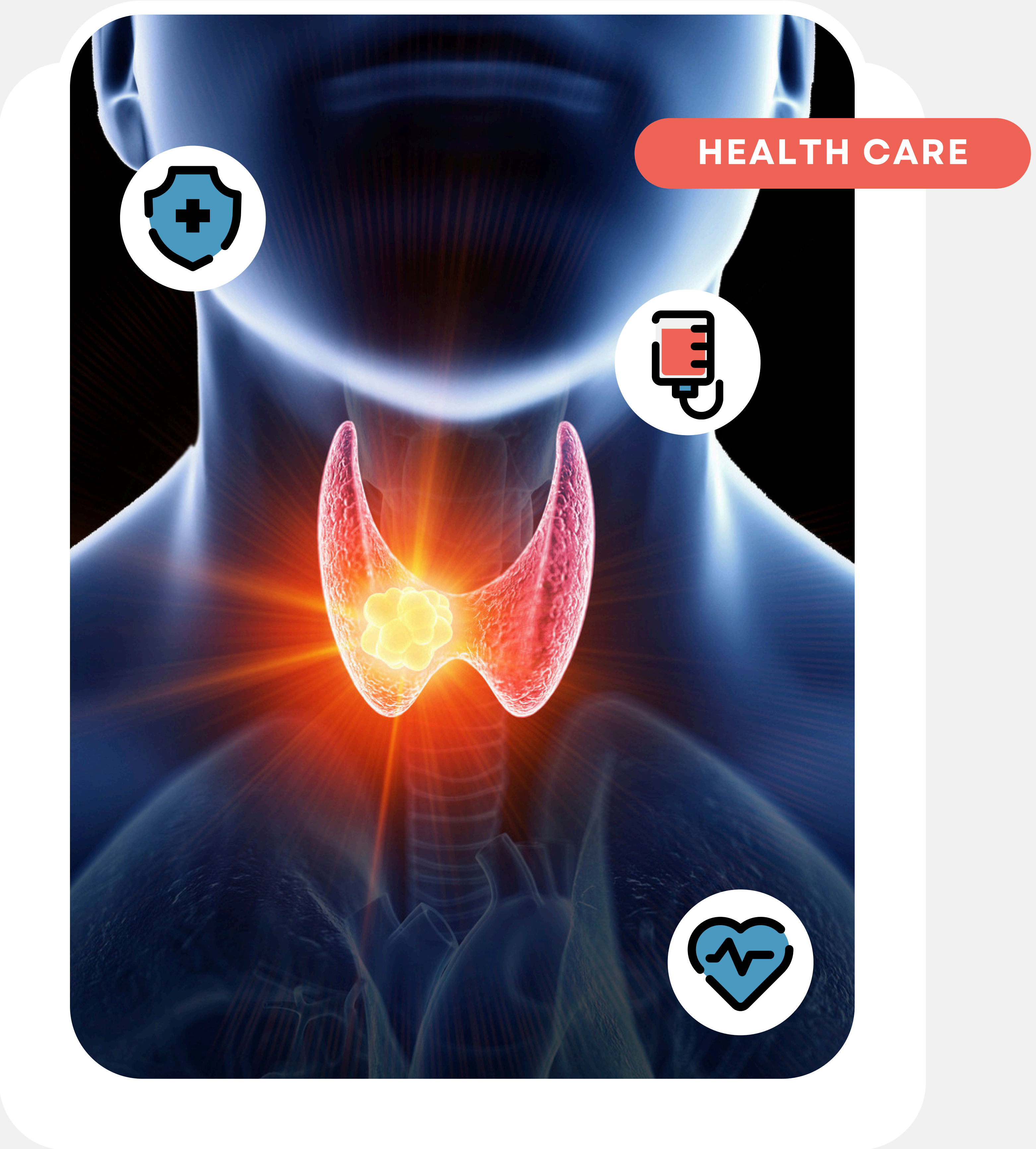
THYROID CANCER RISK DATASET

DATA PREPARATION AND DATA MINING

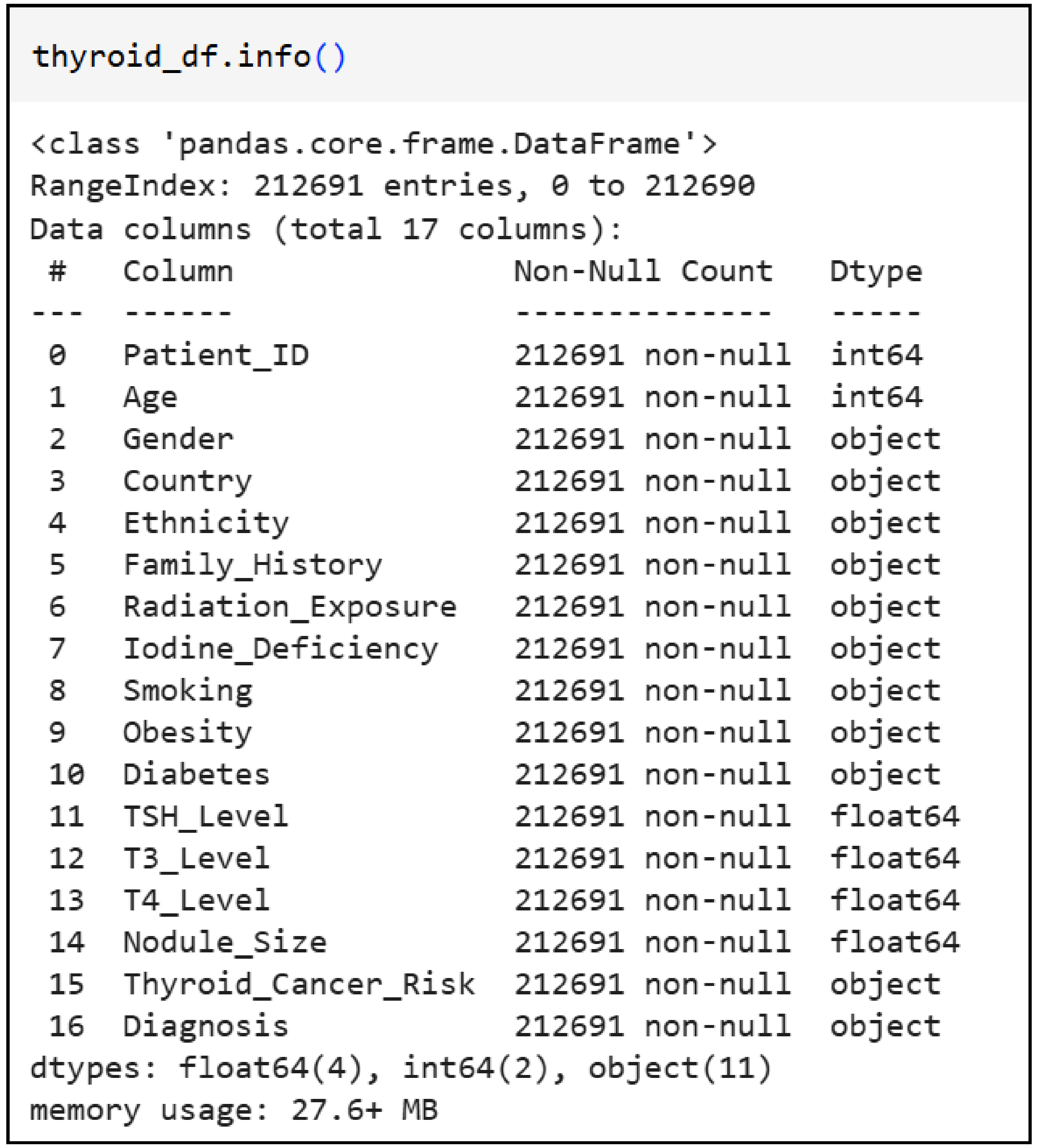


BY NOW WANT TO PRESS GRAB

ABOUT DATASET

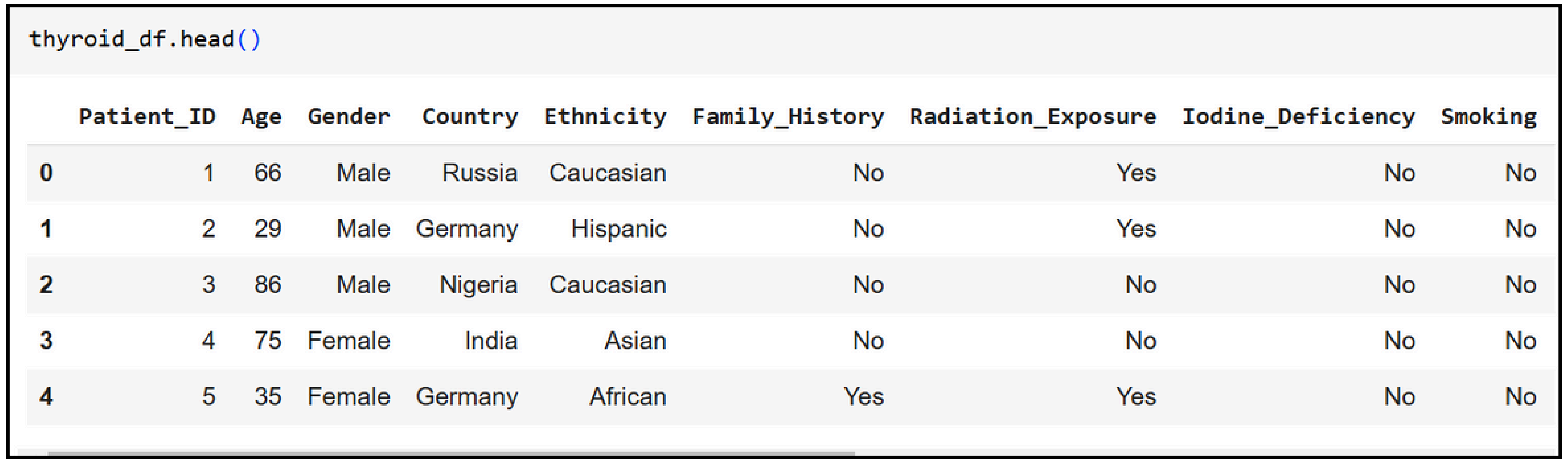
ตัวแปร	คำอธิบายข้อมูล	ตัวแปร	คำอธิบายข้อมูล
Patient_ID	หมายเลขระบุผู้ป่วยแต่ละราย	Radition_Exposure	ประวัติการได้รับรังสี (มี/ไม่มี)
Age	อายุของผู้ป่วย	Iodine_Deficiency	การมีภาวะขาดไอโอดีน (มี/ไม่มี)
Gender	เพศของผู้ป่วย (ชาย/หญิง)	Smoking	ผู้ป่วยสูบบุหรี่หรือไม่ (ใช่/ไม่ใช่)
Country	ประเทศที่อยู่อาศัย	Obesity	ผู้ป่วยเป็นโรคอ้วนหรือไม่ (ใช่/ไม่ใช่)
Ethnicity	สัญชาติของผู้ป่วย	Diabetes	ผู้ป่วยเป็นโรคเบาหวานหรือไม่ (ใช่/ไม่ใช่)
Family_History	ผู้ป่วยมีประวัติครอบครัวเป็นมะเร็งต่อมไทรอยด์หรือไม่ (มี/ไม่มี)	TSH_Level	ระดับฮอร์โมนกระตุ้นต่อมไทรอยด์ (µIU/mL)
T3_Level	ระดับไตรไอโอโดไทรโอนีน (ng/dL)	T4_Level	ระดับไทรอกซิน (µg/dL)
Nodule_Size	ขนาดของก้อนเนื้อในต่อมไทรอยด์ (ซม.)	Thyroid_Cancer_Risk	ความเสี่ยงโดยประมาณของมะเร็งต่อมไทรอยด์ (ต่ำ/กลาง/สูง)
Diagnosis	การวินิจฉัยขั้นสุดท้าย (ไม่ร้ายแรง/ร้ายแรง		

