirement already satisfied: pyzmq>=13 in /usr/local/lib/python3.7/dist-packages (from jupyter-client->ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (22.3.0)
Requirement already satisfied: ptyprocess in /usr/local/lib/python3.7/dist-packages (from terminado>=0.8.1->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.0)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from wordcloud<2.0,>=1.8->dataprep) (3.2.2)
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.7/dist-packages (from yarl<2.0,>=1.0->aiohttp<4.0,>=3.6->dataprep) (2.10)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib->wordcloud<2.0,>=1.8->dataprep) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->wordcloud<2.0,>=1.8->dataprep) (1.3.2)
Requirement already satisfied: entrypoints>=0.2.2 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.3)
Requirement already satisfied: bleach in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (4.1.0)
Requirement already satisfied: testpath in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.5.0)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.1)
Requirement already satisfied: pandocfilters>=1.4.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.5.0)
Requirement already satisfied: mistune<2,>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.8.4)
Requirement already satisfied: webencodings in /usr/local/lib/python3.7/dist-packages (from bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.5.1)
Building wheels for collected packages: metaphone
  Building wheel for metaphone (setup.py) ... done
```

Created wheel for metaphone: filename=Metaphone-0.6-py3-none-any.whl size=13918 sha256=b27e95f0d9f1c2e38167df62d9db51d9a1e83b745a36ca94bb69329b4738be3d
Stored in directory: /root/.cache/pip/wheels/1d/a8/cb/6f8902aa5457bd71344e00665c230e9c45255b3f57f2194a0f
Successfully built metaphone
Installing collected packages: multidict, locket, frozenlist, yarl, regex, python-crfsuite, pure-eval, probableparsing, ply, partd, fsspec, executing, dask, asynctest, async-timeout, asttokens, aiosignal, wordcloud, varname, usaddress, python-stdnum, pydantic, nltk, metaphone, levenshtein, jsonpath-ng, flask-cors, aiohttp, dataprep
Attempting uninstall: regex
Found existing installation: regex 2019.12.20
Uninstalling regex-2019.12.20:
Successfully uninstalled regex-2019.12.20
Attempting uninstall: dask
Found existing installation: dask 2.12.0
Uninstalling dask-2.12.0:
Successfully uninstalled dask-2.12.0
Attempting uninstall: wordcloud
Found existing installation: wordcloud 1.5.0
Uninstalling wordcloud-1.5.0:
Successfully uninstalled wordcloud-1.5.0
Attempting uninstall: nltk
Found existing installation: nltk 3.2.5
Uninstalling nltk-3.2.5:
Successfully uninstalled nltk-3.2.5
Successfully installed aiohttp-3.8.1 aiosignal-1.2.0 asttokens-2.0.5 async-timeout-4.0.2 asynctest-0.13.0 dask-2.30.0 dataprep-0.4.1 executing-0.8.2 flask-cors-3.0.10 frozenlist-1.2.0 fsspec-2022.1.0 jsonpath-ng-1.5.3 levenshtein-0.12.0 locket-0.2.1 metaphone-0.6 multidict-5.2.0 nltk-3.6.3 partd-1.2.0 ply-3.11 probableparsing-0.0.1 pure-eval-0.2.1 pydantic-1.9.0 python-crfsuite-0.9.7 python-stdnum-1.17 regex-2020.11.13 usaddress-0.5.10 varname-0.8.1 wordcloud-1.8.1 yarl-1.7.2

[ ]

```
sns.pairplot(df,diag_kind='kde',hue='stroke')
```

[ ]

```
#Distribution of Targets
from dataprep.eda import plot
plot(df, 'stroke')#From distribution it is clear dataset has highly unbalanced data distribution.
```

Who is more susceptible to infection stroke , women or men?

[ ]

```
# which gender is the most infection to stroke
```

```
sns.countplot(df.gender,hue='stroke',data=df)
```

From the figure above, I think that men have a higher risk of stroke than women.

[]

```
#How does age affect stroke risk?
plt.figure(figsize=[10,8])
sns.countplot(df.age,hue='stroke',data=df)
```

From age features it can be seen that old age people are mostly having strokes, compared to younger ones.

Does smoking affect strokes?

[]

```
# How does smoke affect stroke risk?
sns.countplot(df.smoking_status,hue='stroke',data=df)
```

As we can see, excessive smoking may increase the risk of stroke.

preprocessing for modeling

[]

```
#label encoder
# Convert each of ' Gender,Residence_type and Marrital Status' in
to 0 & 1
df['gender']=df['gender'].apply(lambda x : 1 if x=='Male' else 0)
```

```
df["Residence_type"] = df["Residence_type"].apply(lambda x: 1 if x=="Urban" else 0)
df["ever_married"] = df["ever_married"].apply(lambda x: 1 if x=="Yes" else 0)
df
```

[]

```
#label encoder
#i will use OneHot encoding for smoking_status, work_type coulmns.
data_dummies = df[['smoking_status','work_type']]
data_dummies=pd.get_dummies(data_dummies)
```

[]

```
#remove tha 'smoking_status,work_type' features and reblace it with dummies coulmns
df.drop(columns=['smoking_status','work_type'],inplace=True)
print("data_dummies")
df.merge(data_dummies,left_index=True, right_index=True,how='left')
```

Split Data

[]

```
# detect input and output
X = df.drop('stroke',axis=1)
y = df.stroke
print(X.shape)
print(y.shape)
```

(5110, 8)
(5110,)

[ ]

