```
import pandas as pd
import matplotlib as plt
import seaborn as sns

df = pd.read_csv(r"C:\Users\susmi\Desktop\test_counter\ml_code\
hypothyroid.csv")

df.head()
```

```
  age sex on thyroxine query on thyroxine on antithyroid medication
sick  \
0  41   F            f                  f                          f
f
1  23   F            f                  f                          f
f
2  46   M            f                  f                          f
f
3  70   F            t                  f                          f
f
4  70   F            f                  f                          f
f

  pregnant thyroid surgery I131 treatment query hypothyroid  ... TT4
measured  \
0        f               f              f                 f  ...
t
1        f               f              f                 f  ...
t
2        f               f              f                 f  ...
t
3        f               f              f                 f  ...
t
4        f               f              f                 f  ...
t

   TT4 T4U measured   T4U FTI measured  FTI TBG measured TBG referral
source  \
0  125            t  1.14            t  109            f   ?
SVHC
1  102            f     ?            f    ?            f   ?
other
2  109            t  0.91            t  120            f   ?
other
3  175            f     ?            f    ?            f   ?
other
4   61            t  0.87            t   70            f   ?
SVI

  binaryClass
0           P
```

```
1           P
2           P
3           P
4           P

[5 rows x 30 columns]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3772 entries, 0 to 3771
Data columns (total 30 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   age                        3772 non-null   object
 1   sex                        3772 non-null   object
 2   on thyroxine               3772 non-null   object
 3   query on thyroxine         3772 non-null   object
 4   on antithyroid medication  3772 non-null   object
 5   sick                       3772 non-null   object
 6   pregnant                   3772 non-null   object
 7   thyroid surgery            3772 non-null   object
 8   I131 treatment             3772 non-null   object
 9   query hypothyroid          3772 non-null   object
 10  query hyperthyroid         3772 non-null   object
 11  lithium                    3772 non-null   object
 12  goitre                     3772 non-null   object
 13  tumor                      3772 non-null   object
 14  hypopituitary              3772 non-null   object
 15  psych                      3772 non-null   object
 16  TSH measured               3772 non-null   object
 17  TSH                        3772 non-null   object
 18  T3 measured                3772 non-null   object
 19  T3                         3772 non-null   object
 20  TT4 measured               3772 non-null   object
 21  TT4                        3772 non-null   object
 22  T4U measured               3772 non-null   object
 23  T4U                        3772 non-null   object
 24  FTI measured               3772 non-null   object
 25  FTI                        3772 non-null   object
 26  TBG measured               3772 non-null   object
 27  TBG                        3772 non-null   object
 28  referral source            3772 non-null   object
 29  binaryClass                3772 non-null   object
dtypes: object(30)
memory usage: 884.2+ KB
```

```
df['TBG'].unique()
```

```
array(['?'], dtype=object)
```

```
df = df.drop('referral source', axis =1)

df.head()
```

```
  age sex on thyroxine query on thyroxine on antithyroid medication
sick  \
0  41   F            f                  f                          f
f
1  23   F            f                  f                          f
f
2  46   M            f                  f                          f
f
3  70   F            t                  f                          f
f
4  70   F            f                  f                          f
f

  pregnant thyroid surgery I131 treatment query hypothyroid  ...   T3
\
0        f               f              f                 f  ...  2.5

1        f               f              f                 f  ...    2

2        f               f              f                 f  ...    ?

3        f               f              f                 f  ...  1.9

4        f               f              f                 f  ...  1.2


  TT4 measured  TT4 T4U measured   T4U FTI measured  FTI TBG measured
TBG  \
0            t  125            t  1.14            t  109            f
?
1            t  102            f     ?            f    ?            f
?
2            t  109            t  0.91            t  120            f
?
3            t  175            f     ?            f    ?            f
?
4            t   61            t  0.87            t   70            f
?

  binaryClass
0           P
1           P
2           P
3           P
4           P

[5 rows x 29 columns]
```