DATA PREPARATION



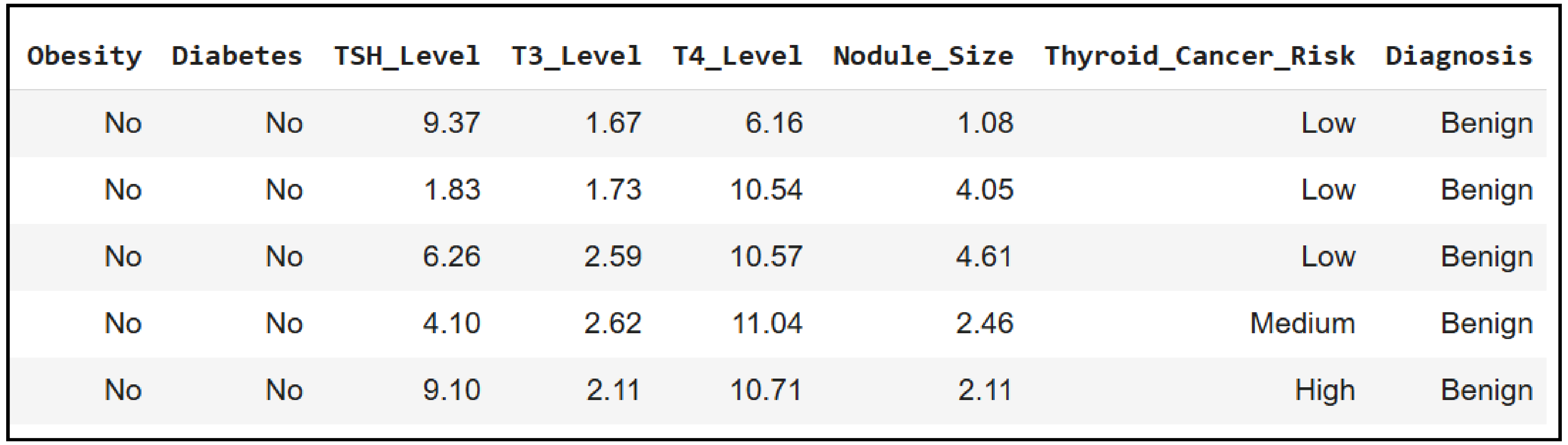
```
thyroid_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 212691 entries, 0 to 212690
Data columns (total 17 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Patient_ID           212691 non-null  int64
 1   Age                  212691 non-null  int64
 2   Gender               212691 non-null  object
 3   Country              212691 non-null  object
 4   Ethnicity            212691 non-null  object
 5   Family_History       212691 non-null  object
 6   Radiation_Exposure   212691 non-null  object
 7   Iodine_Deficiency    212691 non-null  object
 8   Smoking              212691 non-null  object
 9   Obesity              212691 non-null  object
 10  Diabetes             212691 non-null  object
 11  TSH_Level            212691 non-null  float64
 12  T3_Level             212691 non-null  float64
 13  T4_Level             212691 non-null  float64
 14  Nodule_Size          212691 non-null  float64
 15  Thyroid_Cancer_Risk  212691 non-null  object
 16  Diagnosis            212691 non-null  object
dtypes: float64(4), int64(2), object(11)
memory usage: 27.6+ MB
```



```
thyroid_df.head()
```

	Patient_ID	Age	Gender	Country	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking
0	1	66	Male	Russia	Caucasian	No	Yes	No	No
1	2	29	Male	Germany	Hispanic	No	Yes	No	No
2	3	86	Male	Nigeria	Caucasian	No	No	No	No
3	4	75	Female	India	Asian	No	No	No	No
4	5	35	Female	Germany	African	Yes	Yes	No	No

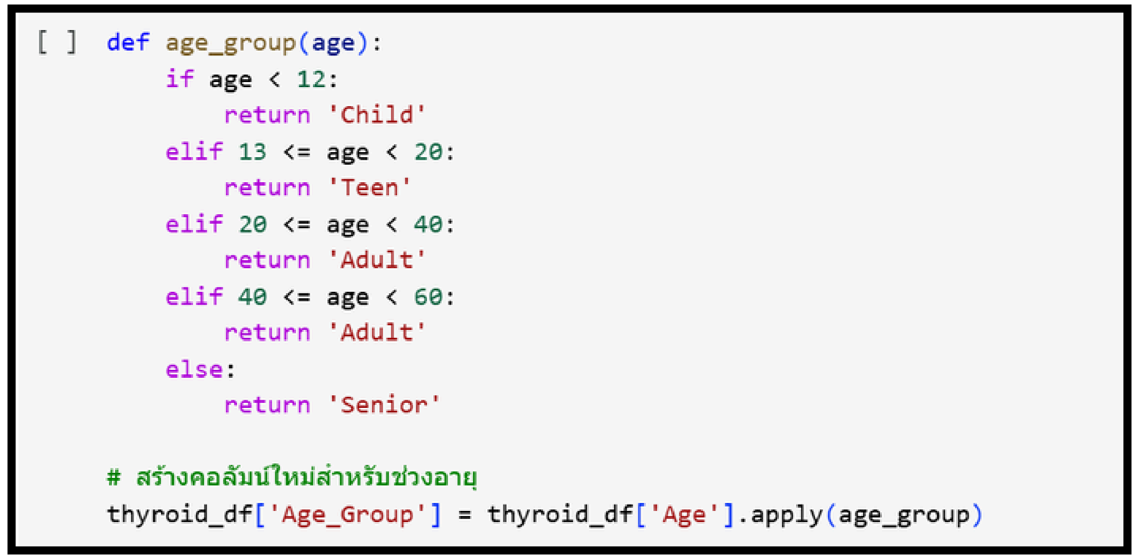


Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
No	No	9.37	1.67	6.16	1.08	Low	Benign
No	No	1.83	1.73	10.54	4.05	Low	Benign
No	No	6.26	2.59	10.57	4.61	Low	Benign
No	No	4.10	2.62	11.04	2.46	Medium	Benign
No	No	9.10	2.11	10.71	2.11	High	Benign

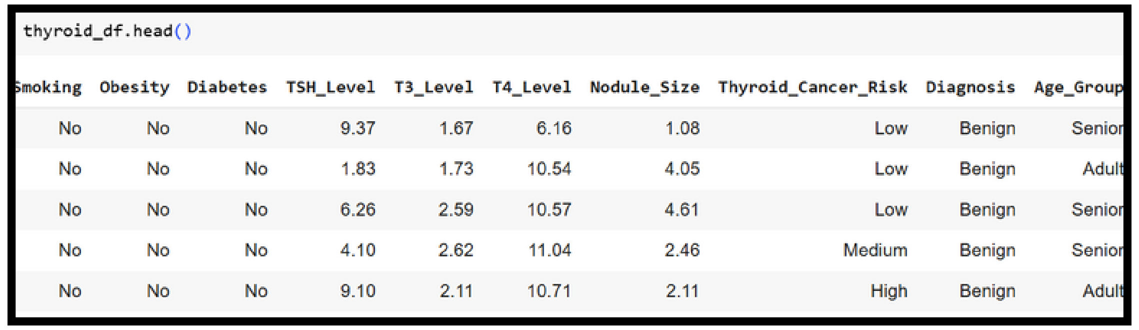
DATA PREPARATION

แบ่งช่วงอายุ

- Child = < 12 yrs
- Teen = 13-19 yrs
- Adult = 20-39 yrs
- Middle Age Adult = 40-59 yrs
- Senior Adult = 60+



```
[ ]  def age_group(age):
         if age < 12:
             return 'Child'
         elif 13 <= age < 20:
             return 'Teen'
         elif 20 <= age < 40:
             return 'Adult'
         elif 40 <= age < 60:
             return 'Adult'
         else:
             return 'Senior'

     # สร้างคอลัมน์ใหม่สำหรับช่วงอายุ
     thyroid_df['Age_Group'] = thyroid_df['Age'].apply(age_group)
```

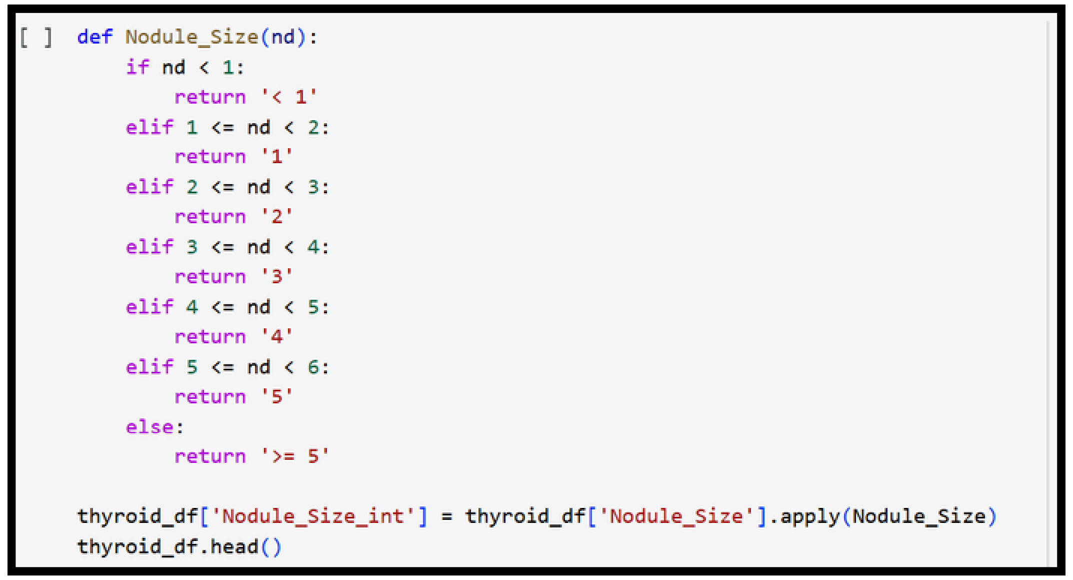


```
thyroid_df.head()
```

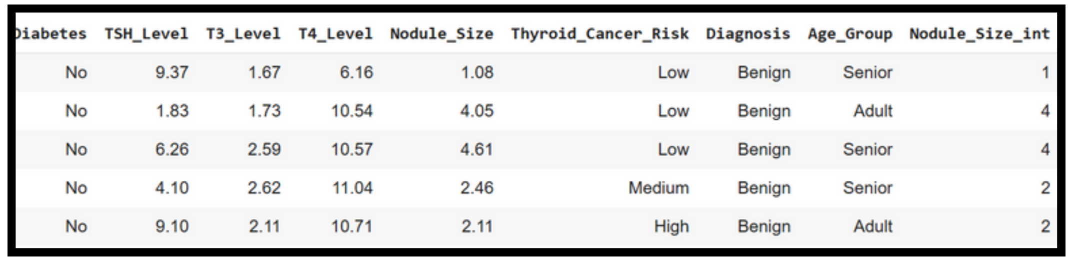
Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis	Age_Group
No	No	No	9.37	1.67	6.16	1.08	Low	Benign	Senior
No	No	No	1.83	1.73	10.54	4.05	Low	Benign	Adult
No	No	No	6.26	2.59	10.57	4.61	Low	Benign	Senior
No	No	No	4.10	2.62	11.04	2.46	Medium	Benign	Senior
No	No	No	9.10	2.11	10.71	2.11	High	Benign	Adult

การแบ่งขนาดก้อนเนื้อ

- ถ้า nd น้อยกว่า 1 → คืนค่า '< 1'
- ถ้า nd อยู่ในช่วง 1 ถึง น้อยกว่า 2 → คืนค่า '1'
- ถ้า nd อยู่ในช่วง 2 ถึง น้อยกว่า 3 → คืนค่า '2'
- ถ้า nd อยู่ในช่วง 3 ถึง น้อยกว่า 4 → คืนค่า '3'
- ถ้า nd อยู่ในช่วง 4 ถึง น้อยกว่า 5 → คืนค่า '4'
- ถ้า nd อยู่ในช่วง 5 ถึง น้อยกว่า 6 → คืนค่า '5'
- ถ้า nd มากกว่าหรือเท่ากับ 6 → คืนค่า '>= 5'



