```
from google.colab import drive
import os,sys
drive.mount('/content/drive')
os.chdir('/content/drive/MyDrive/thyroid_cancer')
sys.path.append('/content/gdrive/MyDrive/thyroid_cancer')
```

```
Mounted at /content/drive
```

```
!pip install feature_engine
```

```
Collecting feature_engine
  Downloading feature_engine-1.8.3-py2.py3-none-any.whl.metadata (9.9 kB)
Requirement already satisfied: numpy>=1.18.2 in /usr/local/lib/python3.11/dist-packages (from feature_engine) (2.0.2)
Requirement already satisfied: pandas>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from feature_engine) (2.2.2)
Requirement already satisfied: scikit-learn>=1.4.0 in /usr/local/lib/python3.11/dist-packages (from feature_engine) (1.6.1)
Requirement already satisfied: scipy>=1.4.1 in /usr/local/lib/python3.11/dist-packages (from feature_engine) (1.14.1)
Requirement already satisfied: statsmodels>=0.11.1 in /usr/local/lib/python3.11/dist-packages (from feature_engine) (0.14.4)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=2.2.0->feature_engine) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas>=2.2.0->feature_engine) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=2.2.0->feature_engine) (2025.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.4.0->feature_engine) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.4.0->feature_engine) (3.6.0)
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.11/dist-packages (from statsmodels>=0.11.1->feature_engine) (1.0.1)
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.11/dist-packages (from statsmodels>=0.11.1->feature_engine) (24.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas>=2.2.0->feature_engine) (1.17.0)
Downloading feature_engine-1.8.3-py2.py3-none-any.whl (378 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 378.6/378.6 kB 5.5 MB/s eta 0:00:00
Installing collected packages: feature_engine
Successfully installed feature_engine-1.8.3
```

```
!pip install --upgrade scikit-plot scipy
```

```
Collecting scikit-plot
  Downloading scikit_plot-0.3.7-py3-none-any.whl.metadata (7.1 kB)
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (1.14.1)
Collecting scipy
  Downloading scipy-1.15.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (61 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.0/62.0 kB 4.2 MB/s eta 0:00:00
Requirement already satisfied: matplotlib>=1.4.0 in /usr/local/lib/python3.11/dist-packages (from scikit-plot) (3.10.0)
Requirement already satisfied: scikit-learn>=0.18 in /usr/local/lib/python3.11/dist-packages (from scikit-plot) (1.6.1)
Requirement already satisfied: joblib>=0.10 in /usr/local/lib/python3.11/dist-packages (from scikit-plot) (1.4.2)
Requirement already satisfied: numpy<2.5,>=1.23.5 in /usr/local/lib/python3.11/dist-packages (from scipy) (2.0.2)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (1.4.8)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (24.2)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (3.2.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (2.8.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=0.18->scikit-plot) (3.6.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib>=1.4.0->scikit-plot) (1.17.0)
Downloading scikit_plot-0.3.7-py3-none-any.whl (33 kB)
Downloading scipy-1.15.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (37.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 37.6/37.6 MB 20.5 MB/s eta 0:00:00
Installing collected packages: scipy, scikit-plot
  Attempting uninstall: scipy
    Found existing installation: scipy 1.14.1
    Uninstalling scipy-1.14.1:
      Successfully uninstalled scipy-1.14.1
Successfully installed scikit-plot-0.3.7 scipy-1.15.2
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
import feature_engine
from sklearn.utils import resample
from imblearn.over_sampling import RandomOverSampler
from sklearn.impute import SimpleImputer
from statsmodels.stats.outliers_influence import variance_inflation_factor
from feature_engine.outliers.winsorizer import Winsorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC
import xgboost as xgb
from sklearn.metrics import auc, accuracy_score, confusion_matrix, mean_squared_error
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_curve
from sklearn.metrics import auc
#import scikitplot as skplt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.preprocessing import StandardScaler
import xgboost
from scipy.stats import norm
```

```
df = pd.read_csv("dataset.csv")
df.head(6)
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
0	27	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indeterminate	No
1	34	F	No	Yes	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
2	30	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
3	62	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
4	62	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Excellent	No
5	52	M	Yes	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Indeterminate	No

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df.columns
```

```
Index(['Age', 'Gender', 'Smoking', 'Hx Smoking', 'Hx Radiothreapy',
       'Thyroid Function', 'Physical Examination', 'Adenopathy', 'Pathology',
       'Focality', 'Risk', 'T', 'N', 'M', 'Stage', 'Response', 'Recurred'],
      dtype='object')
```

```
df.shape
```

```
(383, 17)
```

```
df.isnull().sum()
```

	0
Age	0
Gender	0
Smoking	0
Hx Smoking	0
Hx Radiothreapy	0
Thyroid Function	0
Physical Examination	0
Adenopathy	0
Pathology	0
Focality	0
Risk	0
T	0
N	0
M	0
Stage	0
Response	0
Recurred	0

dtype: int64

```
df = df.replace(to_replace='F',value = 0)
df = df.replace(to_replace='M',value = 1)
```

```
<ipython-input-9-d7579330b916>:2: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
  df = df.replace(to_replace='M',value = 1)
```

```
df = df.replace(to_replace='No',value = 0)
df = df.replace(to_replace='Yes',value = 1)
```

```
<ipython-input-10-b4af9c9b07fa>:2: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
  df = df.replace(to_replace='Yes',value = 1)
```

```
new_df=df.drop(['Recurred'],axis=1)
```

```
# Perform one-hot encoding on categorical columns
df_encoded = pd.get_dummies(df, drop_first=True)

# Display the first few rows of the encoded dataset
df_encoded.head()
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred	Thyroid Function_Clinical Hypothyroidism	Thyroid Function_Euthyroid	Thyroid Function_Subclinical Hyperthyroidism	Thyroid Function_Subclinical Hypothyroidism	...	N_N1a	N_N1b	M_M1	Stage_II	Stage_III	Stage_IVA	Stage_IVB	Response_Excellent	Response_Indeterminate	Response_Structural Incomplete
0	27	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	False	True	False
1	34	0	0	1	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False
2	30	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False
3	62	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False
4	62	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False

5 rows × 41 columns

```
df_encoded.columns
```

```
Index(['Age', 'Gender', 'Smoking', 'Hx Smoking', 'Hx Radiothreapy', 'Recurred',
       'Thyroid Function_Clinical Hypothyroidism',
       'Thyroid Function_Euthyroid',
       'Thyroid Function_Subclinical Hyperthyroidism',
       'Thyroid Function_Subclinical Hypothyroidism',
       'Physical Examination_Multinodular goiter',
       'Physical Examination_Normal',
       'Physical Examination_Single nodular goiter-left',
       'Physical Examination_Single nodular goiter-right',
       'Adenopathy_Bilateral', 'Adenopathy_Extensive', 'Adenopathy_Left',
       'Adenopathy_Posterior', 'Adenopathy_Right', 'Pathology_Hurthel cell',
       'Pathology_Micropapillary', 'Pathology_Papillary', 'Focality_Uni-Focal',
       'Risk_Intermediate', 'Risk_Low', 'T_T1b', 'T_T2', 'T_T3a', 'T_T3b',
       'T_T4a', 'T_T4b', 'N_N1a', 'N_N1b', 'M_M1', 'Stage_II', 'Stage_III',
       'Stage_IVA', 'Stage_IVB', 'Response_Excellent',
       'Response_Indeterminate', 'Response_Structural Incomplete'],
      dtype='object')
```

```
# Re-identify categorical columns
#categorical_columns = df.select_dtypes(include=['object']).columns

# Applying binary encoding again
#binary_encoded_dfs = [binary_encode(col) for col in categorical_columns]
#df_binary_encoded = pd.concat([df.drop(columns=categorical_columns)] + binary_encoded_dfs, axis=1)

# Display the first few rows of the binary encoded dataset
#df_binary_encoded.head()
```

```
new_df = df_encoded.replace(to_replace='False',value = 0)
new_df = df_encoded.replace(to_replace='True',value = 1)
```

