stroke_risk_analysis

June 22, 2024

[ ]:
```
import pandas as pd
import numpy as np
from scipy import stats
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
import joblib
from stroke_risk_utils import *
```

[ ]:
```
stroke_df = pd.read_csv("stroke_dataset.csv")
stroke_df.head()
```

[ ]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

[ ]:
```
stroke_df.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
```

[ ]:
```
print(stroke_df.isnull().sum())
```

```
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
```

This dataset contains 5110 entries and 12 columns related to potential stroke risk factors.

Observations:

- **Data Types:** Includes numerical (int64, float64) and categorical (object) features.
- **Missing Values:** The `bmi` column has 201 missing values.
- **Potential Features:** Age, health conditions (hypertension, heart disease), lifestyle factors (smoking, marriage, work type, residence), and glucose/BMI levels could be predictive.
- **Target Variable:** The `stroke` column (likely binary: 0 or 1) is the target for prediction.

Next Steps:

1. **Data Cleaning:**
   - Rename columns for consistency (using lowercase and underscores).

- Address missing values in `bmi` (dropping rows for simplicity in this case as it only contains 4% of the dataset).

[ ]:
```
stroke_df = stroke_df.rename(columns={'Residence_type': 'residence_type'})
```

[ ]:
```
stroke_df = stroke_df.dropna(subset=['bmi'])
```

[ ]:
```
numerical_features = ['age', 'avg_glucose_level', 'bmi']
categorical_features = ['gender', 'hypertension', 'heart_disease',␣
↪'ever_married', 'work_type', 'residence_type', 'smoking_status']
```

[ ]:
```
plot_combined_histograms(stroke_df, numerical_features)
plot_combined_boxplots(stroke_df, numerical_features)
```