

```
In [1]: import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import metrics
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.tree import DecisionTreeClassifier
```

```
In [2]: diab=pd.read_csv(r"/Users/apple/Downloads/diabetes.csv")
```

```
In [3]: diab
```

```
Out[3]:
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outco
0	6	148	72	35	0	33.6	0.627	50	
1	1	85	66	29	0	26.6	0.351	31	
2	8	183	64	0	0	23.3	0.672	32	
3	1	89	66	23	94	28.1	0.167	21	
4	0	137	40	35	168	43.1	2.288	33	
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	
764	2	122	70	27	0	36.8	0.340	27	
765	5	121	72	23	112	26.2	0.245	30	
766	1	126	60	0	0	30.1	0.349	47	
767	1	93	70	31	0	30.4	0.315	23	

768 rows × 9 columns

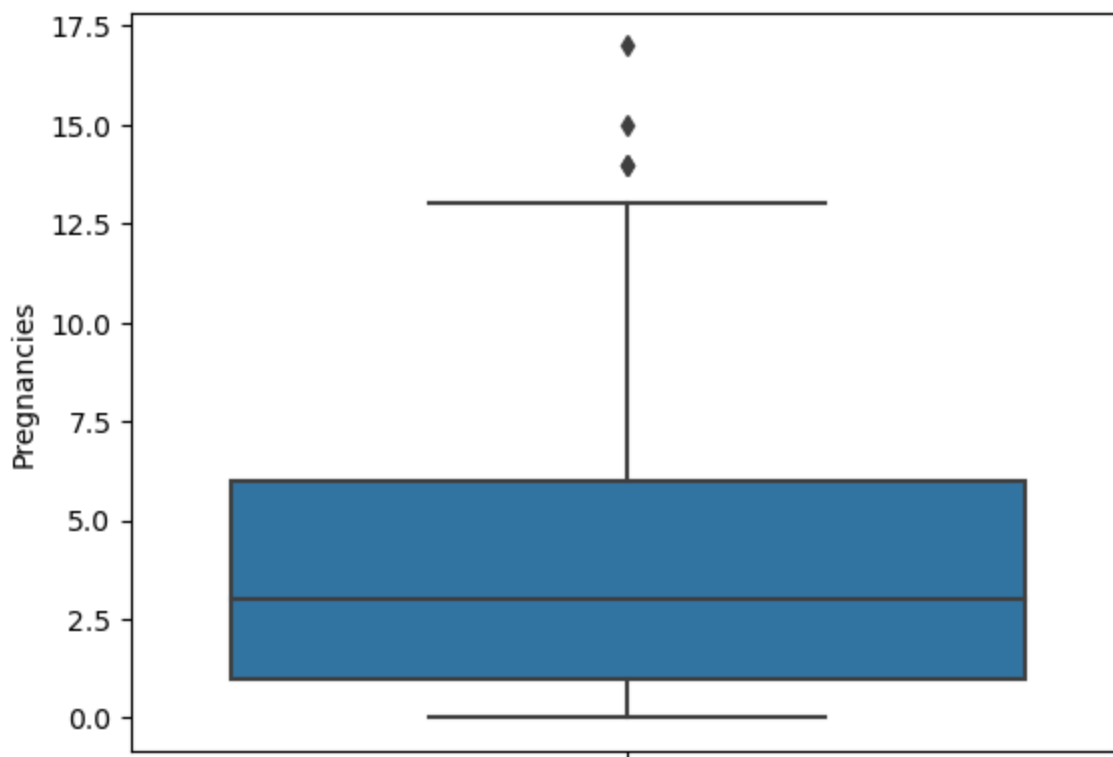
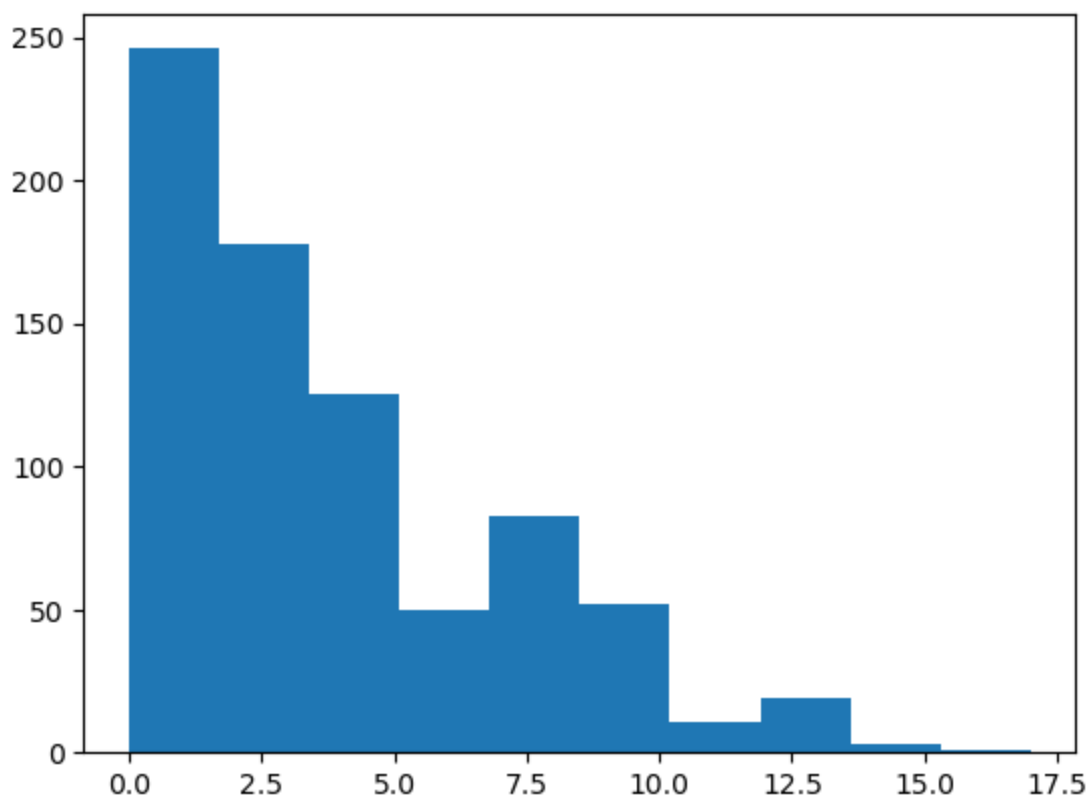
## EDA

```
In [4]: def univariate_num(data,x):
missing=data[x].isnull().sum()
min1=round(data[x].min(), 2)
max1=round(data[x].max(), 2)
mean=round(data[x].mean(), 2)
var=round(data[x].var(), 2)
std=round(data[x].std(),2)
range1=round(max1-min1, 2)
q1=round(data[x].quantile(.25), 2)
q2=round(data[x].quantile(.5),2)
q3=round(data[x].quantile(.75), 2)
skew=round(data[x].skew(), 2)
kurt=round(data[x].kurt(), 2)
