```
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
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
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.4.0->feature_engine)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.4.0->feature_er
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.11/dist-packages (from statsmodels>=0.11.1->feature_engine) (1
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.11/dist-packages (from statsmodels>=0.11.1->feature_engine
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas>=2.2.0->feat
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
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
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (1
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot) (24
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
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib>=1.4.0->scikit-plot
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=0.18->scikit-plot
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib>=1.4.0->sc
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
0	27	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I
1	34	F	No	Yes	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I
2	30	F	No	No	No	Euthyroid	Single nodular	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I

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 ver
  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 ve
  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	Function_Subc Hypothy
0	27	0	0	0	0	0	False	True	False	
1	34	0	0	1	0	0	False	True	False	
2	30	0	0	0	0	0	False	True	False	
3	62	0	0	0	0	0	False	True	False	
4	62	0	0	0	0	0	False	True	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	Function_Subc Hypothy
0	27	0	0	0	0	0	False	True	False	
1	34	0	0	1	0	0	False	True	False	
2	30	0	0	0	0	0	False	True	False	
3	62	0	0	0	0	0	False	True	False	

4 rows × 41 columns

```
new_df.corr()
```

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred	Thyroid Function_Clinical Hypothyroidism	Thyroid Function_Euthyroid	Func
Age	1.000000	0.186457	0.309536	0.134531	0.176588	0.258897	-0.023205	-0.028367	
Gender	0.186457	1.000000	0.621886	0.175755	0.235865	0.328189	-0.047227	-0.050344	
Smoking	0.309536	0.621886	1.000000	0.252773	0.297874	0.333243	-0.024016	-0.010933	
Hx Smoking	0.134531	0.175755	0.252773	1.000000	0.261198	0.136073	0.007065	-0.126106	
Hx Radiothreapy	0.176588	0.235865	0.297874	0.261198	1.000000	0.174407	-0.024539	-0.061267	
Recurred	0.258897	0.328189	0.333243	0.136073	0.174407	1.000000	-0.046091	0.074827	
Thyroid Function_Clinical Hypothyroidism	-0.023205	-0.047227	-0.024016	0.007065	-0.024539	-0.046091	1.000000	-0.458868	
Thyroid Function_Euthyroid	-0.028367	-0.050344	-0.010933	-0.126106	-0.061267	0.074827	-0.458868	1.000000	
Thyroid Function_Subclinical Hyperthyroidism	-0.085732	0.004327	-0.044052	0.056064	-0.015693	-0.072075	-0.020684	-0.293443	
Thyroid Function_Subclinical Hypothyroidism	0.100209	0.086095	0.050354	0.105639	-0.026577	0.032535	-0.035031	-0.496975	
Physical Examination_Multinodular goiter	0.102101	0.084366	0.050136	0.057588	0.017860	0.150881	-0.012026	-0.021666	
Physical Examination_Normal	-0.071016	-0.065089	-0.052261	0.036560	-0.018617	0.001131	-0.024539	-0.118639	
Physical Examination_Single nodular goiter-left	0.020799	0.087516	0.140912	-0.083275	0.017232	0.012412	-0.063466	0.070076	
Physical Examination_Single nodular goiter-right	-0.094108	-0.124909	-0.144643	0.015933	-0.022615	-0.138297	0.081337	0.058124	
Adenopathy_Bilateral	0.131884	0.268738	0.223335	0.132686	-0.041198	0.376962	-0.000141	0.007251	
Adenopathy_Extensive	0.045049	0.135547	0.122806	0.036560	0.417933	0.217726	-0.024539	-0.061267	
Adenopathy_Left	-0.030813	0.027683	0.031315	-0.060527	-0.029406	0.203033	-0.038760	0.047155	
Adenopathy_Posterior	0.029399	-0.034562	-0.027751	-0.020348	-0.009886	0.115613	-0.013030	0.028397	
Adenopathy_Right	-0.008665	0.022360	0.067499	0.075459	0.007225	0.288558	-0.022811	0.055513	
Pathology_Hurthel cell	0.108446	0.069234	0.191187	0.114421	0.055590	0.009398	0.092521	-0.080724	
Pathology_Micropapillary	0.072205	-0.079106	-0.097766	0.014870	-0.051648	-0.237216	0.022456	-0.083755	
Pathology_Papillary	-0.164530	0.012346	-0.121169	-0.092151	-0.056014	0.121444	-0.034311	0.110237	
Focality_Uni-Focal	-0.223847	-0.207634	-0.238494	-0.001204	-0.102415	-0.383776	0.008177	0.014298	
Risk_Intermediate	0.062754	0.153387	0.052174	-0.010368	-0.082206	0.462566	-0.006639	-0.007263	
Risk_Low	-0.228129	-0.269910	-0.276274	-0.088406	-0.145126	-0.708266	0.037660	0.002524	
T_T1b	-0.138038	-0.105800	-0.111453	-0.004562	0.013219	-0.130964	-0.016485	-0.006674	
T_T2	-0.188722	-0.096133	-0.133058	-0.123951	-0.070191	-0.268105	0.038916	0.033135	
T_T3a	0.058107	0.080672	0.030987	-0.046710	-0.078913	0.186500	-0.034852	0.084819	
T_T3b	0.039829	0.068303	0.076302	0.141887	-0.028489	0.275178	0.037356	-0.110217	
T_T4a	0.242001	0.099435	0.261460	0.114421	0.143207	0.348473	-0.042215	-0.011635	
T_T4b	0.206634	0.259198	0.326673	0.169390	0.388970	0.233069	-0.026268	0.003508	
N_N1a	-0.051278	-0.031137	-0.060961	-0.026224	-0.033683	0.094672	0.020013	-0.134440	
N_N1b	0.075087	0.246946	0.220617	0.051487	0.104566	0.605927	-0.066894	0.078570	
M_M1	0.235401	0.211540	0.321233	0.127209	0.430214	0.354360	-0.039939	0.014411	
Stage_II	0.369106	0.147333	0.195086	-0.012303	0.029243	0.335022	-0.054303	0.007251	
Stage_III	0.208210	0.083175	0.191325	0.267138	-0.014017	0.163932	-0.018476	-0.110926	
Stage_IVA	0.141867	0.110044	0.231977	0.088823	0.208984	0.141783	-0.015980	-0.139526	
Stage_IVB	0.336617	0.159335	0.261746	0.191920	0.443356	0.274397	-0.030926	0.021384	
Response_Excellent	-0.258453	-0.263805	-0.276350	-0.084694	-0.109624	-0.671568	0.044619	-0.050955	
Response_Indeterminate	0.055762	-0.005657	-0.038540	-0.067416	-0.059387	-0.161760	0.003636	-0.039426	
Response_Structural									

Response_Structural Incomplete	0.198518	0.302000	0.318792	0.102449	0.152818	0.863540	-0.065186	0.074347

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()
```