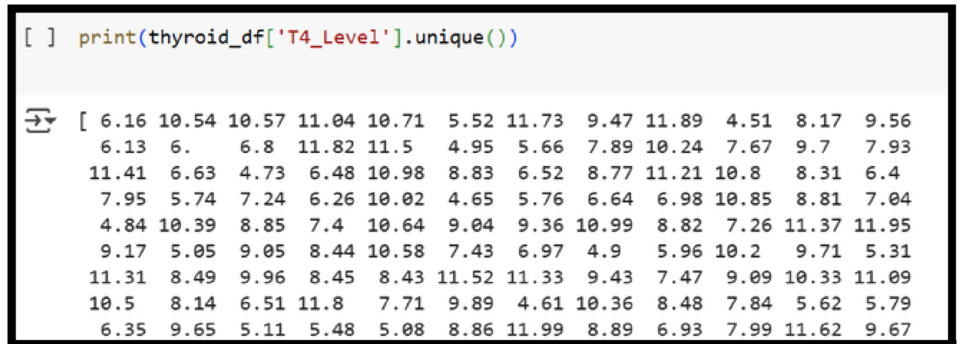
```
[ ]  def Nodule_Size(nd):
         if nd < 1:
             return '< 1'
         elif 1 <= nd < 2:
             return '1'
         elif 2 <= nd < 3:
             return '2'
         elif 3 <= nd < 4:
             return '3'
         elif 4 <= nd < 5:
             return '4'
         elif 5 <= nd < 6:
             return '5'
         else:
             return '>= 5'

     thyroid_df['Nodule_Size_int'] = thyroid_df['Nodule_Size'].apply(Nodule_Size)
     thyroid_df.head()
```



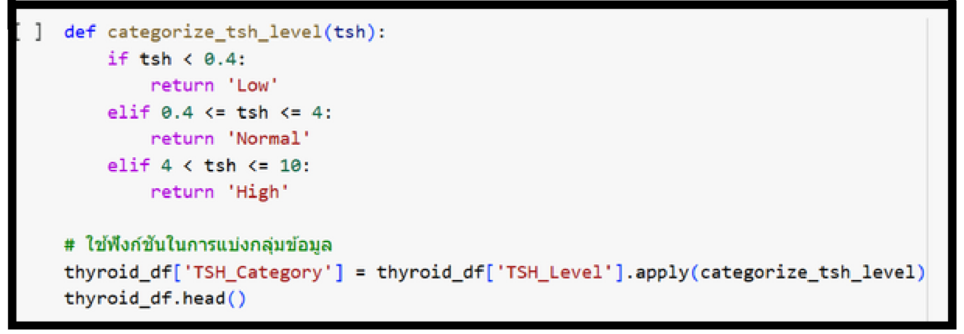
Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis	Age_Group	Nodule_Size_int
No	9.37	1.67	6.16	1.08	Low	Benign	Senior	1
No	1.83	1.73	10.54	4.05	Low	Benign	Adult	4
No	6.26	2.59	10.57	4.61	Low	Benign	Senior	4
No	4.10	2.62	11.04	2.46	Medium	Benign	Senior	2
No	9.10	2.11	10.71	2.11	High	Benign	Adult	2

แบ่งช่วง tsh

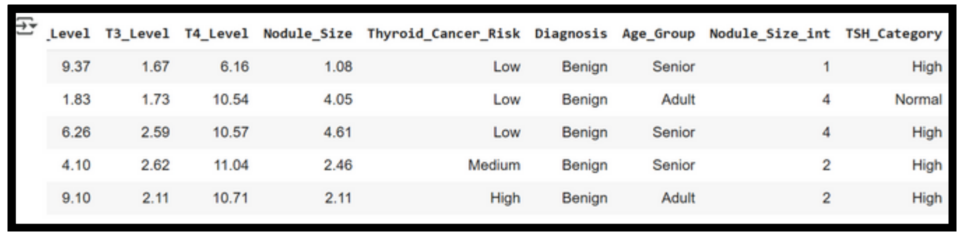
- ถ้า tsh น้อยกว่า 0.4 → คืนค่า 'Low'
- ถ้า tsh อยู่ในช่วง 0.4 ถึง 4 (รวม 4) → คืนค่า 'Normal'
- ถ้า tsh อยู่ในช่วงมากกว่า 4 ถึง 10 (รวม 10) → คืนค่า 'High'



```
[ ]  print(thyroid_df['T4_Level'].unique())

     [ 6.16 10.54 10.57 11.04 10.71  5.52 11.73  9.47 11.89  4.51  8.17  9.56
       6.13  6.    6.8  11.82 11.5   4.95  5.66  7.89 10.24  7.67  9.7   7.93
      11.41  6.63  4.73  6.48 10.98  8.83  6.52  8.77 11.21 10.8   8.31  6.4
       7.95  5.74  7.24  6.26 10.02  4.65  5.76  6.64  6.98 10.85  8.81  7.04
       4.84 10.39  8.85  7.4  10.64  9.04  9.36 10.99  8.82  7.26 11.37 11.95
       9.17  5.05  9.05  8.44 10.58  7.43  6.97  4.9   5.96 10.2   9.71  5.31
      11.31  8.49  9.96  8.45  8.43 11.52 11.33  9.43  7.47  9.09 10.33 11.09
      10.5   8.14  6.51 11.8   7.71  9.89  4.61 10.36  8.48  7.84  5.62  5.79
       6.35  9.65  5.11  5.48  5.08  8.86 11.99  8.89  6.93  7.99 11.62  9.67
```



```
[ ]  def categorize_tsh_level(tsh):
         if tsh < 0.4:
             return 'Low'
         elif 0.4 <= tsh <= 4:
             return 'Normal'
         elif 4 < tsh <= 10:
             return 'High'

     # ใช้ฟังก์ชันในการแบ่งกลุ่มข้อมูล
     thyroid_df['TSH_Category'] = thyroid_df['TSH_Level'].apply(categorize_tsh_level)
     thyroid_df.head()
```

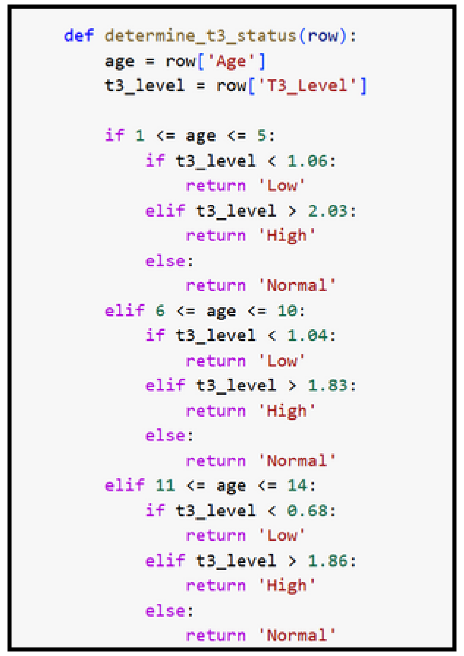


_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis	Age_Group	Nodule_Size_int	TSH_Category
9.37	1.67	6.16	1.08	Low	Benign	Senior	1	High
1.83	1.73	10.54	4.05	Low	Benign	Adult	4	Normal
6.26	2.59	10.57	4.61	Low	Benign	Senior	4	High
4.10	2.62	11.04	2.46	Medium	Benign	Senior	2	High
9.10	2.11	10.71	2.11	High	Benign	Adult	2	High

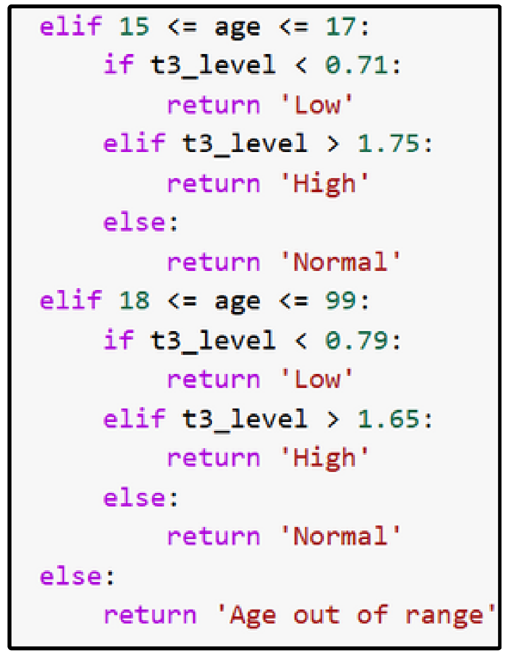
DATA PREPARATION

แบ่งช่วง T3

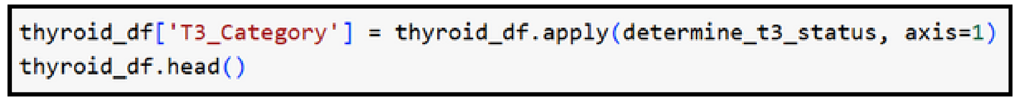
เงื่อนไขการจำแนกตามอายุและค่าปกติของ T3:
อายุ 1–5 ปี: ค่าปกติ 1.06–2.03
อายุ 6–10 ปี: ค่าปกติ 1.04–1.83
อายุ 11–14 ปี: ค่าปกติ 0.68–1.86
อายุ 15–17 ปี: ค่าปกติ 0.71–1.75
อายุ 18–99 ปี: ค่าปกติ 0.79–1.65



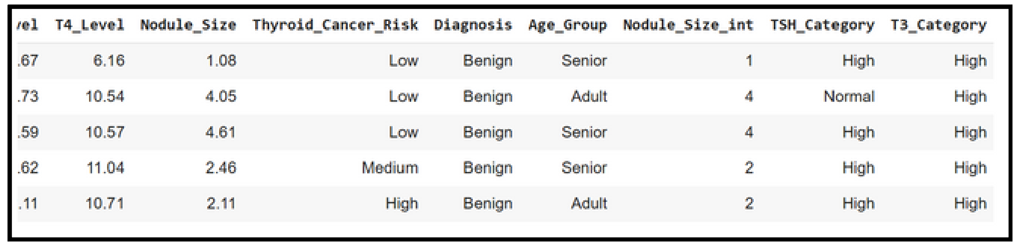
```
def determine_t3_status(row):
    age = row['Age']
    t3_level = row['T3_Level']

    if 1 <= age <= 5:
        if t3_level < 1.06:
            return 'Low'
        elif t3_level > 2.03:
            return 'High'
        else:
            return 'Normal'
    elif 6 <= age <= 10:
        if t3_level < 1.04:
            return 'Low'
        elif t3_level > 1.83:
            return 'High'
        else:
            return 'Normal'
    elif 11 <= age <= 14:
        if t3_level < 0.68:
            return 'Low'
        elif t3_level > 1.86:
            return 'High'
        else:
            return 'Normal'
```



```
    elif 15 <= age <= 17:
        if t3_level < 0.71:
            return 'Low'
        elif t3_level > 1.75:
            return 'High'
        else:
            return 'Normal'
    elif 18 <= age <= 99:
        if t3_level < 0.79:
            return 'Low'
        elif t3_level > 1.65:
            return 'High'
        else:
            return 'Normal'
    else:
        return 'Age out of range'
```



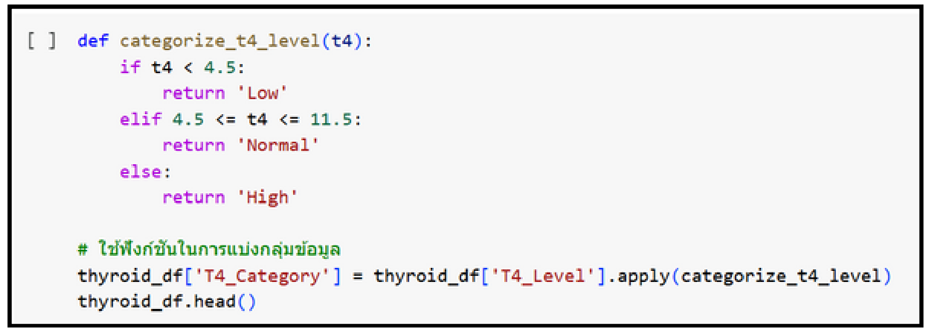
```
thyroid_df['T3_Category'] = thyroid_df.apply(determine_t3_status, axis=1)
thyroid_df.head()
```