myvalue={"missing":missing, "min":min1, "max":max1, "mean":mean,
          "var":var, "std":std, "range":range1, "q1":q1, "q2":q2, "q3":q3,
          "skewness":skew, "kurtosis":kurt}
plt.hist(data[x])
plt.show()
sns.boxplot(data=data, y=data[x])
plt.show()
return myvalue
```

```
In [5]: diab.columns
```

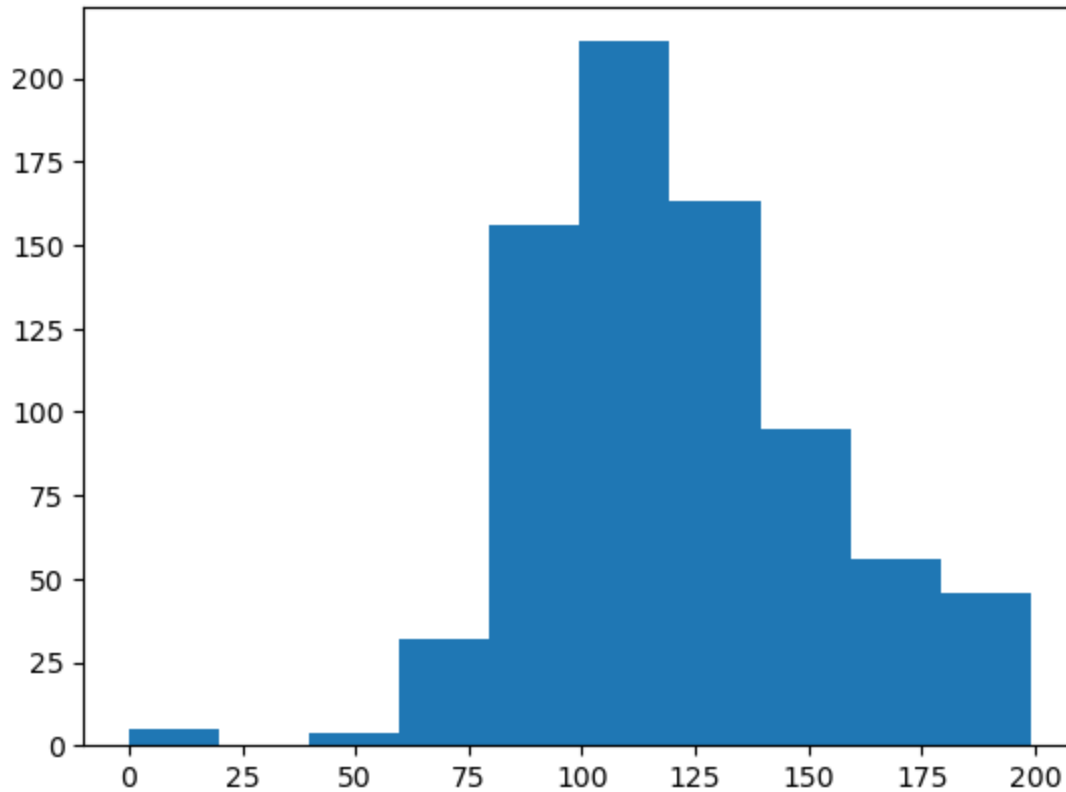
```
Out[5]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
            'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
            dtype='object')
```

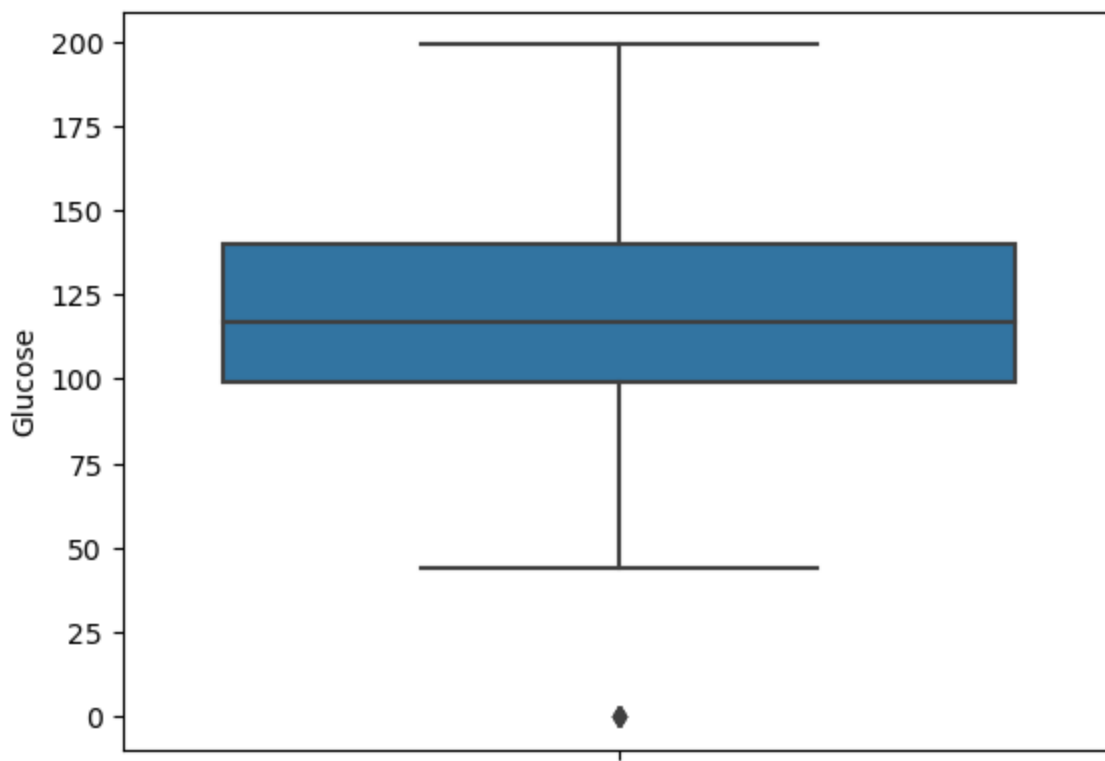
```
In [6]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