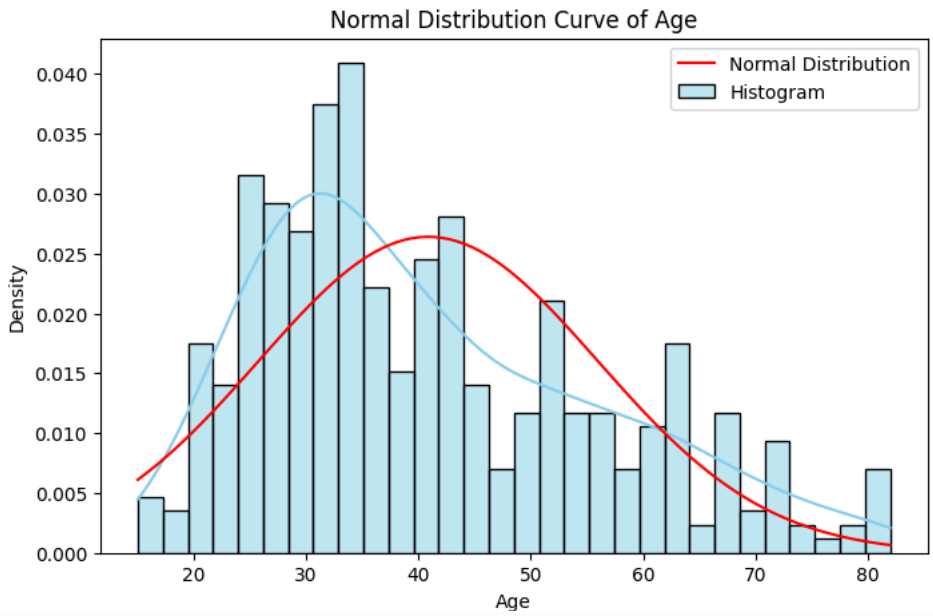
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 383 entries, 0 to 382
Data columns (total 41 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Age                                             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()
```



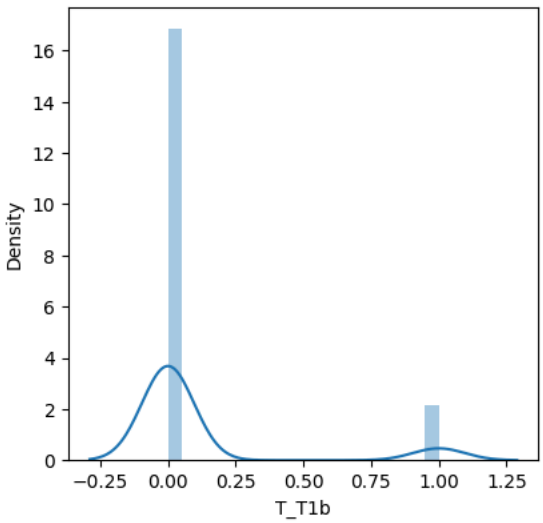