```
new_df.head(4)
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred	Thyroid Function_Clinical Hypothyroidism	Thyroid Function_Euthyroid	Thyroid Function_Subclinical Hyperthyroidism	Thyroid Function_Subclinical Hypothyroidism	...	N_N1a	N_N1b	M_M1	Stage_II	Stage_III	Stage_IVA	Stage_IVB	Response_Excellent	Response_Indeterminate	Response_Structural Incomplete
0	27	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	False	True	False
1	34	0	0	1	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False
2	30	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False
3	62	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False

4 rows × 41 columns

```
new_df.corr()
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred	Thyroid Function_Clinical Hypothyroidism	Thyroid Function_Euthyroid	Thyroid Function_Subclinical Hyperthyroidism	Thyroid Function_Subclinical Hypothyroidism	...	N_N1a	N_N1b	M_M1	Stage_II	Stage_III	Stage_IVA	Stage_IVB	Response_Excellent	Response_Indeterminate	Response_Structural Incomplete
Age	1.000000	0.186457	0.309536	0.134531	0.176588	0.258897	-0.023205	-0.028367	-0.085732	0.100209	...	-0.051278	0.075087	0.235401	0.369106	0.208210	0.141867	0.336617	-0.258453	0.055762	0.198518
Gender	0.186457	1.000000	0.621886	0.175755	0.235865	0.328189	-0.047227	-0.050344	0.004327	0.086095	...	-0.031137	0.246946	0.211540	0.147333	0.083175	0.110044	0.159335	-0.263805	-0.005657	0.302000
Smoking	0.309536	0.621886	1.000000	0.252773	0.297874	0.333243	-0.024016	-0.010933	-0.044052	0.050354	...	-0.060961	0.220617	0.321233	0.195086	0.191325	0.231977	0.261746	-0.276350	-0.038540	0.318792
Hx Smoking	0.134531	0.175755	0.252773	1.000000	0.261198	0.136073	0.007065	-0.126106	0.056064	0.105639	...	-0.026224	0.051487	0.127209	-0.012303	0.267138	0.088823	0.191920	-0.084694	-0.067416	0.102449
Hx Radiothreapy	0.176588	0.235865	0.297874	0.261198	1.000000	0.174407	-0.024539	-0.061267	-0.015693	-0.026577	...	-0.033683	0.104566	0.430214	0.029243	-0.014017	0.208984	0.443356	-0.109624	-0.059387	0.152818
Recurred	0.258897	0.328189	0.333243	0.136073	0.174407	1.000000	-0.046091	0.074827	-0.072075	0.032535	...	0.094672	0.605927	0.354360	0.335022	0.163932	0.141783	0.274397	-0.671568	-0.161760	0.863540
Thyroid Function_Clinical Hypothyroidism	-0.023205	-0.047227	-0.024016	0.007065	-0.024539	-0.046091	1.000000	-0.458868	-0.020684	-0.035031	...	0.020013	-0.066894	-0.039939	-0.054303	-0.018476	-0.015980	-0.030926	0.044619	0.003636	-0.065186
Thyroid Function_Euthyroid	-0.028367	-0.050344	-0.010933	-0.126106	-0.061267	0.074827	-0.458868	1.000000	-0.293443	-0.496975	...	-0.134440	0.078570	0.014411	0.007251	-0.110926	-0.139526	0.021384	-0.050955	-0.039426	0.074347
Thyroid Function_Subclinical Hyperthyroidism	-0.085732	0.004327	-0.044052	0.056064	-0.015693	-0.072075	-0.020684	-0.293443	1.000000	-0.022402	...	-0.028392	-0.065130	-0.025540	-0.034726	-0.011815	-0.010219	-0.019777	0.105494	-0.050058	-0.064205
Thyroid Function_Subclinical Hypothyroidism	0.100209	0.086095	0.050354	0.105639	-0.026577	0.032535	-0.035031	-0.496975	-0.022402	1.000000	...	0.131297	0.051933	0.022486	0.092026	0.253706	0.140516	-0.033495	-0.072706	0.067307	0.022021
Physical Examination_Multinodular goiter	0.102101	0.084366	0.050136	0.057588	0.017860	0.150881	-0.012026	-0.021666	-0.039533	0.025496	...	-0.000973	0.177085	-0.014849	0.045120	0.028685	0.117060	0.064244	-0.174482	0.069671	0.124026
Physical Examination_Normal	-0.071016	-0.065089	-0.052261	0.036560	-0.018617	0.001131	-0.024539	-0.118639	-0.015693	0.181158	...	-0.033683	0.013649	-0.030300	-0.041198	-0.014017	-0.012123	-0.023463	-0.031365	-0.006120	0.015425
Physical Examination_Single nodular goiter-left	0.020799	0.087516	0.140912	-0.083275	0.017232	0.012412	-0.063466	0.070076	-0.008816	0.024599	...	-0.029551	-0.037644	0.140716	0.034941	0.004287	-0.048887	0.016429	-0.041378	0.030833	0.026926
Physical Examination_Single nodular goiter-right	-0.094108	-0.124909	-0.144643	0.015933	-0.022615	-0.138297	0.081337	0.058124	0.055995	-0.090066	...	0.045628	-0.126376	-0.091704	-0.052845	-0.024647	-0.067442	-0.065600	0.195568	-0.093308	-0.130745
Adenopathy_Bilateral	0.131884	0.268738	0.223335	0.132686	-0.041198	0.376962	-0.000141	0.007251	-0.034726	0.092026	...	-0.074538	0.489175	0.066703	0.113426	0.154612	0.080205	0.117561	-0.310239	-0.028277	0.385682
Adenopathy_Extensive	0.045049	0.135547	0.122806	0.036560	0.417933	0.217726	-0.024539	-0.061267	-0.015693	0.077290	...	-0.033683	0.240942	0.153906	0.099685	-0.014017	0.208984	0.093242	-0.148754	-0.059387	0.244414
Adenopathy_Left	-0.030813	0.027683	0.031315	-0.060527	-0.029406	0.203033	-0.038760	0.047155	-0.024787	-0.041979	...	0.110248	0.262312	0.131850	0.026556	-0.022141	-0.019149	0.038844	-0.158612	0.010131	0.207340
Adenopathy_Posterior	0.029399	-0.034562	-0.027751	-0.020348	-0.009886	0.115613	-0.013030	0.028397	-0.008333	-0.014112	...	-0.017886	0.127941	0.155085	0.239956	-0.007443	-0.006438	-0.012459	-0.078989	-0.031535	0.129785
Adenopathy_Right	-0.008665	0.022360	0.067499	0.075459	0.007225	0.288558	-0.022811	0.055513	-0.043535	0.010314	...	0.076013	0.466124	0.027728	0.085199	0.116257	-0.033633	0.029340	-0.238538	-0.013898	0.270431
Pathology_Hurthel cell	0.108446	0.069234	0.191187	0.114421	0.055590	0.009398	0.092521	-0.080724	0.076399	-0.045721	...	-0.007507	-0.050811	-0.052126	0.098780	-0.024114	0.112275	-0.040363	-0.090981	0.090273	0.061990
Pathology_Micropapillary	0.072205	-0.079106	-0.097766	0.014870	-0.051648	-0.237216	0.022456	-0.083755	0.025940	0.010314	...	-0.093445	-0.214358	-0.084060	-0.114293	-0.038887	-0.033633	-0.065091	0.268050	-0.057000	-0.211314
Pathology_Papillary	-0.164530	0.012346	-0.121169	-0.092151	-0.056014	0.121444	-0.034311	0.110237	-0.039633	0.016345	...	0.090992	0.271319	-0.013899	0.022227	0.059416	-0.016951	-0.044829	-0.083016	-0.028156	0.096387
Focality_Uni-Focal	-0.223847	-0.207634	-0.238494	-0.001204	-0.102415	-0.383776	0.008177	0.014298	0.037274	-0.088057	...	-0.027857	-0.368711	-0.221931	-0.268889	-0.084779	-0.119742	-0.199074	0.359902	-0.005061	-0.393386
Risk_Intermediate	0.062754	0.153387	0.052174	-0.010368	-0.082206	0.462566	-0.006639	-0.007263	-0.069292	0.102972	...	0.181276	0.526667	-0.077974	0.330390	-0.003793	-0.053532	-0.103603	-0.443404	0.141315	0.385329
Risk_Low	-0.228129	-0.269910	-0.276274	-0.088406	-0.145126	-0.708266	0.037660	0.002524	0.084371	-0.119659	...	-0.195349	-0.682578	-0.302717	-0.391810	-0.140042	-0.121120	-0.234408	0.601951	-0.054726	-0.632358
T_T1b	-0.138038	-0.105800	-0.111453	-0.004562	0.013219	-0.130964	-0.016485	-0.006674	0.177682	-0.069270	...	-0.087792	-0.124238	-0.078974	-0.107378	-0.036535	-0.031598	-0.061153	0.176776	-0.041780	-0.120802
T_T2	-0.188722	-0.096133	-0.133058	-0.123951	-0.070191	-0.268105	0.038916	0.033135	-0.045717	-0.043263	...	-0.015468	-0.145358	-0.153912	-0.185670	-0.082881	-0.071683	-0.106741	0.225178	-0.015325	-0.236970
T_T3a	0.058107	0.080672	0.030987	-0.046710	-0.078913	0.186500	-0.034852	0.084819	-0.066517	-0.016345	...	0.064357	0.136136	-0.043035	0.239035	-0.059416	-0.051388	-0.099453	-0.316087	0.192794	0.101783
T_T3b	0.039829	0.068303	0.076302	0.141887	-0.028489	0.275178	0.037356	-0.110217	-0.024014	0.167947	...	0.228876	0.125234	0.138620	0.125597	0.106917	-0.018552	0.042229	-0.201438	-0.055216	0.312699
T_T4a	0.242001	0.099435	0.261460	0.114421	0.143207	0.348473	-0.042215	-0.011635	-0.026996	0.079358	...	-0.007507	0.222898	0.336066	0.141193	0.322225	0.112275	0.310981	-0.232342	-0.102164	0.365317
T_T4b	0.206634	0.259198	0.326673	0.169390	0.388970	0.233069	-0.026268	0.003508	-0.016798	-0.028450	...	-0.036057	0.257921	0.312641	0.021878	-0.015005	0.401228	0.412134	-0.159236	-0.063572	0.218741
N_N1a	-0.051278	-0.031137	-0.060961	-0.026224	-0.033683	0.094672	0.020013	-0.134440	-0.028392	0.131297	...	1.000000	-0.139798	0.104243	0.087667	-0.025361	-0.021934	0.024734	-0.156505	0.137872	0.020376
N_N1b	0.075087	0.246946	0.220617	0.051487	0.104566	0.605927	-0.066894	0.078570	-0.065130	0.051933	...	-0.139798	1.000000	0.190739	0.225110	0.181413	0.087829	0.121367	-0.458474	-0.046793	0.613815
M_M1	0.235401	0.211540	0.321233	0.127209	0.430214	0.354360	-0.039939	0.014411	-0.025540	0.022486	...	0.104243	0.190739	1.000000	0.245042	-0.022814	-0.019731	0.700479	-0.242104	-0.096656	0.368809
Stage_II	0.369106	0.147333	0.195086	-0.012303	0.029243	0.335022	-0.054303	0.007251	-0.034726	0.092026	...	0.087667	0.225110	0.245042	1.000000	-0.031019	-0.026828	-0.051921	-0.272355	0.074866	0.252664
Stage_III	0.208210	0.083175	0.191325	0.267138	-0.014017	0.163932	-0.018476	-0.110926	-0.011815	0.253706	...	-0.025361	0.181413	-0.022814	-0.031019	1.000000	-0.009128	-0.017666	-0.112001	-0.044714	0.184027
Stage_IVA	0.141867	0.110044	0.231977	0.088823	0.208984	0.141783	-0.015980	-0.139526	-0.010219	0.140516	...	-0.021934	0.087829	-0.019731	-0.026828	-0.009128	1.000000	-0.015279	-0.096868	-0.038673	0.089574
Stage_IVB	0.336617	0.159335	0.261746	0.191920	0.443356	0.274397	-0.030926	0.021384	-0.019777	-0.033495	...	0.024734	0.121367	0.700479	-0.051921	-0.017666	-0.015279	1.000000	-0.187472	-0.074845	0.271302
Response_Excellent	-0.258453	-0.263805	-0.276350	-0.084694	-0.109624	-0.671568	0.044619	-0.050955	0.105494	-0.072706	...	-0.156505	-0.458474	-0.242104	-0.272355	-0.112001	-0.096868	-0.187472	1.000000	-0.474515	-0.608614
Response_Indeterminate	0.055762	-0.005657	-0.038540	-0.067416	-0.059387	-0.161760	0.003636	-0.039426	-0.050058	0.067307	...	0.137872	-0.046793	-0.096656	0.074866	-0.044714	-0.038673	-0.074845	-0.474515	1.000000	-0.242978
Response_Structural Incomplete	0.198518	0.302000	0.318792	0.102449	0.152818	0.863540	-0.065186	0.074347	-0.064205	0.022021	...	0.020376	0.613815	0.368809	0.252664	0.184027	0.089574	0.271302	-0.608614	-0.242978	1.000000

41 rows × 41 columns