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report

numeric_columns = ['age', 'TSH', 'T3', 'TT4', 'T4U', 'FTI', 'TBG']
df[numeric_columns] = df[numeric_columns].apply(pd.to_numeric, errors
= 'coerce')

binary_columns = ['on thyroxine', 'query on thyroxine', 'on
antithyroid medication', 'sick', 'pregnant', 'thyroid surgery',
                  'I131 treatment', 'query hypothyroid', 'query
hyperthyroid', 'lithium', 'goitre', 'tumor',
                  'hypopituitary', 'psych', 'TSH measured', 'T3
measured', 'TT4 measured', 'T4U measured',
                  'FTI measured', 'TBG measured', 'binaryClass']

le = LabelEncoder()
df[binary_columns] = df[binary_columns].apply(le.fit_transform)

df['sex'] = le.fit_transform(df['sex'])

x = df.drop('binaryClass', axis =1)
y = df['binaryClass']

x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size=0.2, random_state=42)

x_train.head()
```

```
       age  sex  on thyroxine  query on thyroxine  on antithyroid
medication  \
2613  60.0    2             0                   0
0
249   25.0    1             0                   0
0
2663  19.0    1             0                   0
0
3068  71.0    2             1                   0
0
1705  52.0    1             0                   0
0

      sick  pregnant  thyroid surgery  I131 treatment  query
hypothyroid  ...  \
2613     0         0                0               0
0  ...
249      0         0                0               0
0  ...
2663     0         0                0               0
```

```
1  ...
3068     0         0                0                0
1  ...
1705     0         0                0                0
1  ...

          TSH  T3 measured   T3  TT4 measured    TT4  T4U measured
T4U  \
2613   0.005            1  7.1             1  184.0              1
1.03
249    1.600            1  5.4             1  152.0              1
1.50
2663   3.700            1  1.1             1  126.0              1
0.91
3068  25.000            1  0.3             1   31.0              1
0.68
1705  35.000            1  2.1             1   77.0              1
1.10

      FTI measured    FTI  TBG measured
2613             1  177.0             0
249              1  102.0             0
2663             1  138.0             0
3068             1   46.0             0
1705             1   70.0             0

[5 rows x 27 columns]

df.head()

    age  sex  on thyroxine  query on thyroxine  on antithyroid
medication  \
0  41.0    1             0                   0
0
1  23.0    1             0                   0
0
2  46.0    2             0                   0
0
3  70.0    1             1                   0
0
4  70.0    1             0                   0
0

   sick  pregnant  thyroid surgery  I131 treatment  query hypothyroid
...  \
0     0         0                0               0                   0
...
1     0         0                0               0                   0
...
2     0         0                0               0                   0
```

```
...
3     0         0                0               0                   0
...
4     0         0                0               0                   0
...

    T3  TT4 measured    TT4  T4U measured   T4U  FTI measured
FTI  \
0  2.5              1  125.0             1  1.14             1  109.0

1  2.0              1  102.0             0   NaN             0    NaN

2  NaN              1  109.0             1  0.91             1  120.0

3  1.9              1  175.0             0   NaN             0    NaN

4  1.2              1   61.0             1  0.87             1   70.0


   TBG measured  TBG  binaryClass
0             0  NaN            1
1             0  NaN            1
2             0  NaN            1
3             0  NaN            1
4             0  NaN            1

[5 rows x 29 columns]

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3772 entries, 0 to 3771
Data columns (total 29 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   age                        3771 non-null   float64
 1   sex                        3772 non-null   int32
 2   on thyroxine               3772 non-null   int32
 3   query on thyroxine         3772 non-null   int32
 4   on antithyroid medication  3772 non-null   int32
 5   sick                       3772 non-null   int32
 6   pregnant                   3772 non-null   int32
 7   thyroid surgery            3772 non-null   int32
 8   I131 treatment             3772 non-null   int32
 9   query hypothyroid          3772 non-null   int32
 10  query hyperthyroid         3772 non-null   int32
 11  lithium                    3772 non-null   int32
 12  goitre                     3772 non-null   int32
 13  tumor                      3772 non-null   int32
 14  hypopituitary              3772 non-null   int32
```