        #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
        univariate_num(diab, 'Pregnancies') # has outliers but very
```



```
Out[6]: {'missing': 0,  
        'min': 0,  
        'max': 17,  
        'mean': 3.85,  
        'var': 11.35,  
        'std': 3.37,  
        'range': 17,  
        'q1': 1.0,  
        'q2': 3.0,  
        'q3': 6.0,  
        'skewness': 0.9,  
        'kurtosis': 0.16}
```

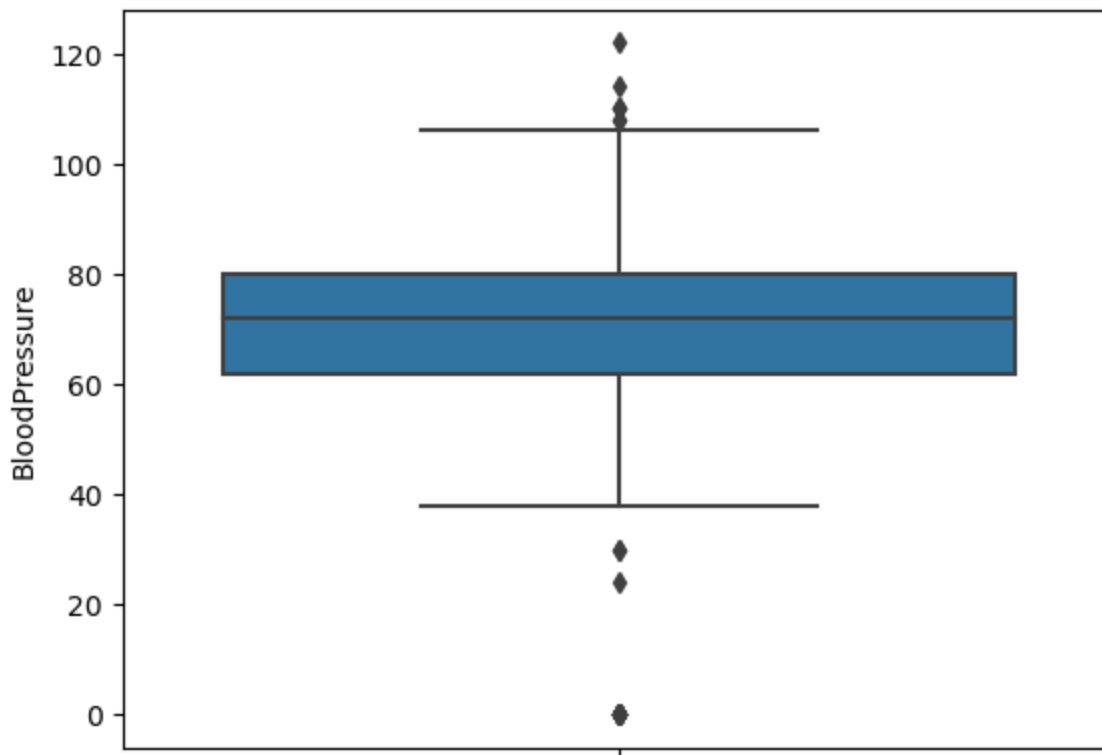
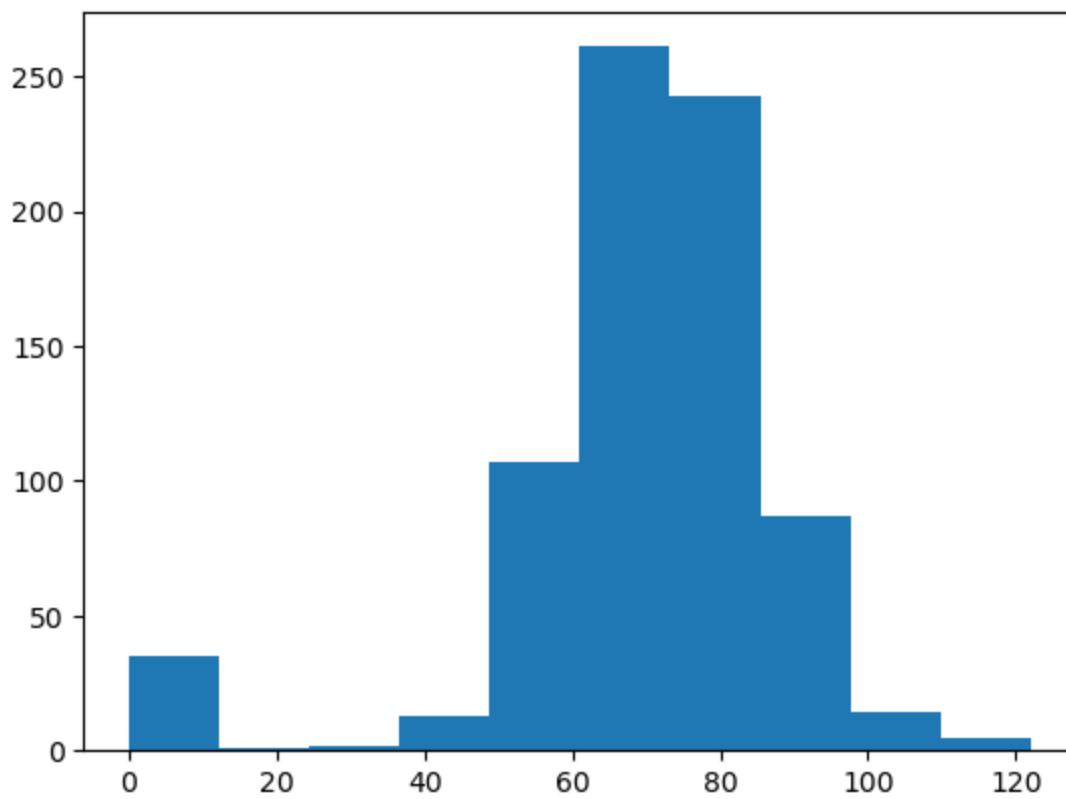
```
In [7]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