```
# train test split
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.
2,random_state=42)
print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)
```

(4088, 8)
(4088,)
(1022, 8)
(1022,)

[ ]

```
# the dataset is embalanced, I will use SMOTE
from imblearn.combine import SMOTETomek
from collections import Counter
print("The number of classes before fit {}".format(Counter(y_trai
n)))
smot =SMOTETomek()
X_train,y_train = smot.fit_resample(X_train,y_train)
print("The number of classes after fit {}".format(Counter(y_train
)))
```

The number of classes before fit Counter({0: 3901, 1: 187})
The number of classes after fit Counter({0: 3845, 1: 3845})

[ ]

```
# feature scaling
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

[ ]

```
print("X_test_scaled:   ", X_test_scaled)
print("X_train_scaled:  ",X_train_scaled)
```

```
X_test_scaled:   [[ 1.50190396 -1.09365377 -0.29742942 ... -0.82938779 -1.00041085
  -0.95557473]
 [ 1.50190396 -0.68728137 -0.29742942 ... -0.82938779 -0.99220527
  -0.16102234]
 [-0.66582154 -2.13216099 -0.29742942 ...  1.20570861 -0.82193958
  -1.0305325 ]
 ...
 [ 1.50190396 -0.28090898  3.36214217 ... -0.82938779  0.01502913
   0.1538003 ]
 [-0.66582154  0.03515843 -0.29742942 ...  1.20570861  0.18902463
  -0.67073519]
 [ 1.50190396  1.02851317  3.36214217 ... -0.82938779 -0.22535693
  -0.31093789]]
X_train_scaled:  [[ 1.50190396  1.07366566 -0.29742942 ... -0.82938779 -0.10917344
  -0.13103923]
 [-0.66582154  0.30607336 -0.29742942 ...  1.20570861 -0.56271799
   1.03830201]
 [-0.66582154 -1.54517865 -0.29742942 ... -0.82938779 -1.09981021
   0.6485216 ]
 ...
 [-0.66582154  0.80290101 -0.29742942 ... -0.82938779  2.16019825
   1.19060575]
 [-0.66582154  0.92371487 -0.29742942 ... -0.82938779  0.95078666
  -0.39740789]
 [-0.66582154  0.96924578 -0.29742942 ...  1.20570861 -0.93507291
   0.83243359]]
```

Modeling

[ ]

```
# apply logistic regression
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
lr = LogisticRegression()
lr.fit(X_train_scaled,y_train)
y_pred = lr.predict(X_test_scaled)
```

```
print(classification_report(y_pred,y_test))
```

```
              precision    recall  f1-score   support

           0       0.79      0.97      0.87       781
           1       0.63      0.16      0.26       241

    accuracy                           0.78      1022
   macro avg       0.71      0.57      0.56      1022
weighted avg       0.75      0.78      0.73      1022
```

[ ]

```
#apply KNN
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report
knn = KNeighborsClassifier(n_neighbors = 7)
knn.fit(X_train_scaled,y_train)
y_pred_knn = knn.predict(X_test_scaled)
print(classification_report(y_pred_knn,y_test))
```

```
              precision    recall  f1-score   support

           0       0.82      0.96      0.89       828
           1       0.42      0.13      0.20       194

    accuracy                           0.80      1022
   macro avg       0.62      0.55      0.54      1022
weighted avg       0.75      0.80      0.76      1022
```

[ ]

```
# applay Random Forest
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
rf= RandomForestClassifier()
rf.fit(X_train_scaled,y_train)
y_pred_rf= rf.predict(X_test_scaled)
print(classification_report(y_pred_rf,y_test))
```

```
              precision    recall  f1-score   support

           0       0.93      0.95      0.94       941
           1       0.24      0.19      0.21        81
```

```
   accuracy                           0.89      1022
  macro avg       0.59      0.57      0.58      1022
weighted avg       0.88      0.89      0.88      1022
```

Model Evaluation

[ ]

```
#comparing between the models
print("logistic regression:",classification_report(y_pred,y_test))
print("KNN:",classification_report(y_pred_knn,y_test))
print("Random Forest:",classification_report(y_pred_rf,y_test))
```

```
logistic regression:               precision    recall  f1-score   support

           0       0.79      0.97      0.87       781
           1       0.63      0.16      0.26       241

    accuracy                           0.78      1022
   macro avg       0.71      0.57      0.56      1022
weighted avg       0.75      0.78      0.73      1022

KNN:               precision    recall  f1-score   support

           0       0.82      0.96      0.89       828
           1       0.42      0.13      0.20       194

    accuracy                           0.80      1022
   macro avg       0.62      0.55      0.54      1022
weighted avg       0.75      0.80      0.76      1022

Random Forest:               precision    recall  f1-score   support

           0       0.93      0.95      0.94       941
           1       0.24      0.19      0.21        81

    accuracy                           0.89      1022
   macro avg       0.59      0.57      0.58      1022
weighted avg       0.88      0.89      0.88      1022
```

---

by compare the results of the different models, i can say RF have the best result,then KNN model and the last one is logistic regression.

---

conclusion

Short summary - if it is important for us to identify all people who may have risk a stroke the best to cope with this task with Random Forest.

---

excuse me am work on colab and I was trying to save it as pdf but I can't. I written everything in the .ipynb file also ,you can see all the details there.