```
new_df.dtypes
```

	0
Age	int64
Gender	int64
Smoking	int64
Hx Smoking	int64
Hx Radiothreapy	int64
Recurred	int64
Thyroid Function_Clinical Hypothyroidism	bool
Thyroid Function_Euthyroid	bool
Thyroid Function_Subclinical Hyperthyroidism	bool
Thyroid Function_Subclinical Hypothyroidism	bool
Physical Examination_Multinodular goiter	bool
Physical Examination_Normal	bool
Physical Examination_Single nodular goiter-left	bool
Physical Examination_Single nodular goiter-right	bool
Adenopathy_Bilateral	bool
Adenopathy_Extensive	bool
Adenopathy_Left	bool
Adenopathy_Posterior	bool
Adenopathy_Right	bool
Pathology_Hurthel cell	bool
Pathology_Micropapillary	bool
Pathology_Papillary	bool
Focality_Uni-Focal	bool
Risk_Intermediate	bool
Risk_Low	bool
T_T1b	bool
T_T2	bool
T_T3a	bool
T_T3b	bool
T_T4a	bool
T_T4b	bool
N_N1a	bool
N_N1b	bool
M_M1	bool
Stage_II	bool
Stage_III	bool
Stage_IVA	bool
Stage_IVB	bool
Response_Excellent	bool
Response_Indeterminate	bool
Response_Structural Incomplete	bool

dtype: object

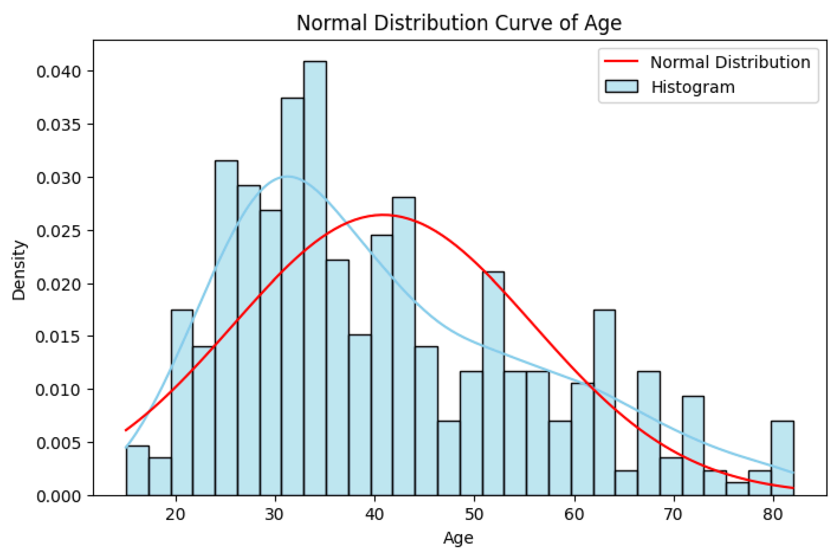
```
new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 383 entries, 0 to 382
Data columns (total 41 columns):
 #   Column                                            Non-Null Count  Dtype
---  ------                                            --------------  -----
 0   Age                                               383 non-null    int64
 1   Gender                                            383 non-null    int64
 2   Smoking                                           383 non-null    int64
 3   Hx Smoking                                        383 non-null    int64
 4   Hx Radiothreapy                                   383 non-null    int64
 5   Recurred                                          383 non-null    int64
 6   Thyroid Function_Clinical Hypothyroidism          383 non-null    bool
 7   Thyroid Function_Euthyroid                        383 non-null    bool
 8   Thyroid Function_Subclinical Hyperthyroidism      383 non-null    bool
 9   Thyroid Function_Subclinical Hypothyroidism       383 non-null    bool
 10  Physical Examination_Multinodular goiter          383 non-null    bool
 11  Physical Examination_Normal                       383 non-null    bool
 12  Physical Examination_Single nodular goiter-left   383 non-null    bool
 13  Physical Examination_Single nodular goiter-right  383 non-null    bool
 14  Adenopathy_Bilateral                              383 non-null    bool
 15  Adenopathy_Extensive                              383 non-null    bool
 16  Adenopathy_Left                                   383 non-null    bool
 17  Adenopathy_Posterior                              383 non-null    bool
 18  Adenopathy_Right                                  383 non-null    bool
 19  Pathology_Hurthel cell                            383 non-null    bool
 20  Pathology_Micropapillary                          383 non-null    bool
 21  Pathology_Papillary                               383 non-null    bool
 22  Focality_Uni-Focal                                383 non-null    bool
 23  Risk_Intermediate                                 383 non-null    bool
 24  Risk_Low                                          383 non-null    bool
 25  T_T1b                                             383 non-null    bool
 26  T_T2                                              383 non-null    bool
 27  T_T3a                                             383 non-null    bool
 28  T_T3b                                             383 non-null    bool
 29  T_T4a                                             383 non-null    bool
 30  T_T4b                                             383 non-null    bool
 31  N_N1a                                             383 non-null    bool
 32  N_N1b                                             383 non-null    bool
 33  M_M1                                              383 non-null    bool
 34  Stage_II                                          383 non-null    bool
 35  Stage_III                                         383 non-null    bool
 36  Stage_IVA                                         383 non-null    bool
 37  Stage_IVB                                         383 non-null    bool
 38  Response_Excellent                                383 non-null    bool
 39  Response_Indeterminate                            383 non-null    bool
 40  Response_Structural Incomplete                    383 non-null    bool
dtypes: bool(35), int64(6)
memory usage: 31.2 KB
```