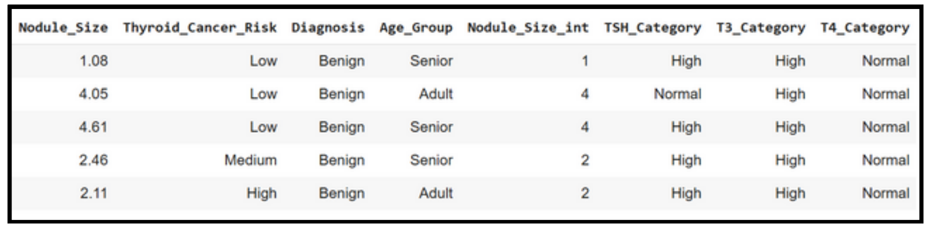
vel	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis	Age_Group	Nodule_Size_int	TSH_Category	T3_Category
.67	6.16	1.08	Low	Benign	Senior	1	High	High
.73	10.54	4.05	Low	Benign	Adult	4	Normal	High
.59	10.57	4.61	Low	Benign	Senior	4	High	High
.62	11.04	2.46	Medium	Benign	Senior	2	High	High
.11	10.71	2.11	High	Benign	Adult	2	High	High

แบ่งช่วง T4

- ถ้า t4 น้อยกว่า 4.5 → คืนค่า 'Low'
- ถ้า t4 อยู่ในช่วง 4.5 ถึง 11.5 (รวม 11.5) → คืนค่า 'Normal'
- ถ้า t4 มากกว่า 11.5 → คืนค่า 'High'

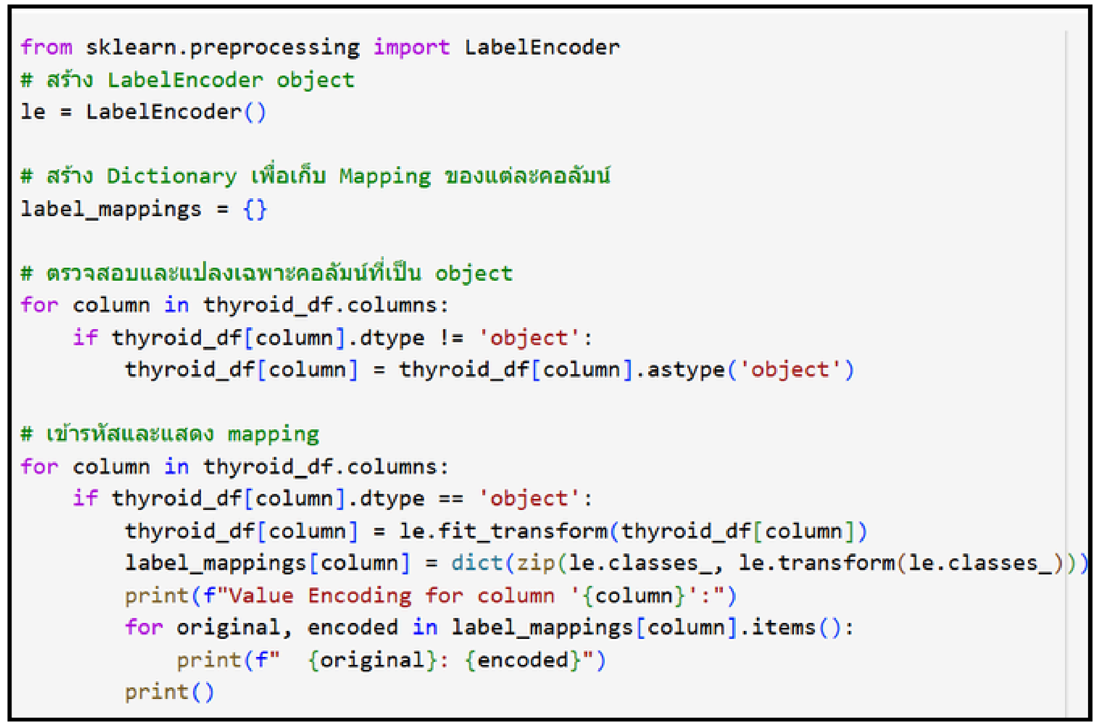


```
[ ]  def categorize_t4_level(t4):
         if t4 < 4.5:
             return 'Low'
         elif 4.5 <= t4 <= 11.5:
             return 'Normal'
         else:
             return 'High'

     # ใช้ฟังก์ชันในการแบ่งกลุ่มข้อมูล
     thyroid_df['T4_Category'] = thyroid_df['T4_Level'].apply(categorize_t4_level)
     thyroid_df.head()
```

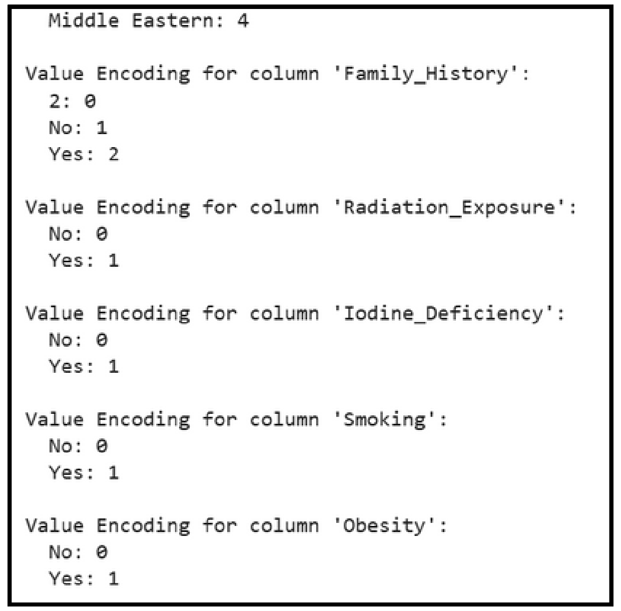


Nodule_Size	Thyroid_Cancer_Risk	Diagnosis	Age_Group	Nodule_Size_int	TSH_Category	T3_Category	T4_Category
1.08	Low	Benign	Senior	1	High	High	Normal
4.05	Low	Benign	Adult	4	Normal	High	Normal
4.61	Low	Benign	Senior	4	High	High	Normal
2.46	Medium	Benign	Senior	2	High	High	Normal
2.11	High	Benign	Adult	2	High	High	Normal

One - hot Encoding



```
from sklearn.preprocessing import LabelEncoder
# สร้าง LabelEncoder object
le = LabelEncoder()

# สร้าง Dictionary เพื่อเก็บ Mapping ของแต่ละคอลัมน์
label_mappings = {}

# ตรวจสอบและแปลงเฉพาะคอลัมน์ที่เป็น object
for column in thyroid_df.columns:
    if thyroid_df[column].dtype != 'object':
        thyroid_df[column] = thyroid_df[column].astype('object')

# เข้ารหัสและแสดง mapping
for column in thyroid_df.columns:
    if thyroid_df[column].dtype == 'object':
        thyroid_df[column] = le.fit_transform(thyroid_df[column])
        label_mappings[column] = dict(zip(le.classes_, le.transform(le.classes_)))
        print(f"Value Encoding for column '{column}':")
        for original, encoded in label_mappings[column].items():
            print(f"  {original}: {encoded}")
        print()
```



```
  Middle Eastern: 4

Value Encoding for column 'Family_History':
  2: 0
  No: 1
  Yes: 2

Value Encoding for column 'Radiation_Exposure':
  No: 0
  Yes: 1

Value Encoding for column 'Iodine_Deficiency':
  No: 0
  Yes: 1

Value Encoding for column 'Smoking':
  No: 0
  Yes: 1

Value Encoding for column 'Obesity':
  No: 0
  Yes: 1
```

Reference

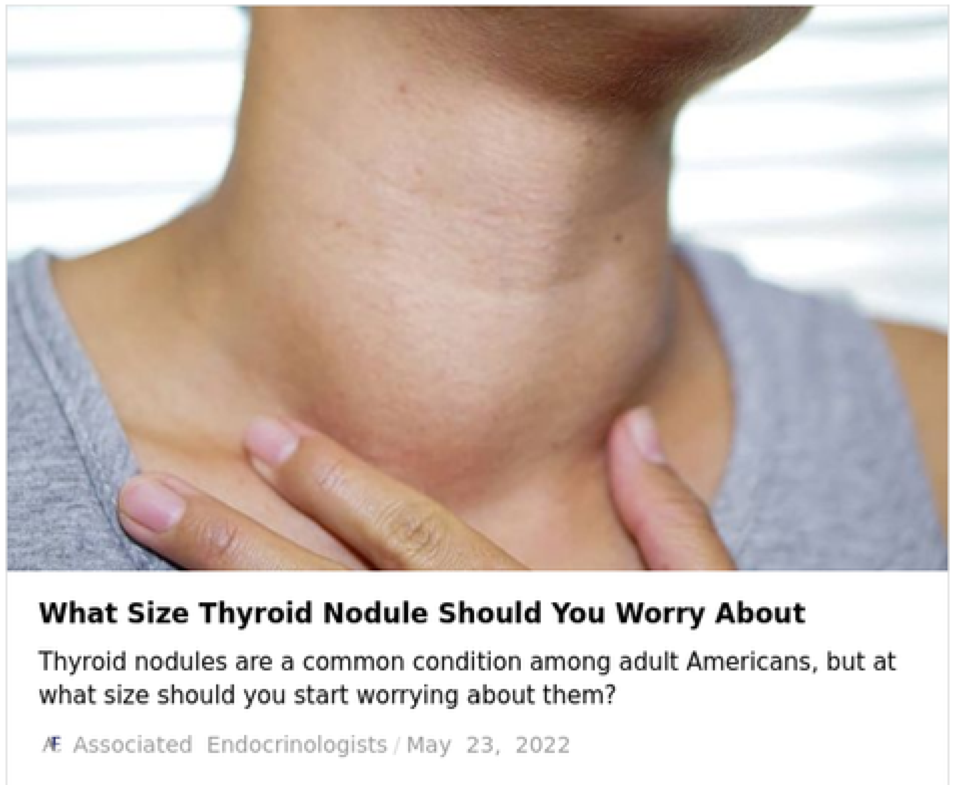
แบ่งช่วงอายุ

Stages of Life: Health for Every Age

Check back to the INTEGRIS Health On Your Health blog for the latest health and wellness news for all Oklahomans.

integrishealth /

การแบ่งขนาดก้อนเนื้อ

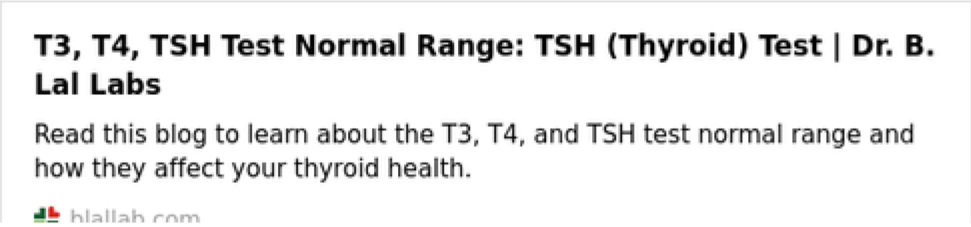


What Size Thyroid Nodule Should You Worry About

Thyroid nodules are a common condition among adult Americans, but at what size should you start worrying about them?