```
 15  psych                      3772 non-null   int32
 16  TSH measured               3772 non-null   int32
 17  TSH                        3403 non-null   float64
 18  T3 measured                3772 non-null   int32
 19  T3                         3003 non-null   float64
 20  TT4 measured               3772 non-null   int32
 21  TT4                        3541 non-null   float64
 22  T4U measured               3772 non-null   int32
 23  T4U                        3385 non-null   float64
 24  FTI measured               3772 non-null   int32
 25  FTI                        3387 non-null   float64
 26  TBG measured               3772 non-null   int32
 27  TBG                        0 non-null      float64
 28  binaryClass                3772 non-null   int32
dtypes: float64(7), int32(22)
memory usage: 530.6 KB
```

```
df = df.drop('TBG', axis =1)
```

```
df = df.dropna()
```

```
clf = RandomForestClassifier(random_state = 42)
clf.fit(x_train, y_train)
```

```
RandomForestClassifier(random_state=42)
```

```
y_pred = clf.predict(x_test)
```

```
accuracy = accuracy_score(y_test, y_pred)
print(f'accuracy_test: {accuracy}')
```

```
accuracy_test: 0.9927404718693285
```

```
report = classification_report(y_test, y_pred)
print(f'Classification Report:\n{report}')
```

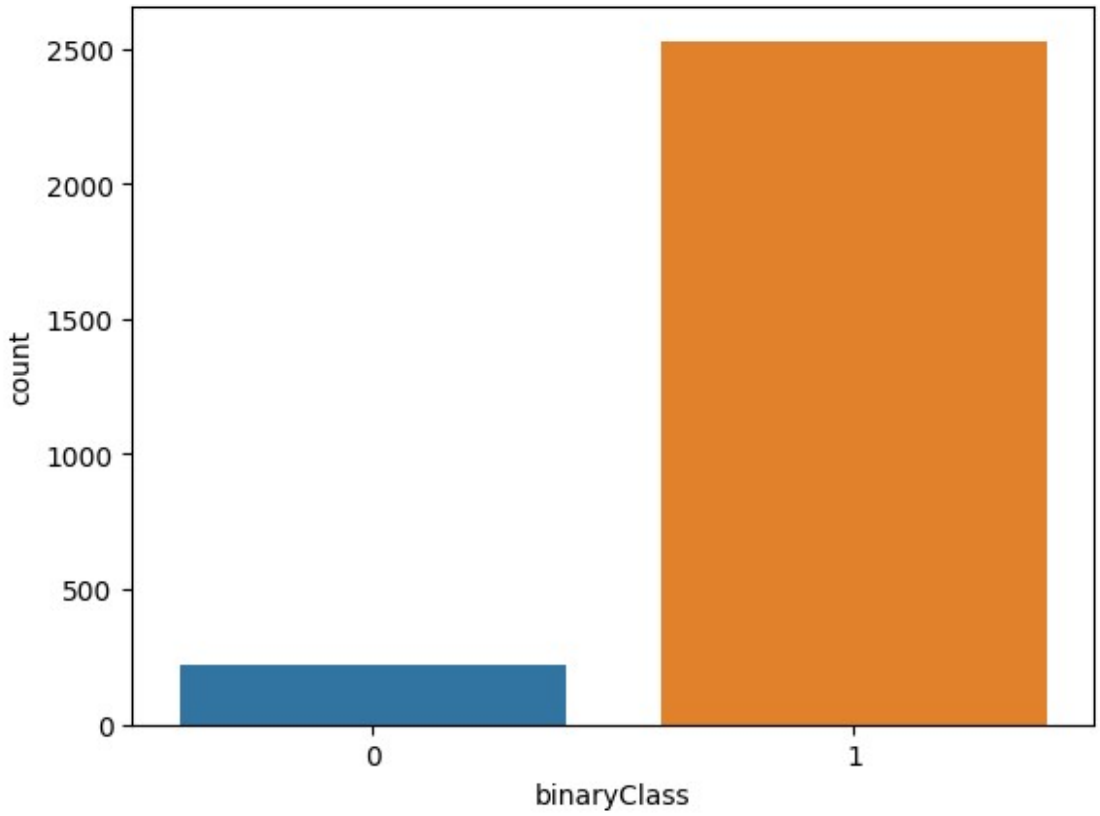
```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        37
           1       1.00      0.99      1.00       514

    accuracy                           0.99       551
   macro avg       0.95      1.00      0.97       551
weighted avg       0.99      0.99      0.99       551
```