```
new_df.describe()
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred
count	383.000000	383.000000	383.000000	383.000000	383.000000	383.000000
mean	40.866841	0.185379	0.127937	0.073107	0.018277	0.281984
std	15.134494	0.389113	0.334457	0.260653	0.134126	0.450554
min	15.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	29.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	37.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	51.000000	0.000000	0.000000	0.000000	0.000000	1.000000
max	82.000000	1.000000	1.000000	1.000000	1.000000	1.000000

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import norm

# Plot the normal distribution curve for "Age"
plt.figure(figsize=(8, 5))
sns.histplot(df["Age"], bins=30, kde=True, stat="density", color="skyblue", label="Histogram")

# Overlay the normal distribution curve
mean_age = np.mean(df["Age"])
std_age = np.std(df["Age"])
x_values = np.linspace(min(df["Age"]), max(df["Age"]), 100)
y_values = norm.pdf(x_values, mean_age, std_age)
plt.plot(x_values, y_values, color="red", label="Normal Distribution")

# Labels and title
plt.title("Normal Distribution Curve of Age")
plt.xlabel("Age")
plt.ylabel("Density")
plt.legend()
plt.show()
```



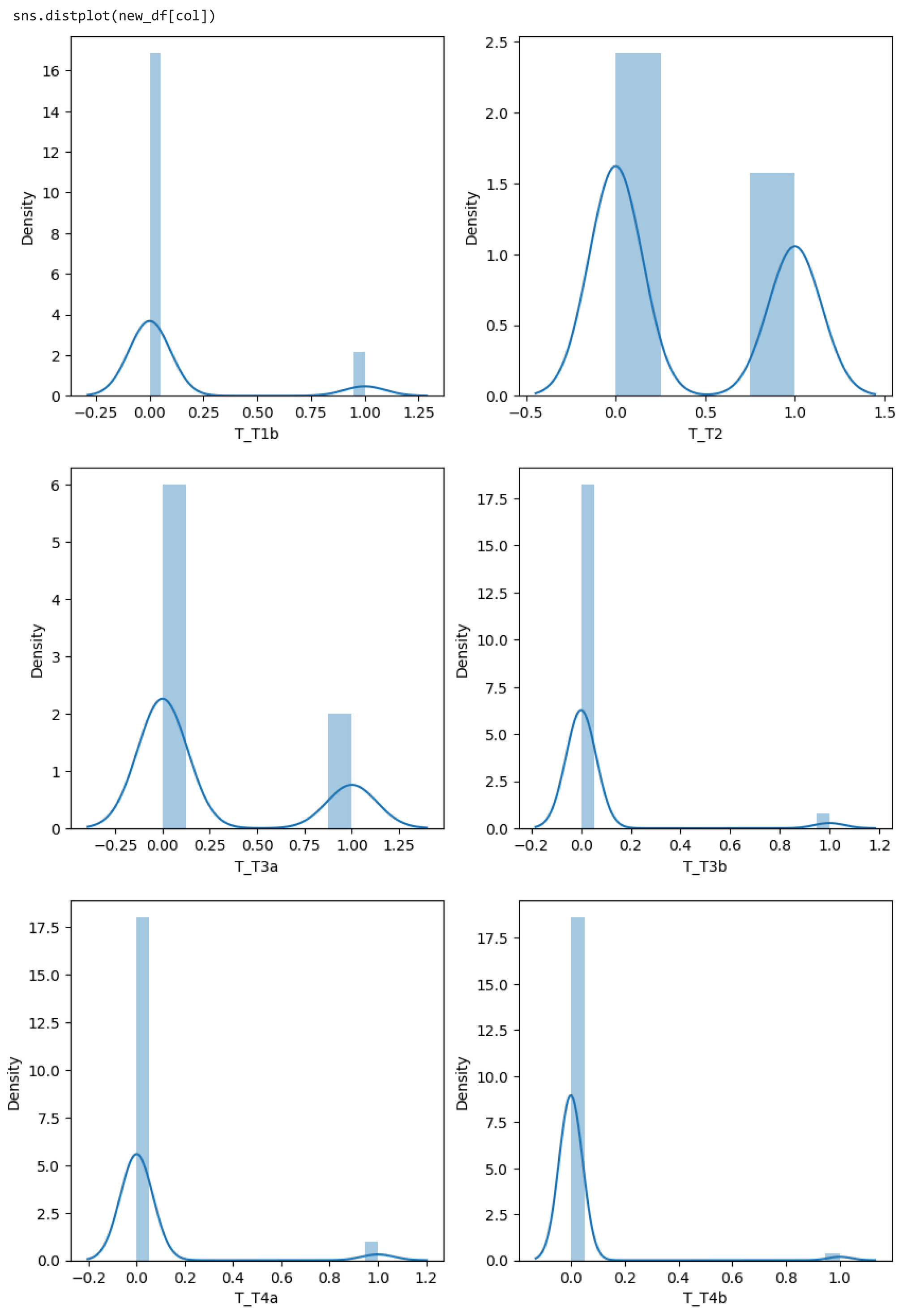

```
new_df['T_T1b'].dtypes
```

```
dtype('bool')
```

```
new_df['T_T1b'] = new_df['T_T1b'].astype(float)
new_df['T_T2'] = new_df['T_T2'].astype(float)
new_df['T_T3a'] = new_df['T_T3a'].astype(float)
new_df['T_T3b'] = new_df['T_T3b'].astype(float)
new_df['T_T4a'] = new_df['T_T4a'].astype(float)
new_df['T_T4b'] = new_df['T_T4b'].astype(float)
```

```
columns = ['T_T1b','T_T2','T_T3a','T_T3b','T_T4a','T_T4b']  # Changed 'T_3a' to 'T_T3a'
plt.figure(figsize= (10,15),facecolor = 'white')
plotnumber = 1
for col in columns:
  ax = plt.subplot(3,2,plotnumber)
  sns.distplot(new_df[col])
  plt.xlabel(col,fontsize = 10)
  plotnumber+=1
plt.show()
```