AE Associated Endocrinologists / May 23, 2022

แบ่งช่วง tsh t3 t4



T3, T4, TSH Test Normal Range: TSH (Thyroid) Test | Dr. B. Lal Labs

Read this blog to learn about the T3, T4, and TSH test normal range and how they affect your thyroid health.

blallab.com

ASSOCIATION RULE

GROUPBY PATIENT_ID

Rule		Support	Confidence	Lift
T4 = Normal	Diagnosis = Malignant	0.93	1.0	1.0
Diabetes = No	Diagnosis = Malignant	0.80	1.0	1.0
T3 = High	Diagnosis = Malignant	0.61	1.0	1.0
TSH = High	Diagnosis = Malignant	0.61	1.0	1.0

Rule		Support	Confidence	Lift

Diabetes = No ,
T4 = Normal
	Diagnosis = Malignant	0.75	1.0	1.0

Gender = Female	Diagnosis = Malignant	0.60	1.0	1.0

T4 = Normal , Smoking = No ,
Diabetes = No
	Diagnosis = Malignant	0.57	1.0	1.0

ASSOCIATION RULE

GROUPBY ETHNICITY

Ethnicity	Rule		Support	Confidence	Lift
Asian / Caucasion / African / Hispanic	T4 = Normal	Diagnosis = Malignant	0.93	1.0	1.0
Asian	Thyroid_Cancer_Risk = High	Diagnosis = Malignant	0.70	1.0	1.0
Caucasian	Family_History = No	Diagnosis = Malignant	0.70	1.0	1.0
African	T3 = High	Diagnosis = Malignant	0.62	1.0	1.0

ASSOCIATION RULE

GROUPBY ETHNICITY

Ethnicity	Rule		Support	Confidence	Lift
African	TSH = High	Diagnosis = Malignant	0.60	1.0	1.0
Middle Eastern / Hispanic	T4 = Normal	Diagnosis = Malignant	0.93	1.0	1.0
Middle Eastern	T3 = High	Diagnosis = Malignant	0.60	1.0	1.0
Hispanic	TSH = High	Diagnosis = Malignant	0.61	1.0	1.0

MODEL

DECISION TREE

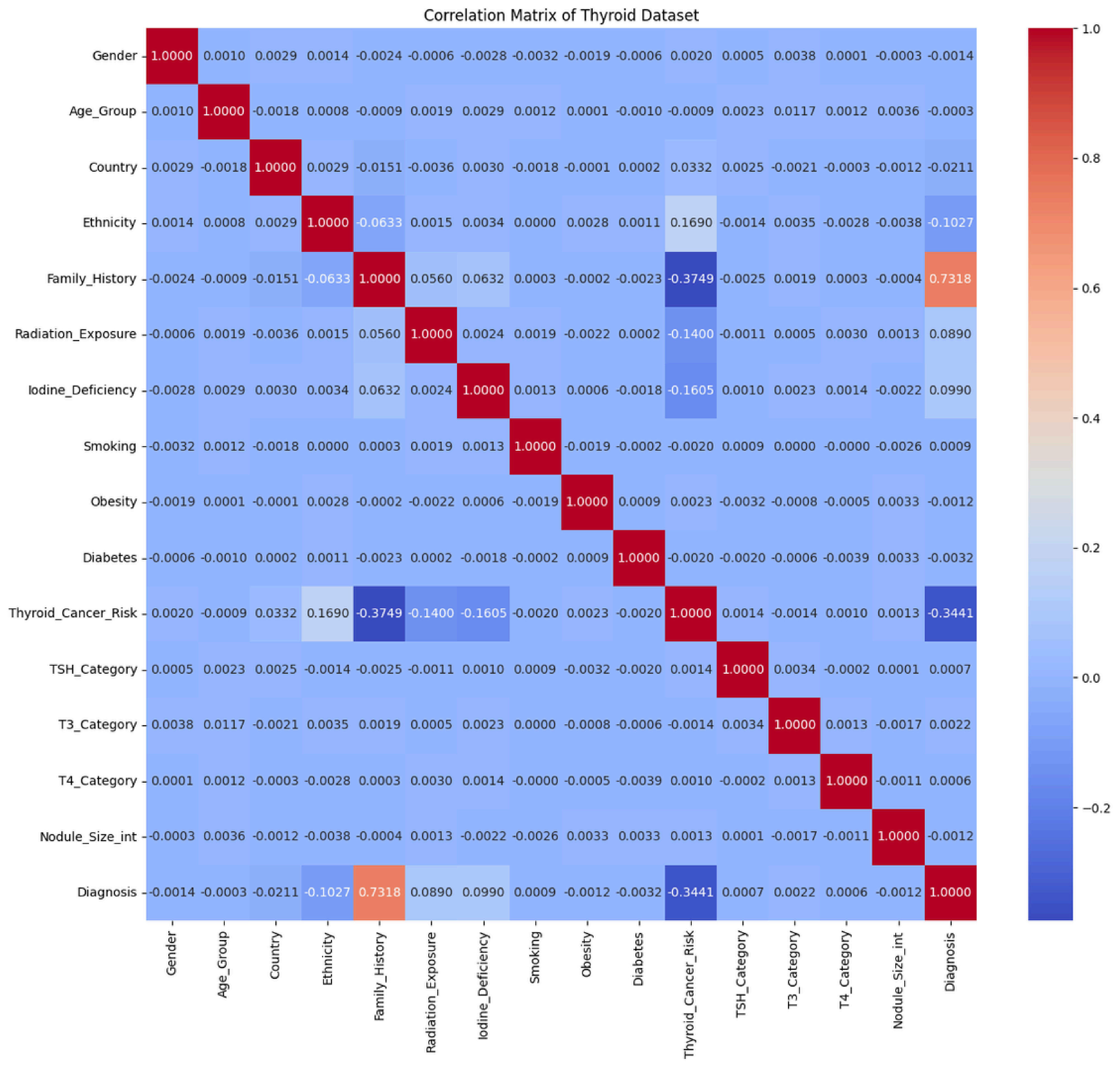
NAIVE BAYES

LOGISTIC REGRESSION

Value Encoding for column

ตัวแปร	คำอธิบายข้อมูล	ตัวแปร	คำอธิบายข้อมูล
Diagnosis	Benign: 0 ,Malignant: 1	Radition_Exposure	No: 0 , Yes: 1
Age	Adult: 0, Senior: 1, Teen: 2	Iodine_Deficiency	No: 0 , Yes: 1
Gender	Female: 0, Male: 1	Smoking	No: 0 , Yes: 1
Country	Brazil: 0 ,China: 1 ,Germany: 2 ,India: 3 ,Japan: 4 ,Nigeria: 5 ,Russia: 6 ,South Korea: 7 ,UK: 8 ,USA: 9	Obesity	No: 0 , Yes: 1
Ethnicity	African: 0 ,Asian: 1 ,Caucasian: 2 ,Hispanic: 3 ,Middle Eastern: 4	Diabetes	No: 0 , Yes: 1
Family_History	No: 0 , Yes: 1	TSH_Category	High: 0 ,Low: 1 ,Normal: 2
T3_Category	High: 0 ,Low: 1 ,Normal: 2	T4_Category	High: 0 ,Low: 1 ,Normal: 2
Nodule_Size	1: 0 ,2: 1 ,3: 2 ,4: 3 ,5: 4 ,< 1: 5	Thyroid_Cancer_Risk	High: 0 , Low: 1 ,Medium: 2

Correlation Matrix of Thyroid Dataset



	Gender	Age_Group	Country	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	Thyroid_Cancer_Risk	TSH_Category	T3_Category	T4_Category	Nodule_Size_int	Diagnosis
Gender	1.0000	0.0010	0.0029	0.0014	-0.0024	-0.0006	-0.0028	-0.0032	-0.0019	-0.0006	0.0020	0.0005	0.0038	0.0001	-0.0003	-0.0014
Age_Group	0.0010	1.0000	-0.0018	0.0008	-0.0009	0.0019	0.0029	0.0012	0.0001	-0.0010	-0.0009	0.0023	0.0117	0.0012	0.0036	-0.0003
Country	0.0029	-0.0018	1.0000	0.0029	-0.0151	-0.0036	0.0030	-0.0018	-0.0001	0.0002	0.0332	0.0025	-0.0021	-0.0003	-0.0012	-0.0211
Ethnicity	0.0014	0.0008	0.0029	1.0000	-0.0633	0.0015	0.0034	0.0000	0.0028	0.0011	0.1690	-0.0014	0.0035	-0.0028	-0.0038	-0.1027
Family_History	-0.0024	-0.0009	-0.0151	-0.0633	1.0000	0.0560	0.0632	0.0003	-0.0002	-0.0023	-0.3749	-0.0025	0.0019	0.0003	-0.0004	0.7318
Radiation_Exposure	-0.0006	0.0019	-0.0036	0.0015	0.0560	1.0000	0.0024	0.0019	-0.0022	0.0002	-0.1400	-0.0011	0.0005	0.0030	0.0013	0.0890
Iodine_Deficiency	-0.0028	0.0029	0.0030	0.0034	0.0632	0.0024	1.0000	0.0013	0.0006	-0.0018	-0.1605	0.0010	0.0023	0.0014	-0.0022	0.0990
Smoking	-0.0032	0.0012	-0.0018	0.0000	0.0003	0.0019	0.0013	1.0000	-0.0019	-0.0002	-0.0020	0.0009	0.0000	-0.0000	-0.0026	0.0009
Obesity	-0.0019	0.0001	-0.0001	0.0028	-0.0002	-0.0022	0.0006	-0.0019	1.0000	0.0009	0.0023	-0.0032	-0.0008	-0.0005	0.0033	-0.0012
Diabetes	-0.0006	-0.0010	0.0002	0.0011	-0.0023	0.0002	-0.0018	-0.0002	0.0009	1.0000	-0.0020	-0.0020	-0.0006	-0.0039	0.0033	-0.0032
Thyroid_Cancer_Risk	0.0020	-0.0009	0.0332	0.1690	-0.3749	-0.1400	-0.1605	-0.0020	0.0023	-0.0020	1.0000	0.0014	-0.0014	0.0010	0.0013	-0.3441
TSH_Category	0.0005	0.0023	0.0025	-0.0014	-0.0025	-0.0011	0.0010	0.0009	-0.0032	-0.0020	0.0014	1.0000	0.0034	-0.0002	0.0001	0.0007
T3_Category	0.0038	0.0117	-0.0021	0.0035	0.0019	0.0005	0.0023	0.0000	-0.0008	-0.0006	-0.0014	0.0034	1.0000	0.0013	-0.0017	0.0022
T4_Category	0.0001	0.0012	-0.0003	-0.0028	0.0003	0.0030	0.0014	-0.0000	-0.0005	-0.0039	0.0010	-0.0002	0.0013	1.0000	-0.0011	0.0006
Nodule_Size_int	-0.0003	0.0036	-0.0012	-0.0038	-0.0004	0.0013	-0.0022	-0.0026	0.0033	0.0033	0.0013	0.0001	-0.0017	-0.0011	1.0000	-0.0012
Diagnosis	-0.0014	-0.0003	-0.0211	-0.1027	0.7318	0.0890	0.0990	0.0009	-0.0012	-0.0032	-0.3441	0.0007	0.0022	0.0006	-0.0012	1.0000

Features ที่นำเข้าและ Target

X	Ethnicity, Family_History, Thyroid_Cancer_Risk, Radiation_Exposure, Iodine_Deficiency
Y	Diagnosis

แบ่งข้อมูล Train/Test

Train	80%
Test	20%

Value counts ของ y_train (Diagnosis)

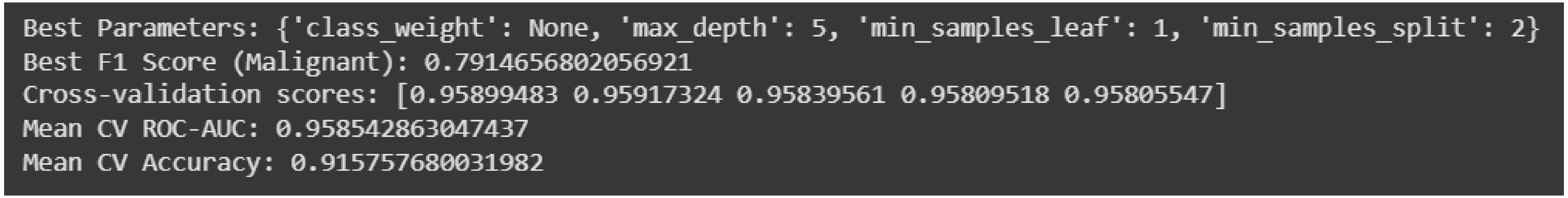
Benign	130631	170,152	212,691
Malignant	39521		