```
new_df['T_T1b'].dtypes
```

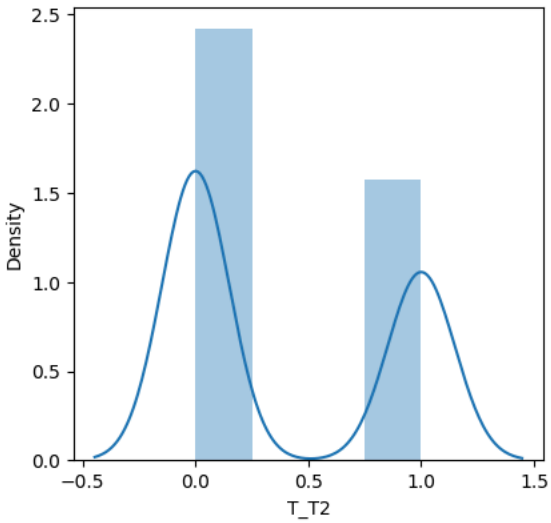
```
dtype('bool')
```

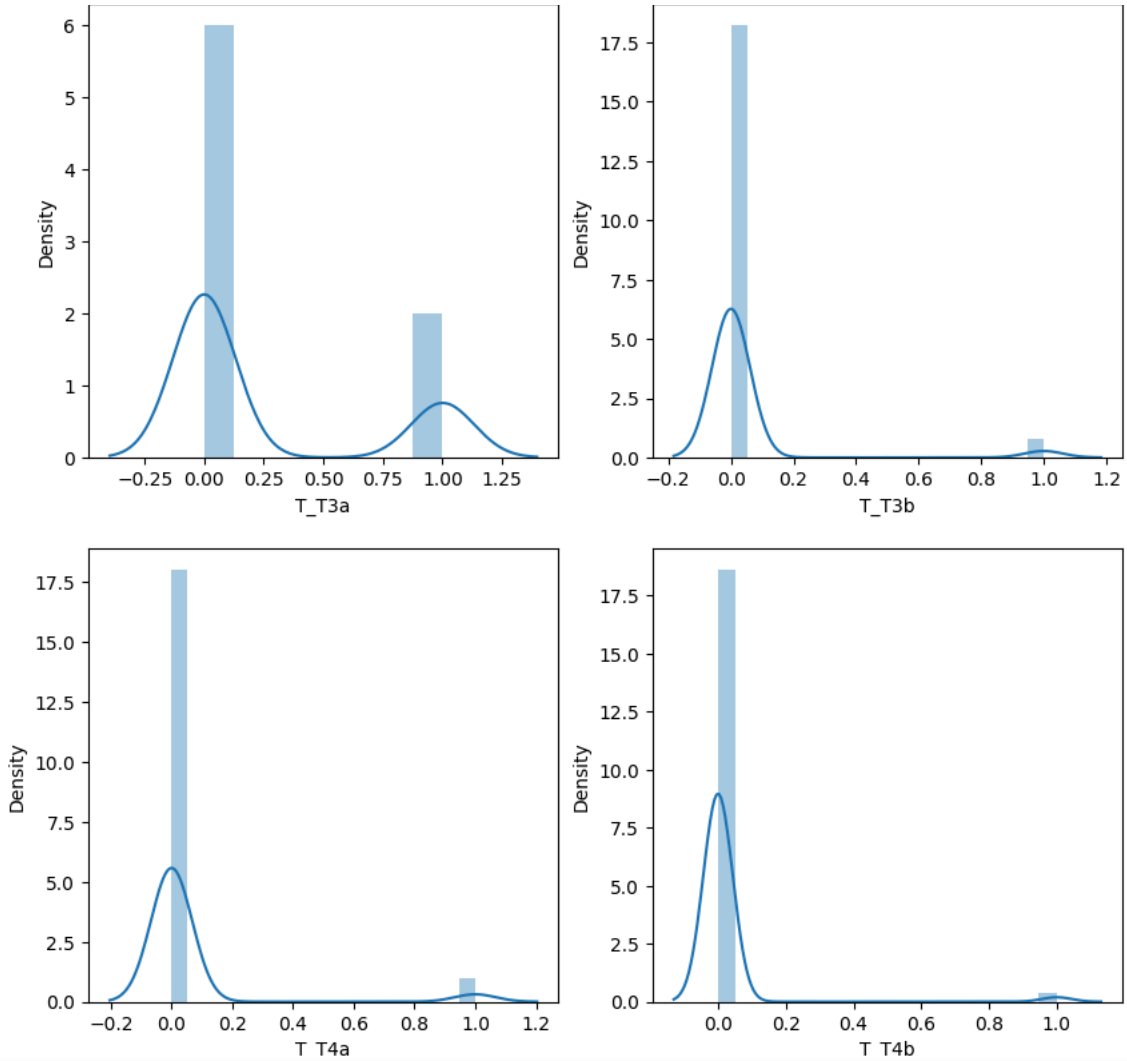
```
new_df['T_T1b'] = new_df['T_T1b'].astype(float)
new_df['T_T2'] = new_df['T_T2'].astype(float)
new_df['T_T3a'] = new_df['T_T3a'].astype(float)
new_df['T_T3b'] = new_df['T_T3b'].astype(float)
new_df['T_T4a'] = new_df['T_T4a'].astype(float)
new_df['T_T4b'] = new_df['T_T4b'].astype(float)
```