```
cm = confusion_matrix(y_test, y_pred)
print(f'Confusion Matrix:\n{cm}')
```

```
Confusion Matrix:
[[ 37   0]
 [  4 510]]
```

```
sns.countplot(x= 'binaryClass', data = df)
```

```
<Axes: xlabel='binaryClass', ylabel='count'>
```




```
df['binaryClass'].value_counts()
```

```
binaryClass
1    2528
0     223
Name: count, dtype: int64
```

```
affected_rate = 223/2751
```

```
affected_rate
```

```
0.08106143220647037
```

DATA_VISUALIZATION

```
df.info()
```

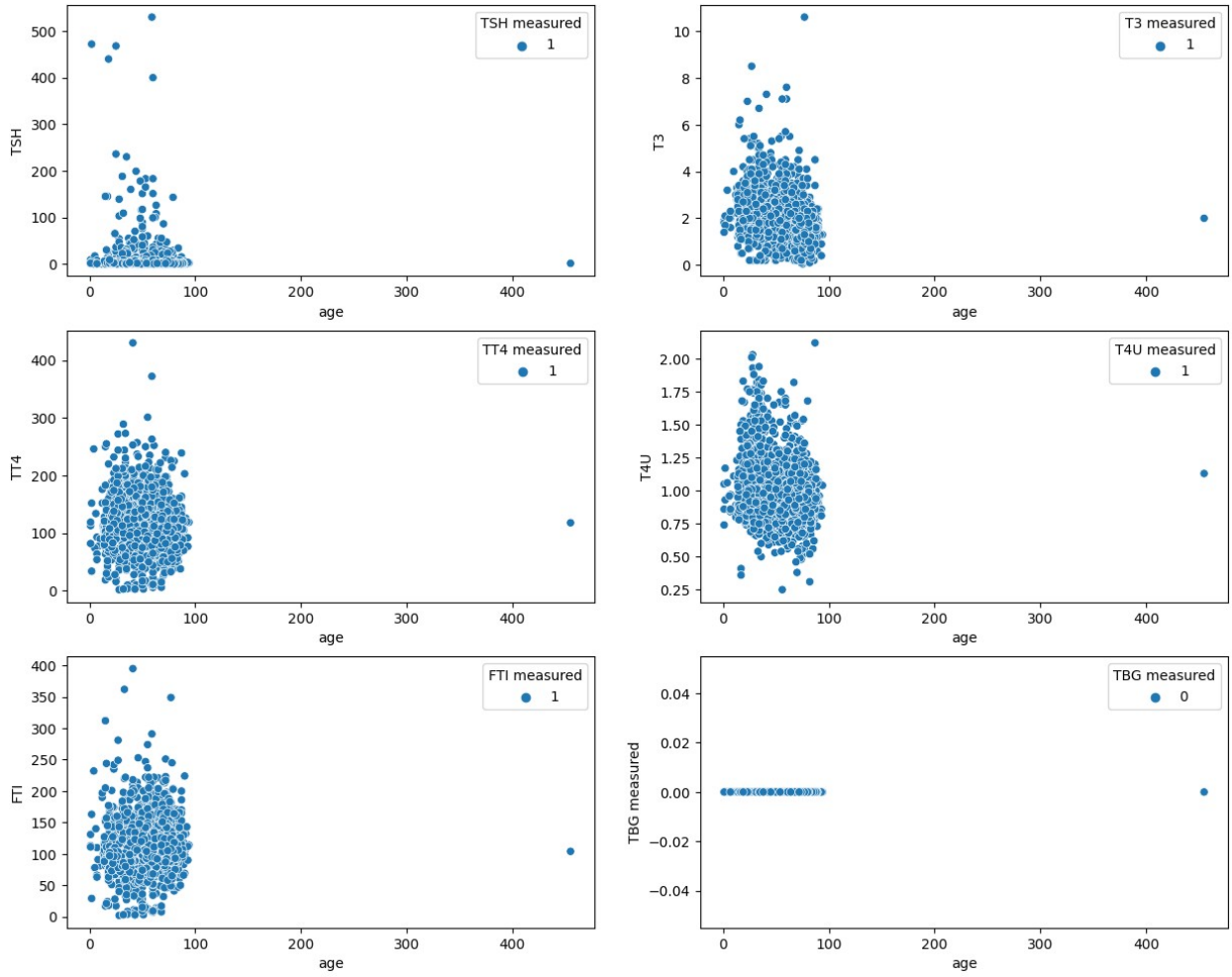
```
<class 'pandas.core.frame.DataFrame'>
Index: 2751 entries, 0 to 3771
Data columns (total 28 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   age                        2751 non-null   float64
 1   sex                        2751 non-null   int32  
 2   on thyroxine               2751 non-null   int32  
 3   query on thyroxine         2751 non-null   int32  
 4   on antithyroid medication  2751 non-null   int32  
 5   sick                       2751 non-null   int32  
 6   pregnant                   2751 non-null   int32  
 7   thyroid surgery            2751 non-null   int32  
 8   I131 treatment             2751 non-null   int32  
 9   query hypothyroid          2751 non-null   int32  
 10  query hyperthyroid         2751 non-null   int32  
 11  lithium                    2751 non-null   int32  
 12  goitre                     2751 non-null   int32  
 13  tumor                      2751 non-null   int32  
 14  hypopituitary              2751 non-null   int32  
 15  psych                      2751 non-null   int32  
 16  TSH measured               2751 non-null   int32  
 17  TSH                        2751 non-null   float64
 18  T3 measured                2751 non-null   int32  
 19  T3                         2751 non-null   float64
 20  TT4 measured               2751 non-null   int32  
 21  TT4                        2751 non-null   float64
 22  T4U measured               2751 non-null   int32  
 23  T4U                        2751 non-null   float64
 24  FTI measured               2751 non-null   int32  
 25  FTI                        2751 non-null   float64
 26  TBG measured               2751 non-null   int32  
 27  binaryClass                2751 non-null   int32  
dtypes: float64(6), int32(22)
memory usage: 386.9 KB
```