        #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
univariate_num(diab, 'Glucose') # no outliers
```





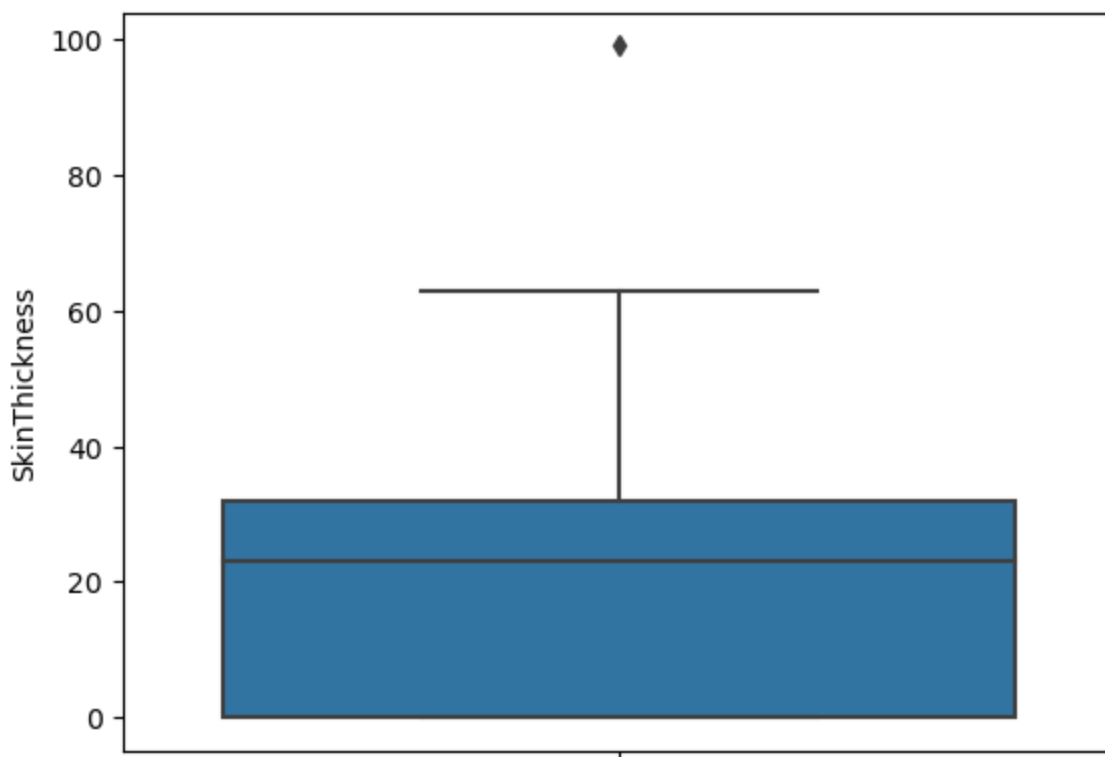
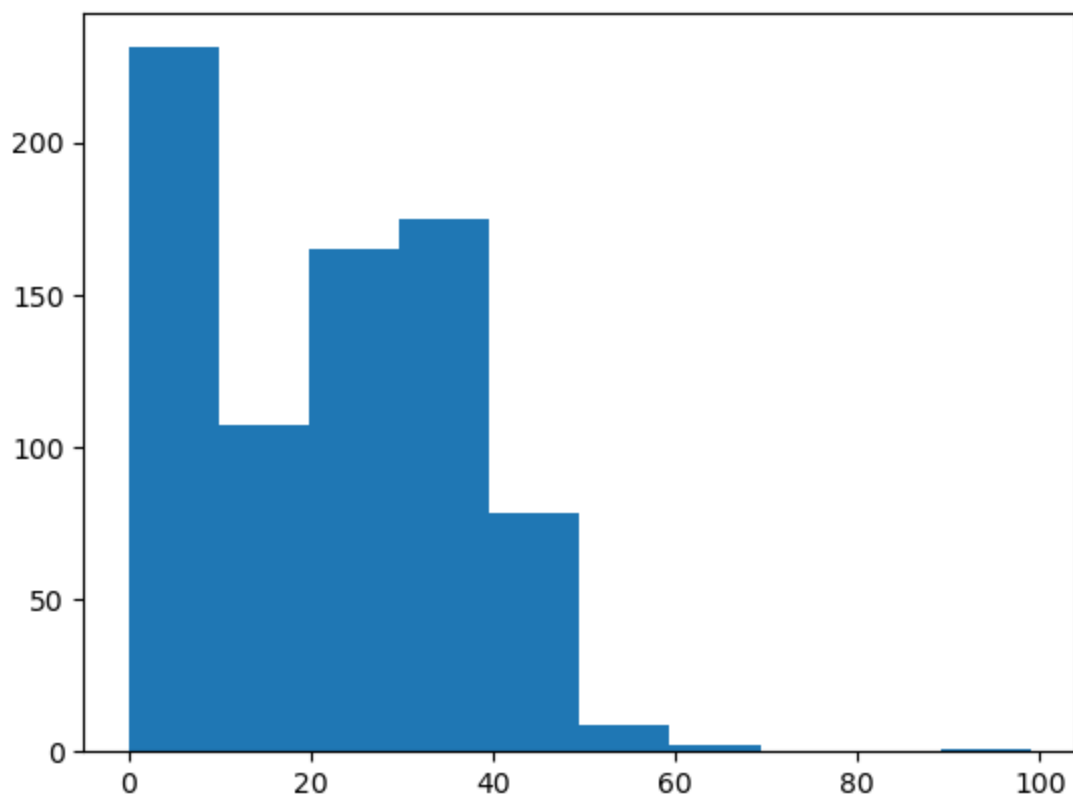
```
Out[7]: {'missing': 0,  
        'min': 0,  
        'max': 199,  
        'mean': 120.89,  
        'var': 1022.25,  
        'std': 31.97,  
        'range': 199,  
        'q1': 99.0,  
        'q2': 117.0,  
        'q3': 140.25,  
        'skewness': 0.17,  
        'kurtosis': 0.64}
```

```
In [8]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
        #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