Value counts ของ y_test (Diagnosis)

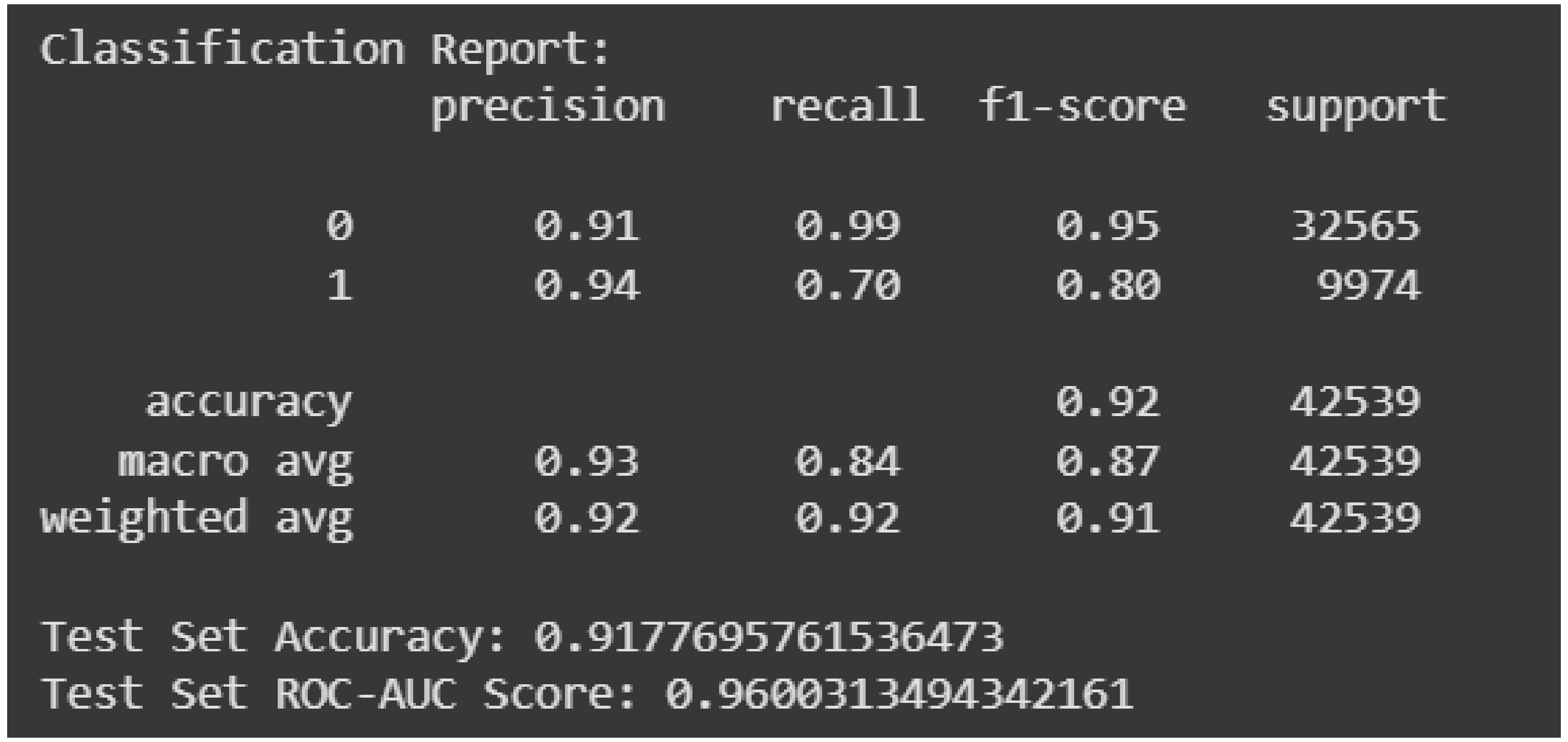
Benign	32565	42,539
Malignant	9974	

DECISION TREE

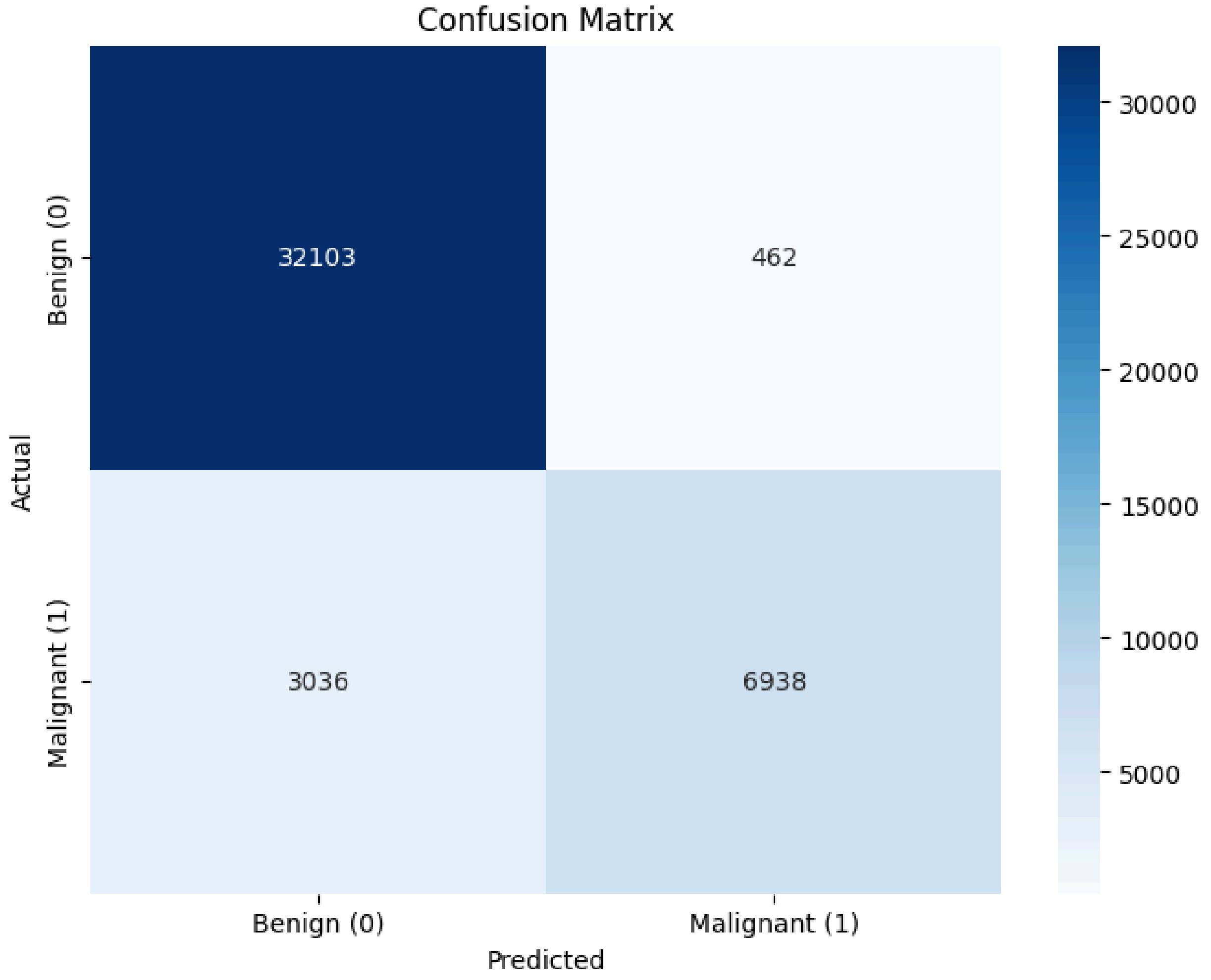
Result

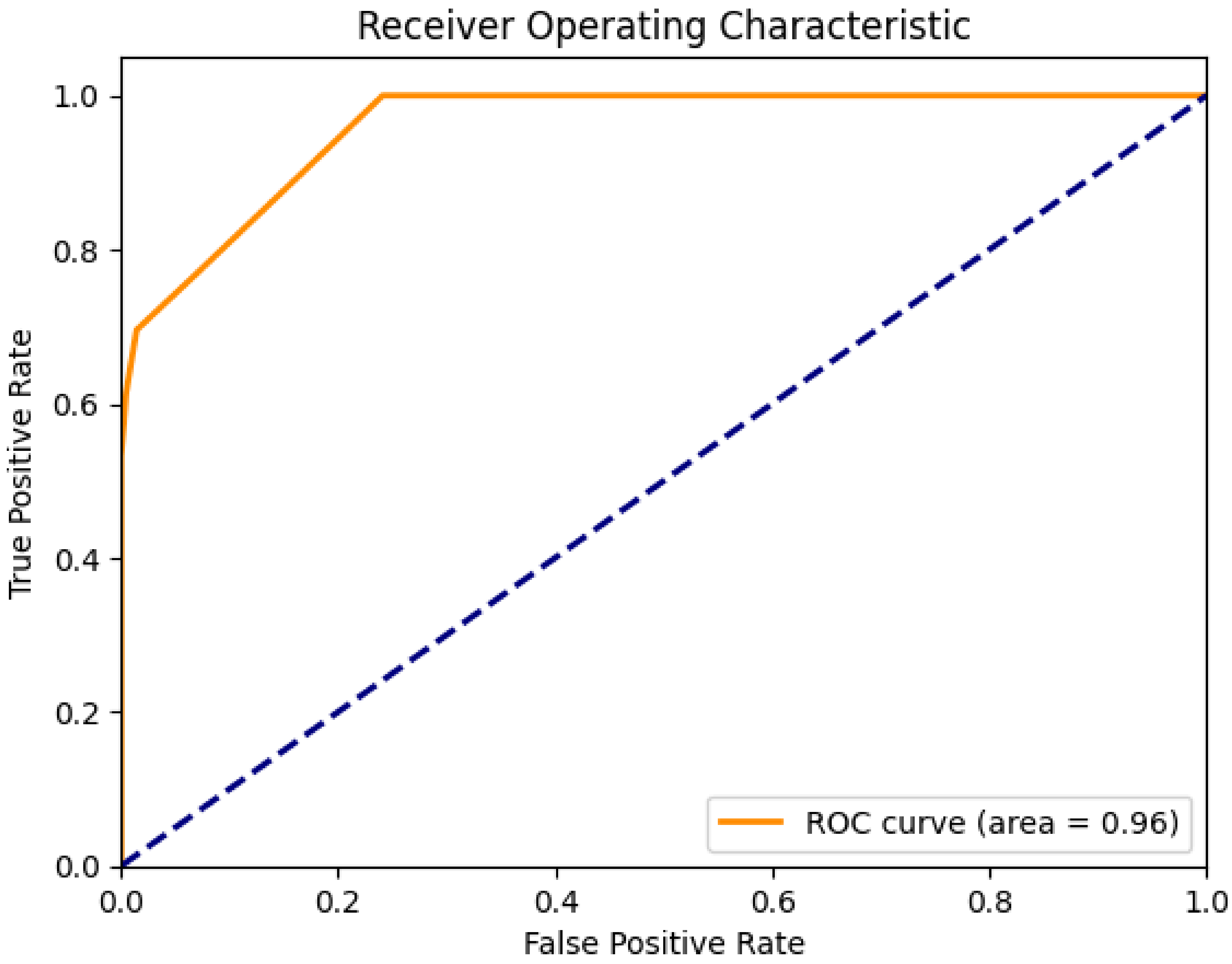


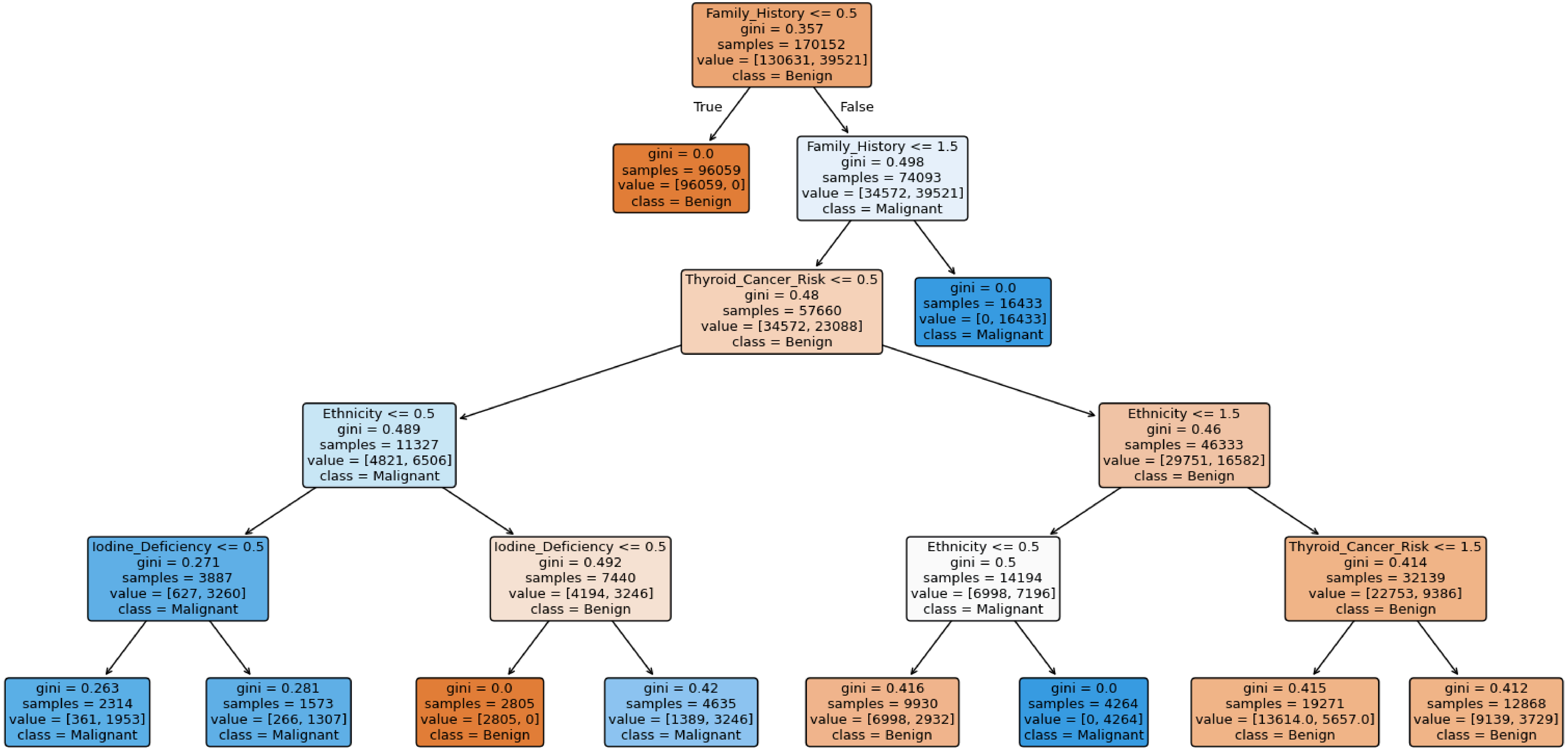
```
Best Parameters: {'class_weight': None, 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}
Best F1 Score (Malignant): 0.7914656802056921
Cross-validation scores: [0.95899483 0.95917324 0.95839561 0.95809518 0.95805547]
Mean CV ROC-AUC: 0.958542863047437
Mean CV Accuracy: 0.915757680031982
```



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

           0       0.91      0.99      0.95     32565
           1       0.94      0.70      0.80      9974

    accuracy                           0.92     42539
   macro avg       0.93      0.84      0.87     42539
weighted avg       0.92      0.92      0.91     42539

Test Set Accuracy: 0.9177695761536473
Test Set ROC-AUC Score: 0.9600313494342161
```

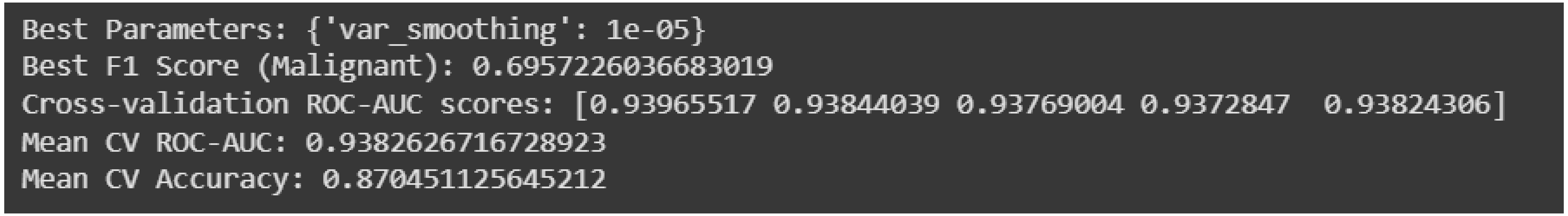



Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.96)

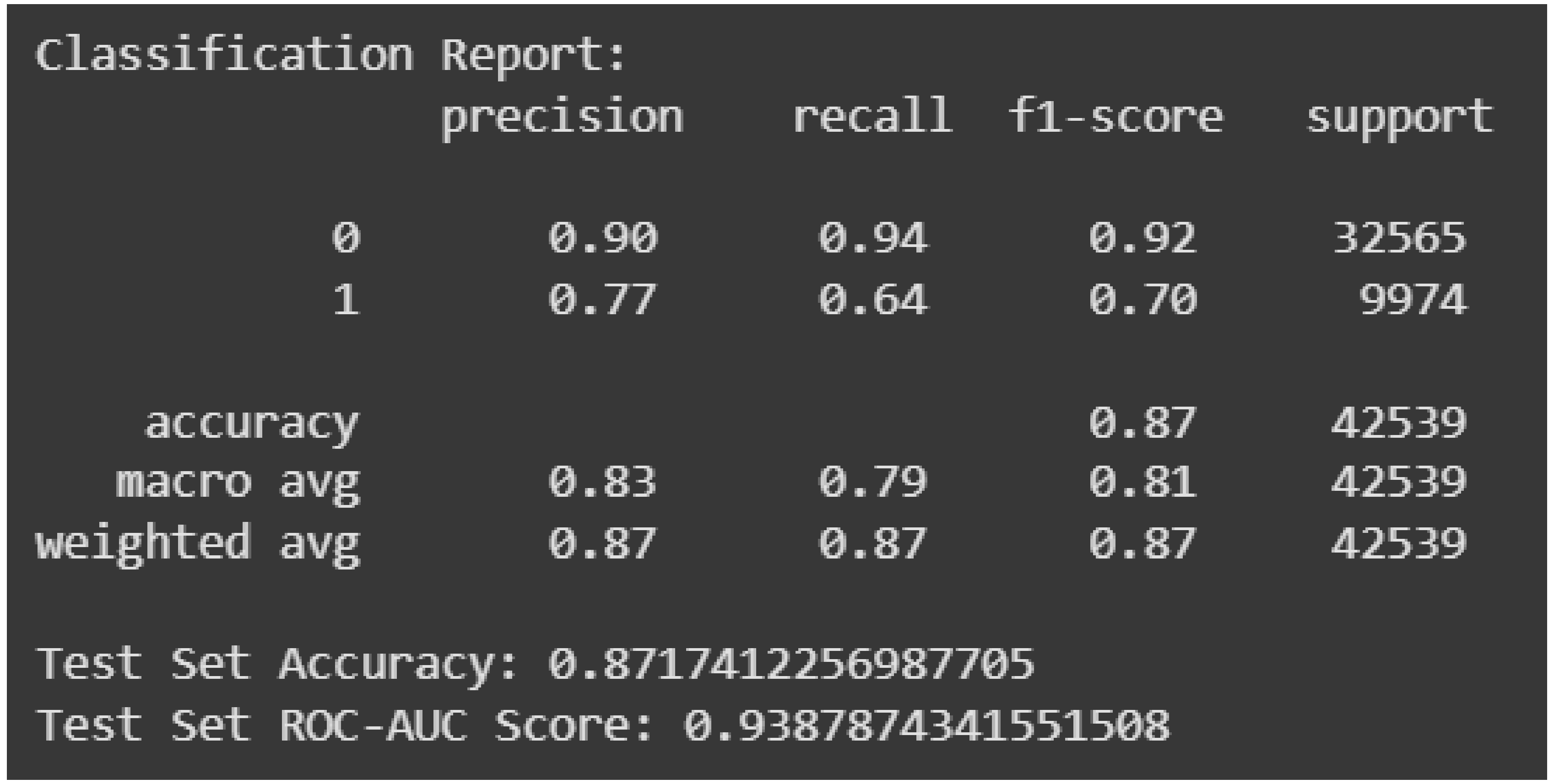

Family_History <= 0.5
gini = 0.357
samples = 170152
value = [130631, 39521]
class = Benign
True
False
gini = 0.0
samples = 96059
value = [96059, 0]
class = Benign
Family_History <= 1.5
gini = 0.498