```
<ipython-input-24-01a342c1cb55>:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(new_df[col])
<ipython-input-24-01a342c1cb55>:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(new_df[col])
<ipython-input-24-01a342c1cb55>:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(new_df[col])
<ipython-input-24-01a342c1cb55>:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(new_df[col])
<ipython-input-24-01a342c1cb55>:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(new_df[col])
<ipython-input-24-01a342c1cb55>:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(new_df[col])
```

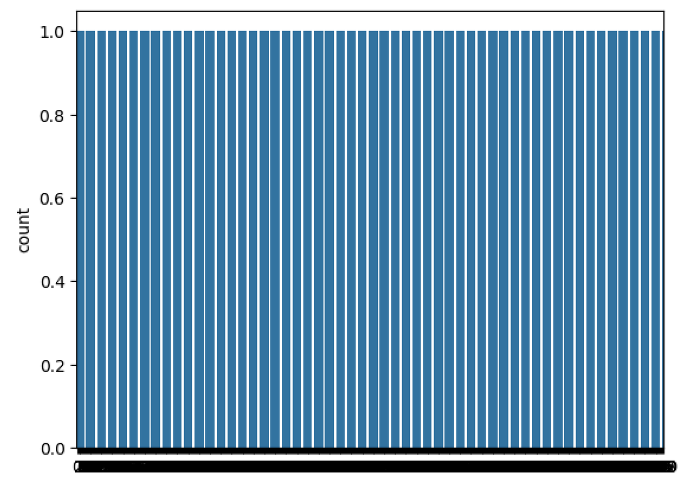



```
target = df['Recurred']
```

```
rdsample=RandomOverSampler()
target = df['Recurred']
x_sampled,y_sampled=rdsample.fit_resample(df,target)
```

```
sns.countplot(y_sampled)
```

```
<Axes: ylabel='count'>
```




```
target.value_counts()
```

Recurred	count
0	275
1	108

dtype: int64

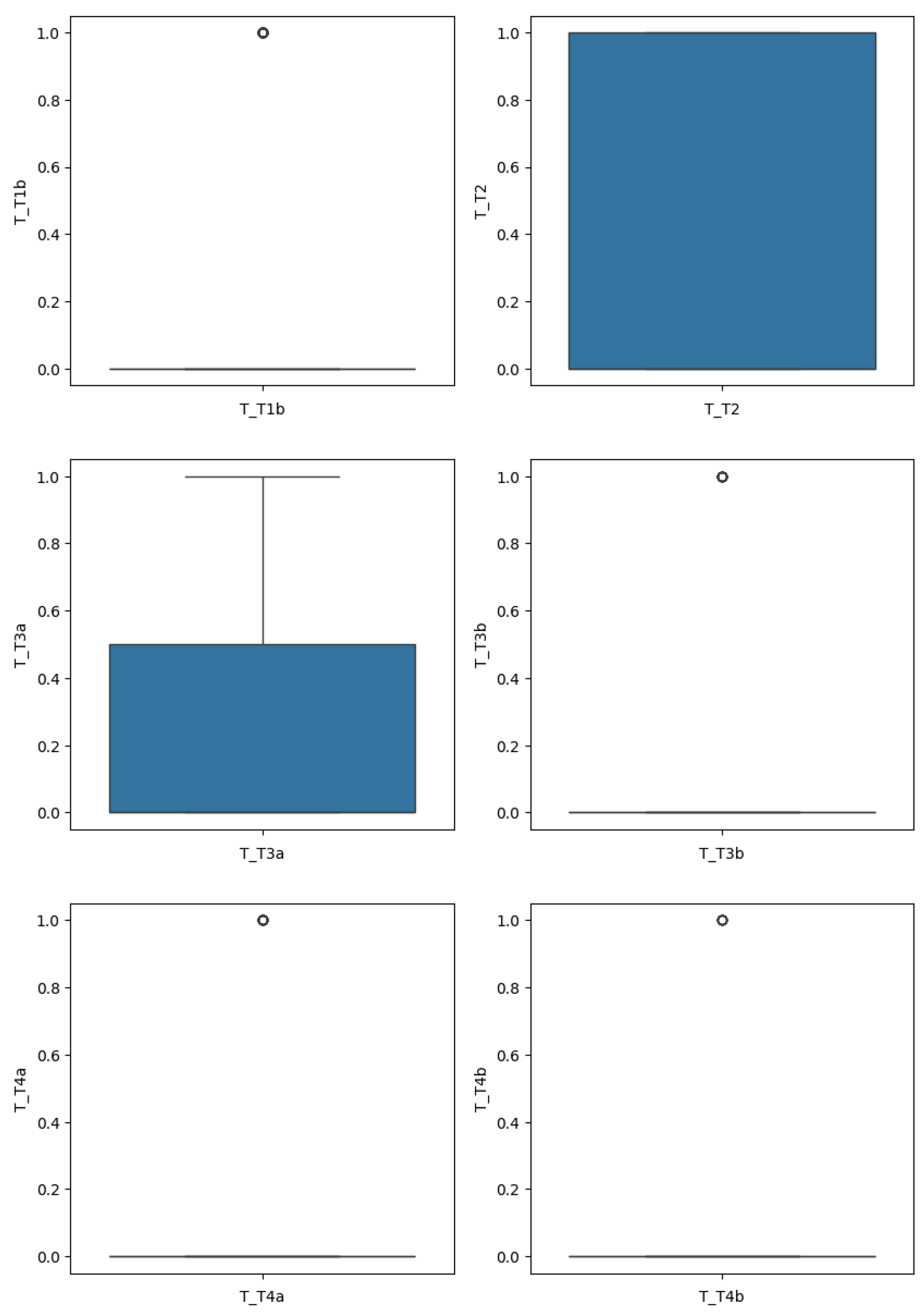
```
x_sampled=new_df
```

```
x_sampled.head(3)
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred	Thyroid Function_Clinical Hypothyroidism	Thyroid Function_Euthyroid	Thyroid Function_Subclinical Hyperthyroidism	Thyroid Function_Subclinical Hypothyroidism	...	N_N1a	N_N1b	M_M1	Stage_II	Stage_III	Stage_IVA	Stage_IVB	Response_Excellent	Response_Indeterminate	Response_Structural Incomplete
0	27	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	False	True	False
1	34	0	0	1	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False
2	30	0	0	0	0	0	False	True	False	False	...	False	False	False	False	False	False	False	True	False	False

3 rows × 41 columns

```
columns = ['T_T1b','T_T2','T_T3a','T_T3b','T_T4a','T_T4b']  # Changed 'T_3a' to 'T_T3a'
plt.figure(figsize= (10,15),facecolor = 'white')
plotnumber = 1
for col in columns:
  ax = plt.subplot(3,2,plotnumber)
  sns.boxplot(new_df[col])
  plt.xlabel(col,fontsize = 10)
  plotnumber+=1
plt.show()
#T_T1b, T_T3b,T_T4a,T_T4b has outliers so we have to remove it.
```




```
winsorizer=Winsorizer(capping_method='gaussian',tail='both',fold=1.5,variables=['T_T4a'])
x_sampled['T_T4a']=winsorizer.fit_transform(x_sampled[['T_T4a']])
winsorizer=Winsorizer(capping_method='gaussian',tail='both',fold=1.5,variables=['T_T1b'])
x_sampled['T_T1b']=winsorizer.fit_transform(x_sampled[['T_T1b']])
winsorizer=Winsorizer(capping_method='gaussian',tail='both',fold=1.5,variables=['T_T4b'])
x_sampled['T_T4b']=winsorizer.fit_transform(x_sampled[['T_T4b']])
winsorizer=Winsorizer(capping_method='gaussian',tail='both',fold=1.5,variables=['T_T3b'])
x_sampled['T_T3b']=winsorizer.fit_transform(x_sampled[['T_T3b']])
```

```
# Convert all columns of x_sampled to numeric, coercing errors to NaN
for col in x_sampled.columns:
    x_sampled[col] = pd.to_numeric(x_sampled[col], errors='coerce')

# Impute NaN values if any (replace with mean, median, or other strategy)
imputer = SimpleImputer(strategy='mean') # Choose an appropriate strategy
x_sampled = pd.DataFrame(imputer.fit_transform(x_sampled), columns=x_sampled.columns)


def calc_vif(X):
    # Calculating VIF
    vif = pd.DataFrame()
    vif["variables"] = X.columns
    vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
    return(vif)

calc_vif(x_sampled)
```