univariate_num(diab, "BloodPressure") # less outliers
```



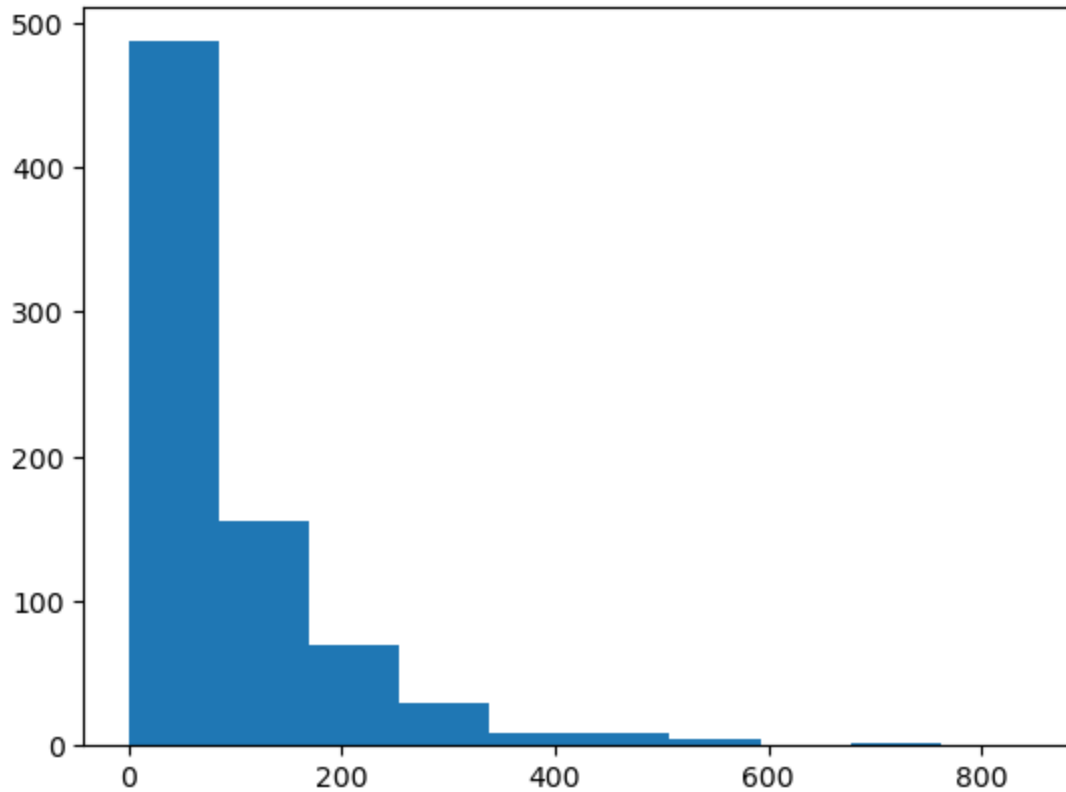
```
Out[8]: {'missing': 0,  
        'min': 0,  
        'max': 122,  
        'mean': 69.11,  
        'var': 374.65,  
        'std': 19.36,  
        'range': 122,  
        'q1': 62.0,  
        'q2': 72.0,  
        'q3': 80.0,  
        'skewness': -1.84,  
        'kurtosis': 5.18}
```

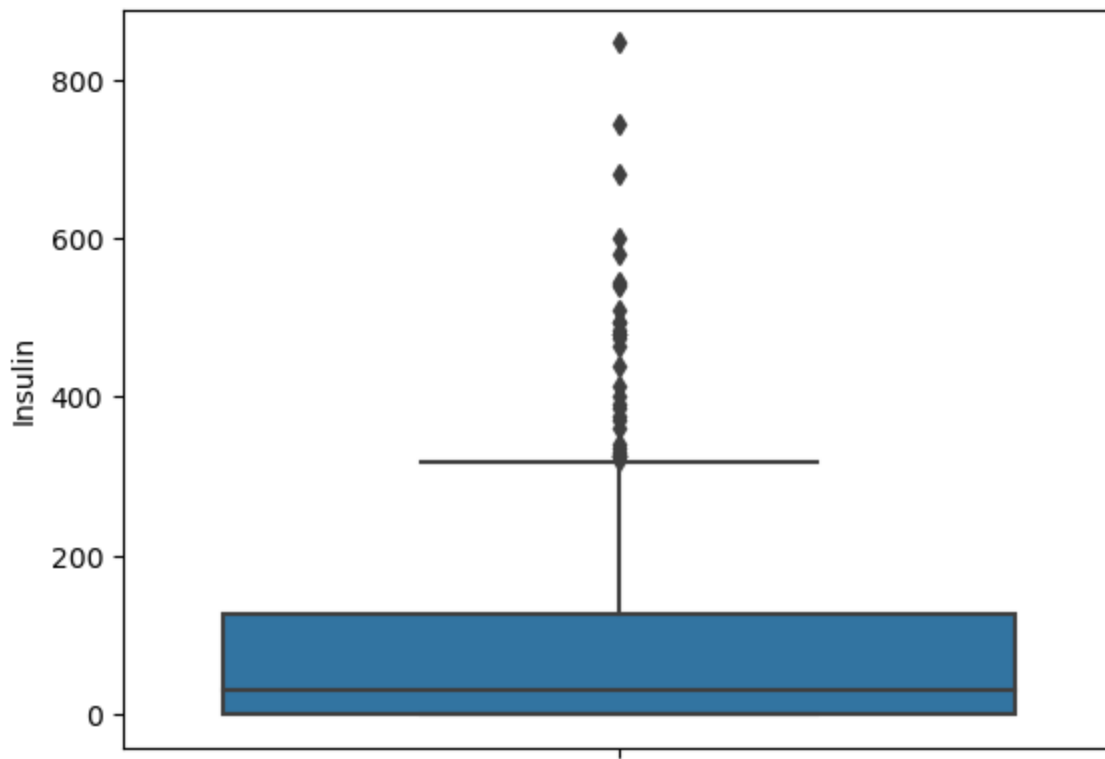
```
In [10]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
         #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