```
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])
```





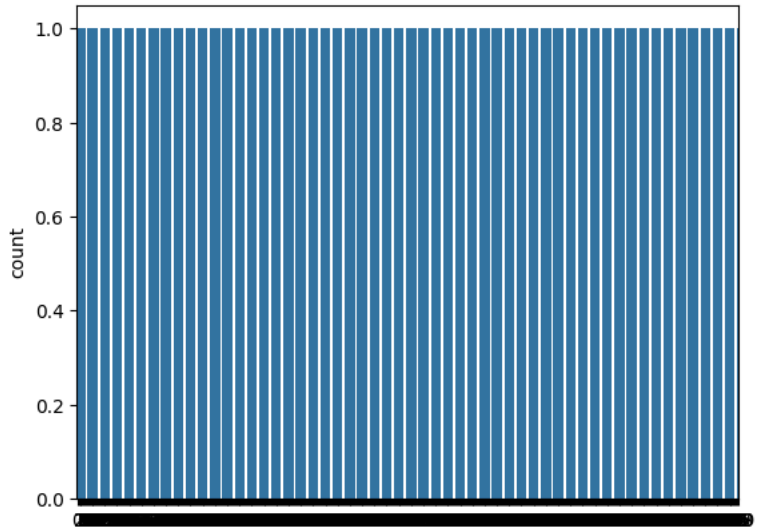

Density
T_T3a
Density
T_T3b
Density
T_T4a
Density
T_T4b