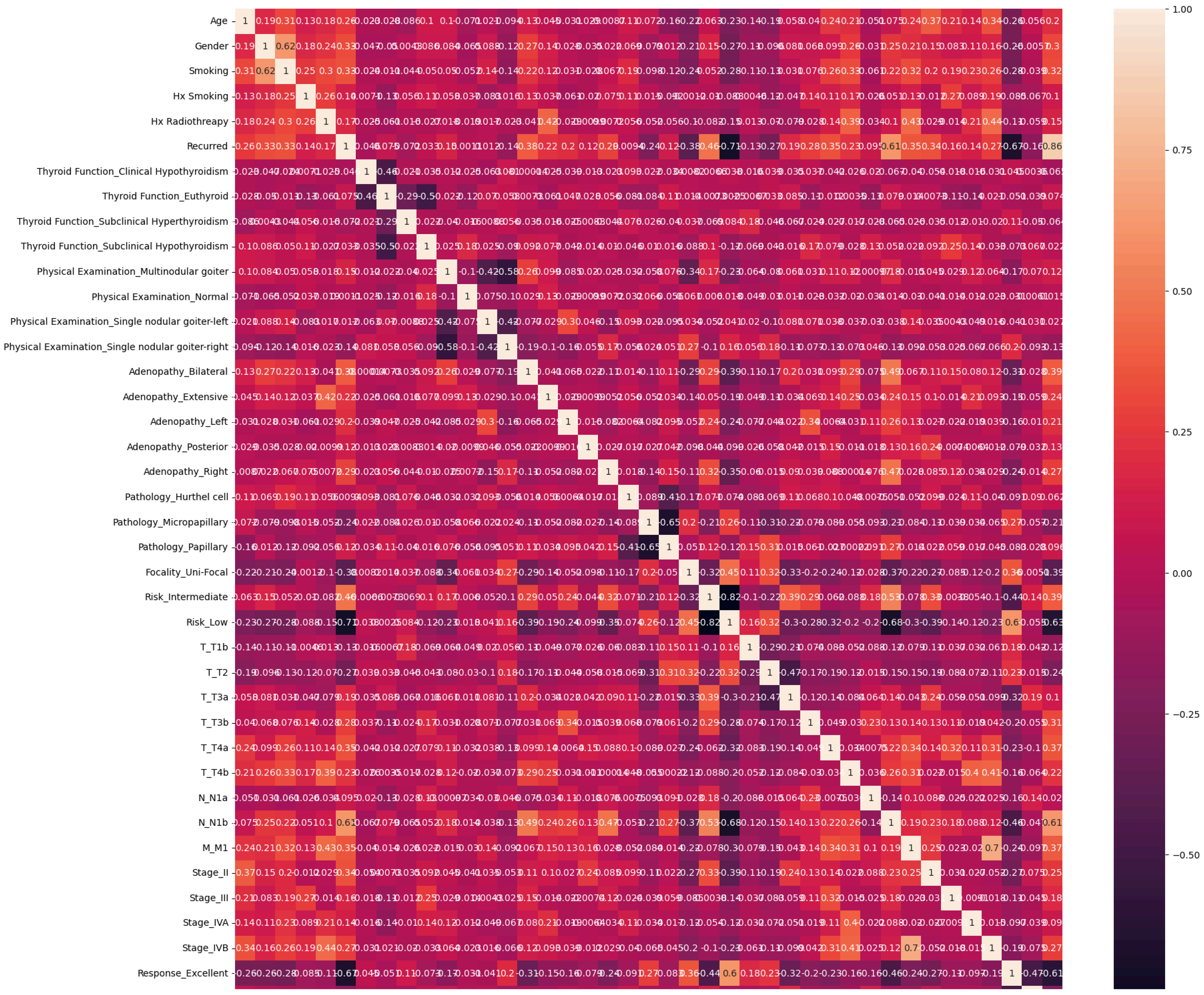
	variables	VIF
0	Age	14.618290
1	Gender	2.326394
2	Smoking	2.604143
3	Hx Smoking	1.513784
4	Hx Radiothreapy	1.936078
5	Recurred	8.885628
6	Thyroid Function_Clinical Hypothyroidism	1.893348
7	Thyroid Function_Euthyroid	23.811943
8	Thyroid Function_Subclinical Hyperthyroidism	1.409991
9	Thyroid Function_Subclinical Hypothyroidism	2.134751
10	Physical Examination_Multinodular goiter	23.651938
11	Physical Examination_Normal	2.019277
12	Physical Examination_Single nodular goiter-left	15.829113
13	Physical Examination_Single nodular goiter-right	24.198110
14	Adenopathy_Bilateral	3.564836
15	Adenopathy_Extensive	1.979320
16	Adenopathy_Left	2.206877
17	Adenopathy_Posterior	1.337788
18	Adenopathy_Right	3.450868
19	Pathology_Hurthel cell	2.035612
20	Pathology_Micropapillary	12.129033
21	Pathology_Papillary	14.015717
22	Focality_Uni-Focal	4.996959
23	Risk_Intermediate	15.091287
24	Risk_Low	43.818909
25	T_T1b	10.413646
26	T_T2	35.405907
27	T_T3a	21.334791
28	T_T3b	4.916254
29	T_T4a	5.269054
30	T_T4b	3.619553
31	N_N1a	1.647335
32	N_N1b	8.476848
33	M_M1	3.587724
34	Stage_II	2.221883
35	Stage_III	1.679713
36	Stage_IVA	1.710554
37	Stage_IVB	3.656950
38	Response_Excellent	13.842402
39	Response_Indeterminate	4.350889
40	Response_Structural Incomplete	8.054639

```
x_sampled.corr()
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred	Thyroid Function_Clinical Hypothyroidism	Thyroid Function_Euthyroid	Thyroid Function_Subclinical Hyperthyroidism	Thyroid Function_Subclinical Hypothyroidism	...	N_N1a	N_N1b	M_M1	Stage_II	Stage_III	Stage_IVA	Stage_IVB	Response_Excellent	Response_Indeterminate	Response_Structural Incomplete
Age	1.000000	0.186457	0.309536	0.134531	0.176588	0.258897	-0.023205	-0.028367	-0.085732	0.100209	...	-0.051278	0.075087	0.235401	0.369106	0.208210	0.141867	0.336617	-0.258453	0.055762	0.198518
Gender	0.186457	1.000000	0.621886	0.175755	0.235865	0.328189	-0.047227	-0.050344	0.004327	0.086095	...	-0.031137	0.246946	0.211540	0.147333	0.083175	0.110044	0.159335	-0.263805	-0.005657	0.302000
Smoking	0.309536	0.621886	1.000000	0.252773	0.297874	0.333243	-0.024016	-0.010933	-0.044052	0.050354	...	-0.060961	0.220617	0.321233	0.195086	0.191325	0.231977	0.261746	-0.276350	-0.038540	0.318792
Hx Smoking	0.134531	0.175755	0.252773	1.000000	0.261198	0.136073	0.007065	-0.126106	0.056064	0.105639	...	-0.026224	0.051487	0.127209	-0.012303	0.267138	0.088823	0.191920	-0.084694	-0.067416	0.102449
Hx Radiothreapy	0.176588	0.235865	0.297874	0.261198	1.000000	0.174407	-0.024539	-0.061267	-0.015693	-0.026577	...	-0.033683	0.104566	0.430214	0.029243	-0.014017	0.208984	0.443356	-0.109624	-0.059387	0.152818
Recurred	0.258897	0.328189	0.333243	0.136073	0.174407	1.000000	-0.046091	0.074827	-0.072075	0.032535	...	0.094672	0.605927	0.354360	0.335022	0.163932	0.141783	0.274397	-0.671568	-0.161760	0.863540
Thyroid Function_Clinical Hypothyroidism	-0.023205	-0.047227	-0.024016	0.007065	-0.024539	-0.046091	1.000000	-0.458868	-0.020684	-0.035031	...	0.020013	-0.066894	-0.039939	-0.054303	-0.018476	-0.015980	-0.030926	0.044619	0.003636	-0.065186
Thyroid Function_Euthyroid	-0.028367	-0.050344	-0.010933	-0.126106	-0.061267	0.074827	-0.458868	1.000000	-0.293443	-0.496975	...	-0.134440	0.078570	0.014411	0.007251	-0.110926	-0.139526	0.021384	-0.050955	-0.039426	0.074347
Thyroid Function_Subclinical Hyperthyroidism	-0.085732	0.004327	-0.044052	0.056064	-0.015693	-0.072075	-0.020684	-0.293443	1.000000	-0.022402	...	-0.028392	-0.065130	-0.025540	-0.034726	-0.011815	-0.010219	-0.019777	0.105494	-0.050058	-0.064205
Thyroid Function_Subclinical Hypothyroidism	0.100209	0.086095	0.050354	0.105639	-0.026577	0.032535	-0.035031	-0.496975	-0.022402	1.000000	...	0.131297	0.051933	0.022486	0.092026	0.253706	0.140516	-0.033495	-0.072706	0.067307	0.022021
Physical Examination_Multinodular goiter	0.102101	0.084366	0.050136	0.057588	0.017860	0.150881	-0.012026	-0.021666	-0.039533	0.025496	...	-0.000973	0.177085	-0.014849	0.045120	0.028685	0.117060	0.064244	-0.174482	0.069671	0.124026
Physical Examination_Normal	-0.071016	-0.065089	-0.052261	0.036560	-0.018617	0.001131	-0.024539	-0.118639	-0.015693	0.181158	...	-0.033683	0.013649	-0.030300	-0.041198	-0.014017	-0.012123	-0.023463	-0.031365	-0.006120	0.015425
Physical Examination_Single nodular goiter-left	0.020799	0.087516	0.140912	-0.083275	0.017232	0.012412	-0.063466	0.070076	-0.008816	0.024599	...	-0.029551	-0.037644	0.140716	0.034941	0.004287	-0.048887	0.016429	-0.041378	0.030833	0.026926
Physical Examination_Single nodular goiter-right	-0.094108	-0.124909	-0.144643	0.015933	-0.022615	-0.138297	0.081337	0.058124	0.055995	-0.090066	...	0.045628	-0.126376	-0.091704	-0.052845	-0.024647	-0.067442	-0.065600	0.195568	-0.093308	-0.130745
Adenopathy_Bilateral	0.131884	0.268738	0.223335	0.132686	-0.041198	0.376962	-0.000141	0.007251	-0.034726	0.092026	...	-0.074538	0.489175	0.066703	0.113426	0.154612	0.080205	0.117561	-0.310239	-0.028277	0.385682
Adenopathy_Extensive	0.045049	0.135547	0.122806	0.036560	0.417933	0.217726	-0.024539	-0.061267	-0.015693	0.077290	...	-0.033683	0.240942	0.153906	0.099685	-0.014017	0.208984	0.093242	-0.148754	-0.059387	0.244414