univariate_num(diab, "SkinThickness") # no outliers
```



```
Out[10]: {'missing': 0,  
          'min': 0,  
          'max': 99,  
          'mean': 20.54,  
          'var': 254.47,  
          'std': 15.95,  
          'range': 99,  
          'q1': 0.0,  
          'q2': 23.0,  
          'q3': 32.0,  
          'skewness': 0.11,  
          'kurtosis': -0.52}
```

```
In [11]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
          #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
univariate_num(diab, "Insulin") # has outliers
```

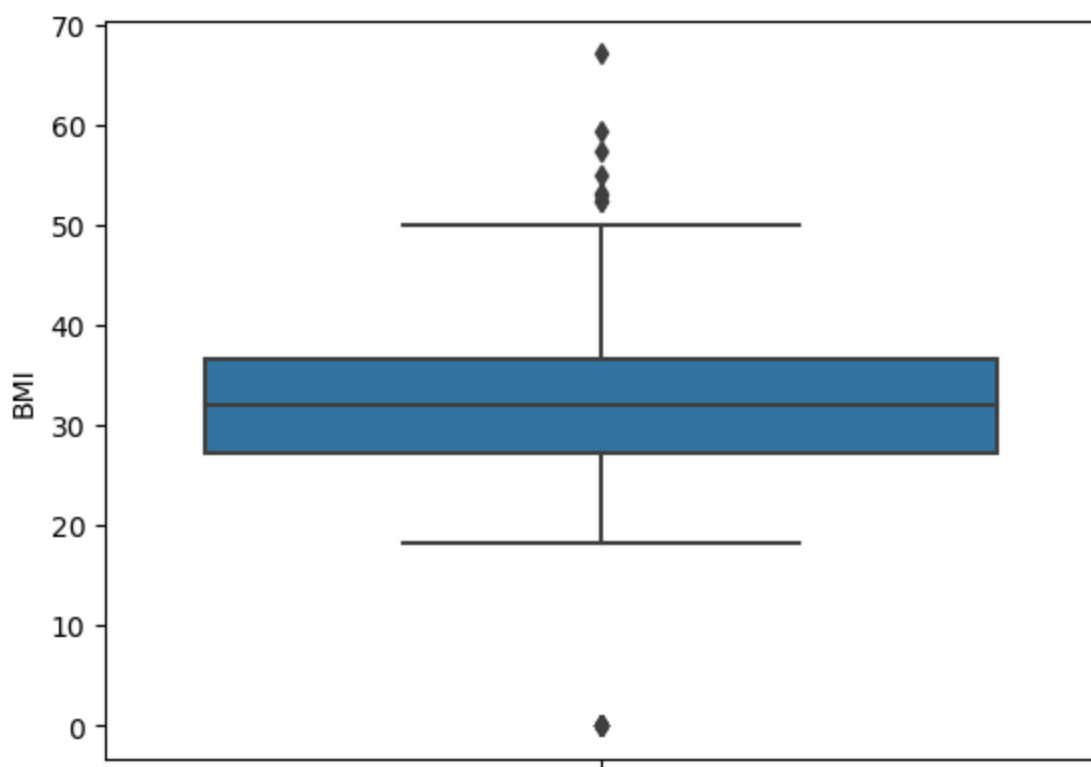