```
df.head(1)
```

```
    age  sex  on thyroxine  query on thyroxine  on antithyroid
medication  \
0  41.0    1             0                   0
0

   sick  pregnant  thyroid surgery  I131 treatment  query hypothyroid
...  \
0     0         0                0               0                   0
```

```
...

   T3 measured   T3  TT4 measured    TT4  T4U measured   T4U  FTI
measured  \
0            1  2.5             1  125.0             1  1.14
1

     FTI  TBG measured  binaryClass
0  109.0             0            1

[1 rows x 28 columns]
```

```
thyroid_data = df[['TSH measured', 'TSH', 'T3 measured', 'T3', 'TT4
measured', 'TT4', 'T4U measured', 'T4U', 'FTI measured', 'FTI', 'TBG
measured', 'age']]



def func_plot(col):
    fig, axes = plt.subplots(nrows=3, ncols=2, figsize=(15, 12))
    sns.scatterplot(x=col, y='TSH', hue='TSH measured',
data=thyroid_data, ax=axes[0, 0])
    sns.scatterplot(x=col, y='T3', hue='T3 measured',
data=thyroid_data, ax=axes[0, 1])
    sns.scatterplot(x=col, y='TT4', hue='TT4 measured',
data=thyroid_data, ax=axes[1, 0])
    sns.scatterplot(x=col, y='T4U', hue='T4U measured',
data=thyroid_data, ax=axes[1, 1])
    sns.scatterplot(x=col, y='FTI', hue='FTI measured',
data=thyroid_data, ax=axes[2, 0])
    sns.scatterplot(x=col, y='TBG measured', hue='TBG measured',
data=thyroid_data, ax=axes[2, 1])


func_plot("age")
plt.show()
```




```
df.head(0)
```

```
Empty DataFrame
Columns: [age, sex, on thyroxine, query on thyroxine, on antithyroid
medication, sick, pregnant, thyroid surgery, I131 treatment, query
hypothyroid, query hyperthyroid, lithium, goitre, tumor,
hypopituitary, psych, TSH measured, TSH, T3 measured, T3, TT4
measured, TT4, T4U measured, T4U, FTI measured, FTI, TBG measured,
binaryClass]
Index: []

[0 rows x 28 columns]
```