samples = 74093
value = [34572, 39521]
class = Malignant
Thyroid_Cancer_Risk <= 0.5
gini = 0.48
samples = 57660
value = [34572, 23088]
class = Benign
gini = 0.0
samples = 16433
value = [0, 16433]
class = Malignant
Ethnicity <= 0.5
gini = 0.489
samples = 11327
value = [4821, 6506]
class = Malignant
Ethnicity <= 1.5
gini = 0.46
samples = 46333
value = [29751, 16582]
class = Benign
Iodine_Deficiency <= 0.5
gini = 0.271
samples = 3887
value = [627, 3260]
class = Malignant
Iodine_Deficiency <= 0.5
gini = 0.492
samples = 7440
value = [4194, 3246]
class = Benign
Ethnicity <= 0.5
gini = 0.5
samples = 14194
value = [6998, 7196]
class = Malignant
Thyroid_Cancer_Risk <= 1.5
gini = 0.414
samples = 32139
value = [22753, 9386]
class = Benign
gini = 0.263
samples = 2314
value = [361, 1953]
class = Malignant
gini = 0.281
samples = 1573
value = [266, 1307]
class = Malignant
gini = 0.0
samples = 2805
value = [2805, 0]
class = Benign
gini = 0.42
samples = 4635
value = [1389, 3246]
class = Malignant
gini = 0.416
samples = 9930
value = [6998, 2932]
class = Benign
gini = 0.0
samples = 4264
value = [0, 4264]
class = Malignant
gini = 0.415
samples = 19271
value = [13614.0, 5657.0]
class = Benign
gini = 0.412
samples = 12868
value = [9139, 3729]
class = Benign

NAIVE BAYES

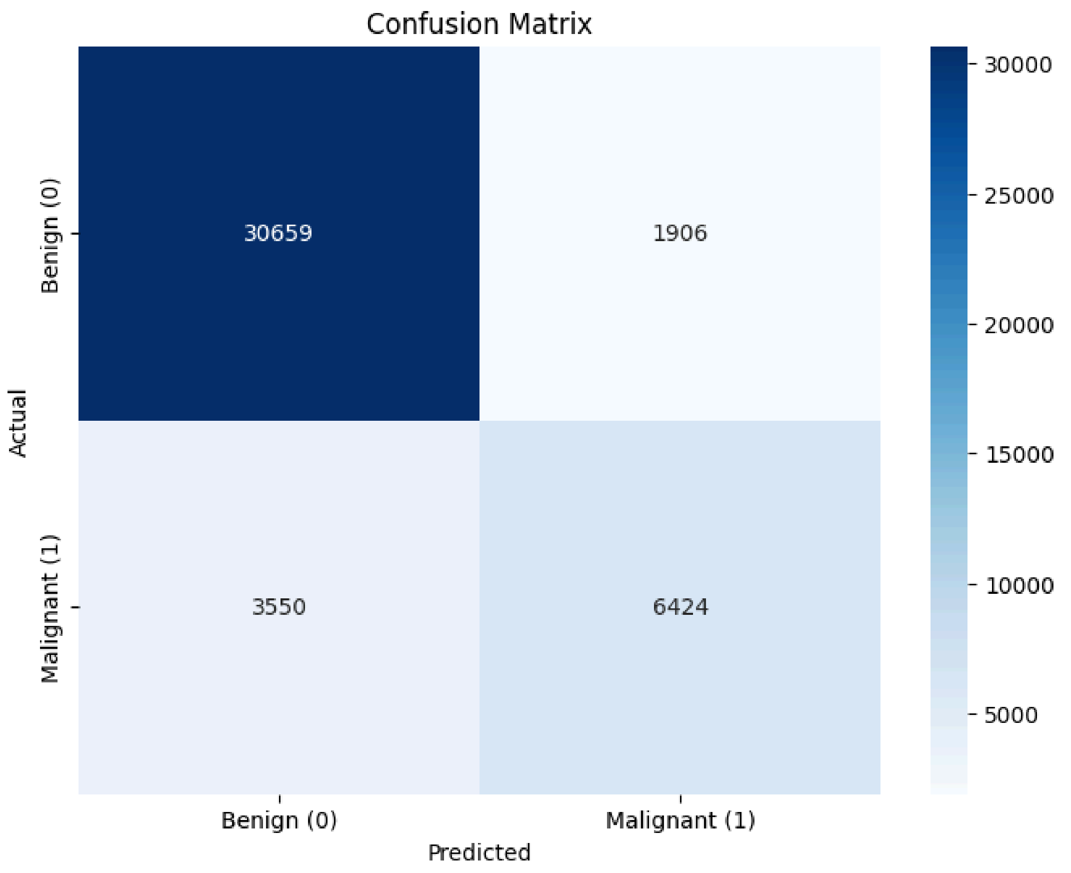
Result