Adenopathy_Left	-0.030813	0.027683	0.031315	-0.060527	-0.029406	0.203033	-0.038760	0.047155	-0.024787	-0.041979	...	0.110248	0.262312	0.131850	0.026556	-0.022141	-0.019149	0.038844	-0.158612	0.010131	0.207340
Adenopathy_Posterior	0.029399	-0.034562	-0.027751	-0.020348	-0.009886	0.115613	-0.013030	0.028397	-0.008333	-0.014112	...	-0.017886	0.127941	0.155085	0.239956	-0.007443	-0.006438	-0.012459	-0.078989	-0.031535	0.129785
Adenopathy_Right	-0.008665	0.022360	0.067499	0.075459	0.007225	0.288558	-0.022811	0.055513	-0.043535	0.010314	...	0.076013	0.466124	0.027728	0.085199	0.116257	-0.033633	0.029340	-0.238538	-0.013898	0.270431
Pathology_Hurthel cell	0.108446	0.069234	0.191187	0.114421	0.055590	0.009398	0.092521	-0.080724	0.076399	-0.045721	...	-0.007507	-0.050811	-0.052126	0.098780	-0.024114	0.112275	-0.040363	-0.090981	0.090273	0.061990
Pathology_Micropapillary	0.072205	-0.079106	-0.097766	0.014870	-0.051648	-0.237216	0.022456	-0.083755	0.025940	0.010314	...	-0.093445	-0.214358	-0.084060	-0.114293	-0.038887	-0.033633	-0.065091	0.268050	-0.057000	-0.211314
Pathology_Papillary	-0.164530	0.012346	-0.121169	-0.092151	-0.056014	0.121444	-0.034311	0.110237	-0.039633	0.016345	...	0.090992	0.271319	-0.013899	0.022227	0.059416	-0.016951	-0.044829	-0.083016	-0.028156	0.096387
Focality_Uni-Focal	-0.223847	-0.207634	-0.238494	-0.001204	-0.102415	-0.383776	0.008177	0.014298	0.037274	-0.088057	...	-0.027857	-0.368711	-0.221931	-0.268889	-0.084779	-0.119742	-0.199074	0.359902	-0.005061	-0.393386
Risk_Intermediate	0.062754	0.153387	0.052174	-0.010368	-0.082206	0.462566	-0.006639	-0.007263	-0.069292	0.102972	...	0.181276	0.526667	-0.077974	0.330390	-0.003793	-0.053532	-0.103603	-0.443404	0.141315	0.385329
Risk_Low	-0.228129	-0.269910	-0.276274	-0.088406	-0.145126	-0.708266	0.037660	0.002524	0.084371	-0.119659	...	-0.195349	-0.682578	-0.302717	-0.391810	-0.140042	-0.121120	-0.234408	0.601951	-0.054726	-0.632358
T_T1b	-0.138038	-0.105800	-0.111453	-0.004562	0.013219	-0.130964	-0.016485	-0.006674	0.177682	-0.069270	...	-0.087792	-0.124238	-0.078974	-0.107378	-0.036535	-0.031598	-0.061153	0.176776	-0.041780	-0.120802
T_T2	-0.188722	-0.096133	-0.133058	-0.123951	-0.070191	-0.268105	0.038916	0.033135	-0.045717	-0.043263	...	-0.015468	-0.145358	-0.153912	-0.185670	-0.082881	-0.071683	-0.106741	0.225178	-0.015325	-0.236970
T_T3a	0.058107	0.080672	0.030987	-0.046710	-0.078913	0.186500	-0.034852	0.084819	-0.066517	-0.016345	...	0.064357	0.136136	-0.043035	0.239035	-0.059416	-0.051388	-0.099453	-0.316087	0.192794	0.101783
T_T3b	0.039829	0.068303	0.076302	0.141887	-0.028489	0.275178	0.037356	-0.110217	-0.024014	0.167947	...	0.228876	0.125234	0.138620	0.125597	0.106917	-0.018552	0.042229	-0.201438	-0.055216	0.312699
T_T4a	0.242001	0.099435	0.261460	0.114421	0.143207	0.348473	-0.042215	-0.011635	-0.026996	0.079358	...	-0.007507	0.222898	0.336066	0.141193	0.322225	0.112275	0.310981	-0.232342	-0.102164	0.365317
T_T4b	0.206634	0.259198	0.326673	0.169390	0.388970	0.233069	-0.026268	0.003508	-0.016798	-0.028450	...	-0.036057	0.257921	0.312641	0.021878	-0.015005	0.401228	0.412134	-0.159236	-0.063572	0.218741
N_N1a	-0.051278	-0.031137	-0.060961	-0.026224	-0.033683	0.094672	0.020013	-0.134440	-0.028392	0.131297	...	1.000000	-0.139798	0.104243	0.087667	-0.025361	-0.021934	0.024734	-0.156505	0.137872	0.020376
N_N1b	0.075087	0.246946	0.220617	0.051487	0.104566	0.605927	-0.066894	0.078570	-0.065130	0.051933	...	-0.139798	1.000000	0.190739	0.225110	0.181413	0.087829	0.121367	-0.458474	-0.046793	0.613815
M_M1	0.235401	0.211540	0.321233	0.127209	0.430214	0.354360	-0.039939	0.014411	-0.025540	0.022486	...	0.104243	0.190739	1.000000	0.245042	-0.022814	-0.019731	0.700479	-0.242104	-0.096656	0.368809
Stage_II	0.369106	0.147333	0.195086	-0.012303	0.029243	0.335022	-0.054303	0.007251	-0.034726	0.092026	...	0.087667	0.225110	0.245042	1.000000	-0.031019	-0.026828	-0.051921	-0.272355	0.074866	0.252664
Stage_III	0.208210	0.083175	0.191325	0.267138	-0.014017	0.163932	-0.018476	-0.110926	-0.011815	0.253706	...	-0.025361	0.181413	-0.022814	-0.031019	1.000000	-0.009128	-0.017666	-0.112001	-0.044714	0.184027
Stage_IVA	0.141867	0.110044	0.231977	0.088823	0.208984	0.141783	-0.015980	-0.139526	-0.010219	0.140516	...	-0.021934	0.087829	-0.019731	-0.026828	-0.009128	1.000000	-0.015279	-0.096868	-0.038673	0.089574
Stage_IVB	0.336617	0.159335	0.261746	0.191920	0.443356	0.274397	-0.030926	0.021384	-0.019777	-0.033495	...	0.024734	0.121367	0.700479	-0.051921	-0.017666	-0.015279	1.000000	-0.187472	-0.074845	0.271302
Response_Excellent	-0.258453	-0.263805	-0.276350	-0.084694	-0.109624	-0.671568	0.044619	-0.050955	0.105494	-0.072706	...	-0.156505	-0.458474	-0.242104	-0.272355	-0.112001	-0.096868	-0.187472	1.000000	-0.474515	-0.608614
Response_Indeterminate	0.055762	-0.005657	-0.038540	-0.067416	-0.059387	-0.161760	0.003636	-0.039426	-0.050058	0.067307	...	0.137872	-0.046793	-0.096656	0.074866	-0.044714	-0.038673	-0.074845	-0.474515	1.000000	-0.242978
Response_Structural Incomplete	0.198518	0.302000	0.318792	0.102449	0.152818	0.863540	-0.065186	0.074347	-0.064205	0.022021	...	0.020376	0.613815	0.368809	0.252664	0.184027	0.089574	0.271302	-0.608614	-0.242978	1.000000

41 rows × 41 columns