```
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

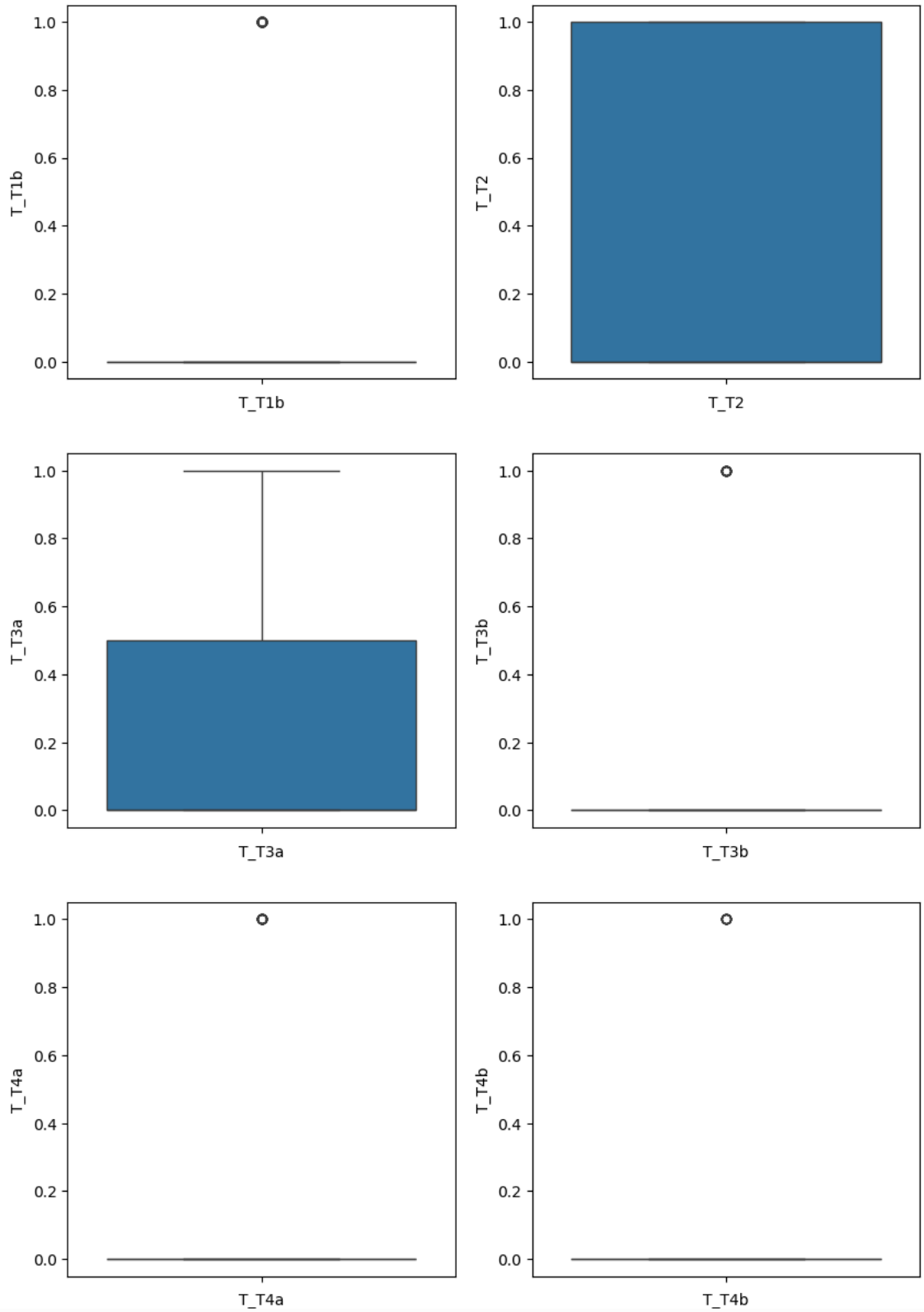
```
x_sampled=new_df
```

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

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Recurred	Thyroid Function_Clinical Hypothyroidism	Thyroid Function_Euthyroid	Thyroid Function_Subclinical Hyperthyroidism	Function_Subc Hypothy
0	27	0	0	0	0	0	False	True	False	
1	34	0	0	1	0	0	False	True	False	
2	30	0	0	0	0	0	False	True	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	Func
Age	1.000000	0.186457	0.309536	0.134531	0.176588	0.258897	-0.023205	-0.028367	
Gender	0.186457	1.000000	0.621886	0.175755	0.235865	0.328189	-0.047227	-0.050344	
Smoking	0.309536	0.621886	1.000000	0.252773	0.297874	0.333243	-0.024016	-0.010933	
Hx Smoking	0.134531	0.175755	0.252773	1.000000	0.261198	0.136073	0.007065	-0.126106	
Hx Radiothreapy	0.176588	0.235865	0.297874	0.261198	1.000000	0.174407	-0.024539	-0.061267	
Recurred	0.258897	0.328189	0.333243	0.136073	0.174407	1.000000	-0.046091	0.074827	
Thyroid Function_Clinical Hypothyroidism	-0.023205	-0.047227	-0.024016	0.007065	-0.024539	-0.046091	1.000000	-0.458868	
Thyroid Function_Euthyroid	-0.028367	-0.050344	-0.010933	-0.126106	-0.061267	0.074827	-0.458868	1.000000	
Thyroid Function_Subclinical Hyperthyroidism	-0.085732	0.004327	-0.044052	0.056064	-0.015693	-0.072075	-0.020684	-0.293443	
Thyroid Function_Subclinical Hypothyroidism	0.100209	0.086095	0.050354	0.105639	-0.026577	0.032535	-0.035031	-0.496975	
Physical Examination_Multinodular goiter	0.102101	0.084366	0.050136	0.057588	0.017860	0.150881	-0.012026	-0.021666	
Physical Examination_Normal	-0.071016	-0.065089	-0.052261	0.036560	-0.018617	0.001131	-0.024539	-0.118639	
Physical Examination_Single nodular goiter-left	0.020799	0.087516	0.140912	-0.083275	0.017232	0.012412	-0.063466	0.070076	
Physical Examination_Single nodular goiter-right	-0.094108	-0.124909	-0.144643	0.015933	-0.022615	-0.138297	0.081337	0.058124	
Adenopathy_Bilateral	0.131884	0.268738	0.223335	0.132686	-0.041198	0.376962	-0.000141	0.007251	
Adenopathy_Extensive	0.045049	0.135547	0.122806	0.036560	0.417933	0.217726	-0.024539	-0.061267	
Adenopathy_Left	-0.030813	0.027683	0.031315	-0.060527	-0.029406	0.203033	-0.038760	0.047155	