```
Best Parameters: {'var_smoothing': 1e-05}
Best F1 Score (Malignant): 0.6957226036683019
Cross-validation ROC-AUC scores: [0.93965517 0.93844039 0.93769004 0.9372847  0.93824306]
Mean CV ROC-AUC: 0.9382626716728923
Mean CV Accuracy: 0.870451125645212
```

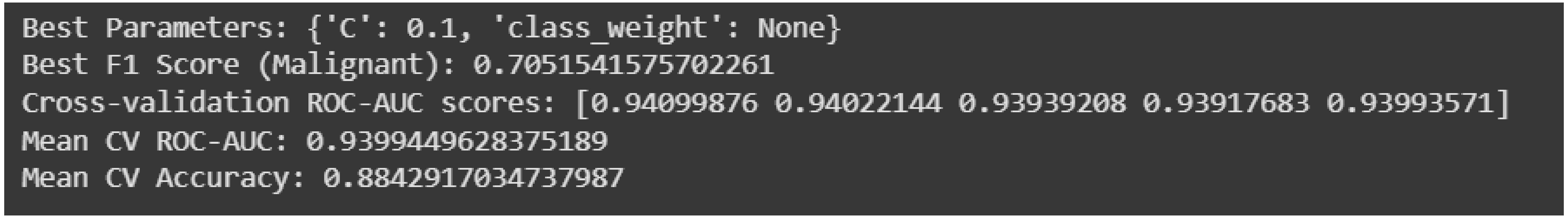


```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.94      0.92     32565
           1       0.77      0.64      0.70      9974

    accuracy                           0.87     42539
   macro avg       0.83      0.79      0.81     42539
weighted avg       0.87      0.87      0.87     42539

Test Set Accuracy: 0.8717412256987705
Test Set ROC-AUC Score: 0.9387874341551508
```

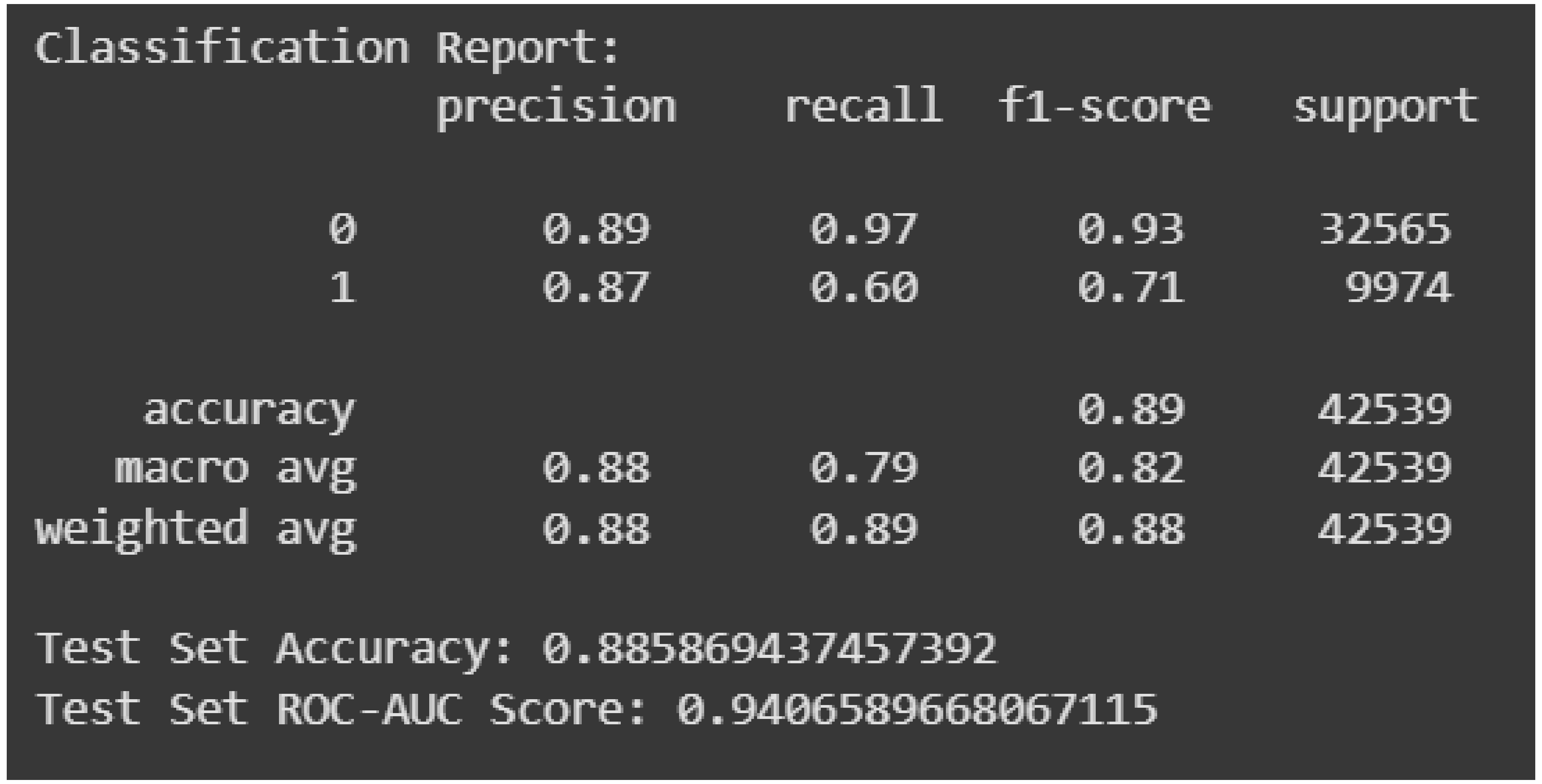


LOGISTIC REGRESSION

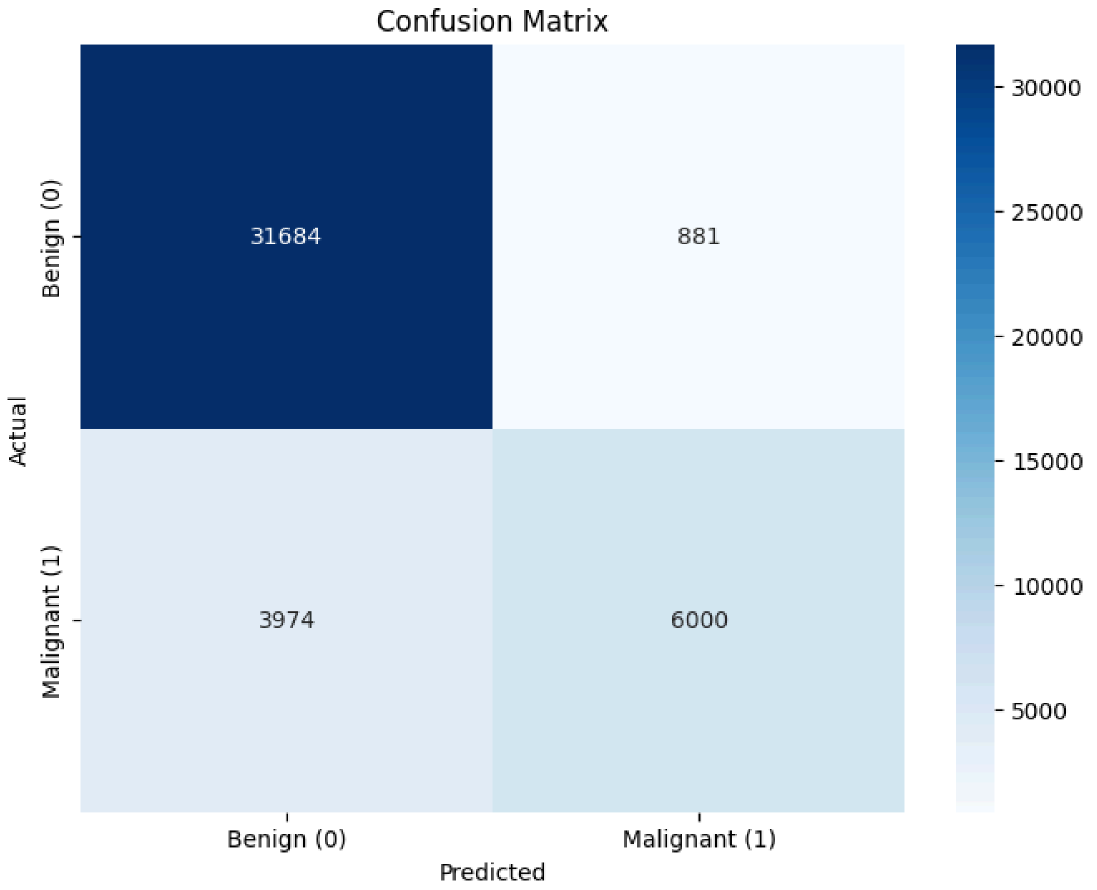
Result



```
Best Parameters: {'C': 0.1, 'class_weight': None}
Best F1 Score (Malignant): 0.7051541575702261
Cross-validation ROC-AUC scores: [0.94099876 0.94022144 0.93939208 0.93917683 0.93993571]
Mean CV ROC-AUC: 0.9399449628375189
Mean CV Accuracy: 0.8842917034737987
```



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

           0       0.89      0.97      0.93     32565
           1       0.87      0.60      0.71      9974

    accuracy                           0.89     42539
   macro avg       0.88      0.79      0.82     42539
weighted avg       0.88      0.89      0.88     42539

Test Set Accuracy: 0.885869437457392
Test Set ROC-AUC Score: 0.9406589668067115
```



MODEL SELECTION

Model	Precision	Recall	F1-Score	Accuracy	ROC-AUC
Decision Tree	**0.93**	**0.84**	**0.87**	**0.9157**	**0.9157**
Naive Bayes (Gaussian)	0.83	0.79	0.81	0.8705	0.9383
Logistic Regression	0.88	0.79	0.82	0.8843	0.9399

MODEL SELECTION

Model	Accuracy	ROC-AUC
Decision Tree	**0.9178**	**0.9600**
Naive Bayes (Gaussian)	0.8717	0.9388
Logistic Regression	0.8859	0.9407

THANK YOU

FOR YOUR ATTENTION

นางสาวพรรณรมณ ราชคมน์	653020213-2
นางสาวสิริญาพร รสจันทร์	653020218-2
นางสาวพรวลัย ฟ็อกซ์ออล	653020573-2