```
plt.figure(figsize=(20,20))
sns.heatmap(data=x_sampled.corr(),annot=True)
plt.show()
```




```
rdsample=RandomOverSampler()
target = df['Recurred']
x_sampled,y_sampled=rdsample.fit_resample(df,target)
train_set, test_set,train_label,test_label = train_test_split(x_sampled,y_sampled,test_size = 0.33,random_state = 42)
print(train_set.shape)
```

```
(368, 17)
```

```
y = y_sampled.copy()
```

```
scaler = StandardScaler()
```

```
numerical_cols = train_set.select_dtypes(include=np.number).columns
train_set_numerical = train_set[numerical_cols]
train_set_scaled = scaler.fit_transform(train_set_numerical)
```

```
from sklearn import svm
```

```
params = {'kernel':['linear','poly','rbf'],'degree':[3,4]}
```

```
KNN = KNeighborsClassifier(n_neighbors=2)
# Fit the KNN model using the scaled training data
KNN.fit(train_set_scaled, train_label)
# Predict using the scaled test data
predicted_values_KNN = KNN.predict(test_set_scaled)
print(predicted_values_KNN)
accuracy_KNN = accuracy_score(test_label, predicted_values_KNN)
print(accuracy_KNN)
```

```
[0 0 1 0 1 1 0 1 0 1 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 0 1 1 0 1 1 0 0 0 0
 0 1 1 0 0 1 0 0 1 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 1 0 0 0 0 0 1 1 1 0 1 1 0
 0 1 0 0 1 0 0 1 1 0 1 0 1 1 1 1 1 0 1 0 0 0 0 0 1 0 0 1 1 0 0 1 0 1 0 1 1
 0 0 1 0 0 1 0 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 0 0 1 1 0 0 1 1 1 0 1 1 1 1
 0 1 1 1 1 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 1 1 1 0 0 0 1 0 0 0 1 0 1 0]
1.0
```

```
train_set
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
57	23	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
255	37	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
24	60	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
17	44	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
237	36	1	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
71	69	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
106	26	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
270	39	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
435	63	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1
102	27	0	0	0	0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0

368 rows × 17 columns

Next steps: Generate code with `train_set` View recommended plots New interactive sheet

```
test_set
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
195	61	1	0	0	0	Subclinical Hypothyroidism	Single nodular goiter-left	0	Papillary	Uni-Focal	Low	T2	N0	M0	I	Excellent	0
79	50	0	0	0	0	Euthyroid	Multinodular goiter	0	Papillary	Multi-Focal	Low	T1b	N0	M0	I	Excellent	0
480	40	1	1	0	0	Euthyroid	Multinodular goiter	Bilateral	Papillary	Multi-Focal	High	T4b	N1b	M0	I	Structural Incomplete	1
109	60	0	0	0	0	Euthyroid	Single nodular goiter-right	0	Papillary	Uni-Focal	Low	T2	N0	M0	I	Biochemical Incomplete	0
522	51	0	0	0	0	Euthyroid	Multinodular goiter	0	Papillary	Multi-Focal	High	T4a	N1a	M1	II	Structural Incomplete	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
113	32	0	0	0	0	Euthyroid	Single nodular goiter-right	0	Papillary	Uni-Focal	Low	T2	N0	M0	I	Excellent	0
304	26	0	0	0	0	Euthyroid	Single nodular goiter-left	Left	Papillary	Uni-Focal	Intermediate	T3a	N1b	M1	I	Structural Incomplete	1
173	30	0	0	0	0	Euthyroid	Normal	0	Papillary	Uni-Focal	Low	T2	N0	M0	I	Indeterminate	0
362	80	0	1	1	0	Euthyroid	Multinodular goiter	Right	Papillary	Uni-Focal	High	T4a	N1b	M0	III	Structural Incomplete	1
208	24	0	0	0	0	Clinical Hypothyroidism	Multinodular goiter	Bilateral	Papillary	Multi-Focal	Intermediate	T2	N1b	M0	I	Excellent	0

182 rows × 17 columns

Next steps: Generate code with `test_set` View recommended plots New interactive sheet

```
import pandas as pd
from sklearn.impute import SimpleImputer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score


rdsample = RandomOverSampler()
target = df['Recurred']
x_sampled, y_sampled = rdsample.fit_resample(new_df, target)  # Use new_df here
train_set, test_set, train_label, test_label = train_test_split(x_sampled, y_sampled, test_size=0.33, random_state=42)

# Get numerical columns from train_set (after one-hot encoding)
numerical_cols = train_set.select_dtypes(include=np.number).columns

# Impute missing values in numerical columns of both train_set and test_set
imputer = SimpleImputer(strategy='mean')  # Or other strategies like 'median', 'most_frequent'
train_set[numerical_cols] = imputer.fit_transform(train_set[numerical_cols])
test_set[numerical_cols] = imputer.transform(test_set[numerical_cols])

# Now fit the KNN model using the imputed data
KNN_1 = KNeighborsClassifier(n_neighbors=2)
KNN_1.fit(train_set, train_label)

# And predict using the imputed test data
predicted_values_KNN_1 = KNN_1.predict(test_set)
print(predicted_values_KNN_1)
accuracy_KNN_1 = accuracy_score(test_label, predicted_values_KNN_1)
print(accuracy_KNN_1)
```

```
[0 0 1 0 1 1 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 1 1 0 1 1 0 0 0 0
 0 1 1 0 0 1 0 0 1 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 0 0 0 0 0 1 1 1 0 1 1 0
 0 1 0 0 1 0 0 1 1 0 1 0 1 1 1 1 1 0 1 1 0 0 0 0 1 0 0 1 0 0 0 1 0 1 0 1 1
 0 0 0 0 0 1 0 0 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 0 0 1 1 0 0 0 1 1 0 1 1 1 1
 0 1 1 1 1 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 1 1 1 0 0 0 1 0 0 0 1 0 1 0]
0.9560439560439561
```

```
predict_knn = pd.DataFrame(predicted_values_KNN)
```

```
predict_knn.value_counts()
```

	count
0	
1	93
0	89

dtype: int64

Start coding or generate with AI.