Adenopathy_Posterior	0.029399	-0.034562	-0.027751	-0.020348	-0.009886	0.115613	-0.013030	0.028397	
Adenopathy_Right	-0.008665	0.022360	0.067499	0.075459	0.007225	0.288558	-0.022811	0.055513	
Pathology_Hurthel cell	0.108446	0.069234	0.191187	0.114421	0.055590	0.009398	0.092521	-0.080724	
Pathology_Micropapillary	0.072205	-0.079106	-0.097766	0.014870	-0.051648	-0.237216	0.022456	-0.083755	
Pathology_Papillary	-0.164530	0.012346	-0.121169	-0.092151	-0.056014	0.121444	-0.034311	0.110237	
Focality_Uni-Focal	-0.223847	-0.207634	-0.238494	-0.001204	-0.102415	-0.383776	0.008177	0.014298	
Risk_Intermediate	0.062754	0.153387	0.052174	-0.010368	-0.082206	0.462566	-0.006639	-0.007263	
Risk_Low	-0.228129	-0.269910	-0.276274	-0.088406	-0.145126	-0.708266	0.037660	0.002524	
T_T1b	-0.138038	-0.105800	-0.111453	-0.004562	0.013219	-0.130964	-0.016485	-0.006674	
T_T2	-0.188722	-0.096133	-0.133058	-0.123951	-0.070191	-0.268105	0.038916	0.033135	
T_T3a	0.058107	0.080672	0.030987	-0.046710	-0.078913	0.186500	-0.034852	0.084819	
T_T3b	0.039829	0.068303	0.076302	0.141887	-0.028489	0.275178	0.037356	-0.110217	
T_T4a	0.242001	0.099435	0.261460	0.114421	0.143207	0.348473	-0.042215	-0.011635	
T_T4b	0.206634	0.259198	0.326673	0.169390	0.388970	0.233069	-0.026268	0.003508	
N_N1a	-0.051278	-0.031137	-0.060961	-0.026224	-0.033683	0.094672	0.020013	-0.134440	
N_N1b	0.075087	0.246946	0.220617	0.051487	0.104566	0.605927	-0.066894	0.078570	
M_M1	0.235401	0.211540	0.321233	0.127209	0.430214	0.354360	-0.039939	0.014411	
Stage_II	0.369106	0.147333	0.195086	-0.012303	0.029243	0.335022	-0.054303	0.007251	
Stage_III	0.208210	0.083175	0.191325	0.267138	-0.014017	0.163932	-0.018476	-0.110926	
Stage_IVA	0.141867	0.110044	0.231977	0.088823	0.208984	0.141783	-0.015980	-0.139526	
Stage_IVB	0.336617	0.159335	0.261746	0.191920	0.443356	0.274397	-0.030926	0.021384	
Response_Excellent	-0.258453	-0.263805	-0.276350	-0.084694	-0.109624	-0.671568	0.044619	-0.050955	
Response_Indeterminate	0.055762	-0.005657	-0.038540	-0.067416	-0.059387	-0.161760	0.003636	-0.039426	
Response_Structural									

Response_Structural Incomplete	0.198518	0.302000	0.318792	0.102449	0.152818	0.863540	-0.065186	0.074347

41 rows × 41 columns

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