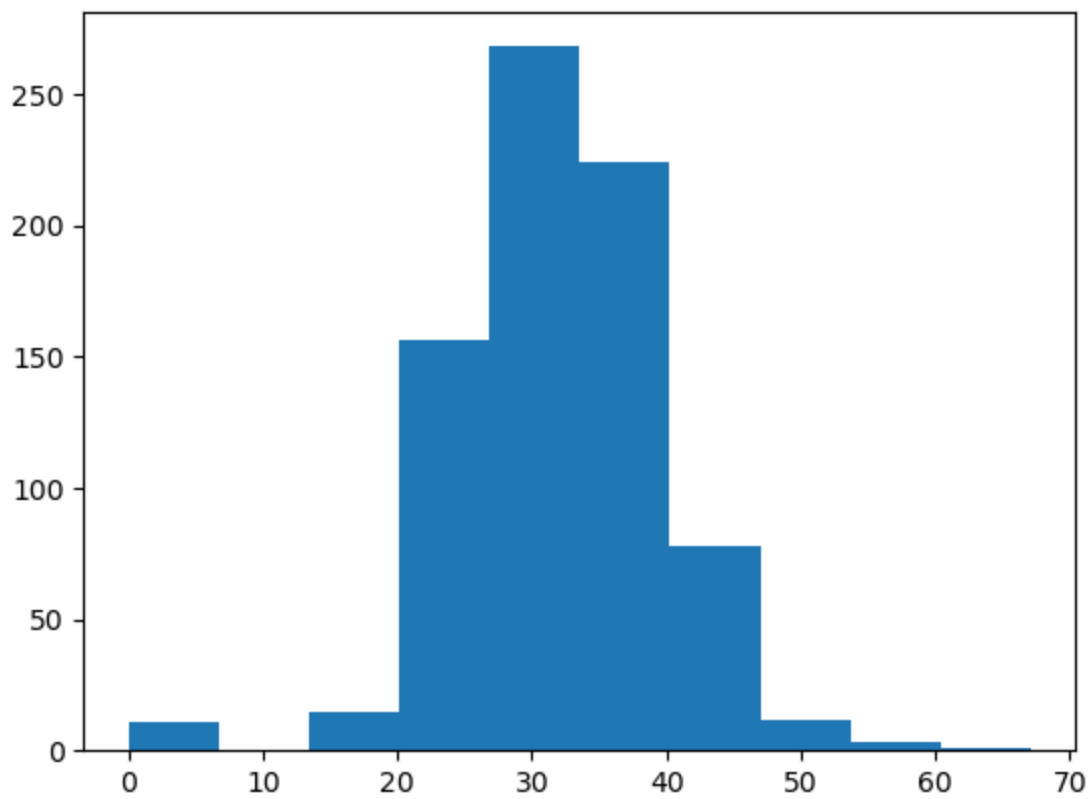




```
Out[11]: {'missing': 0,
          'min': 0,
          'max': 846,
          'mean': 79.8,
          'var': 13281.18,
          'std': 115.24,
          'range': 846,
          'q1': 0.0,
          'q2': 30.5,
          'q3': 127.25,
          'skewness': 2.27,
          'kurtosis': 7.21}
```

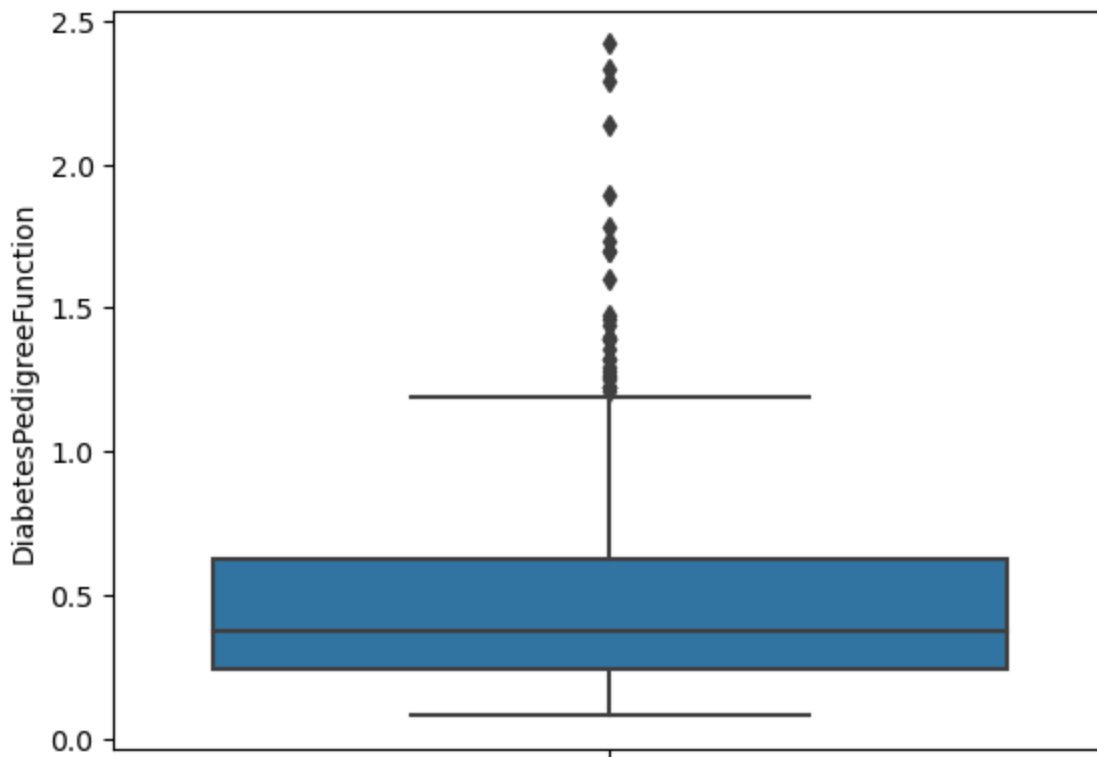
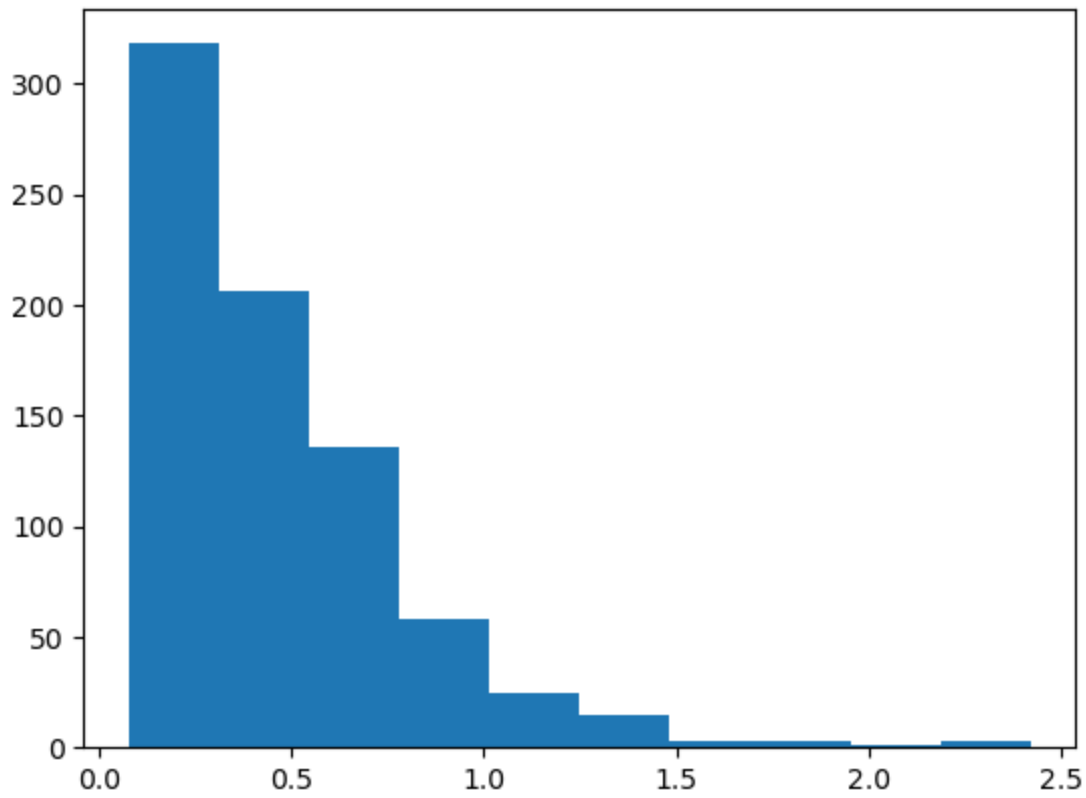
```
In [12]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
          #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
          univariate_num(diab, "BMI") # has outliers
```





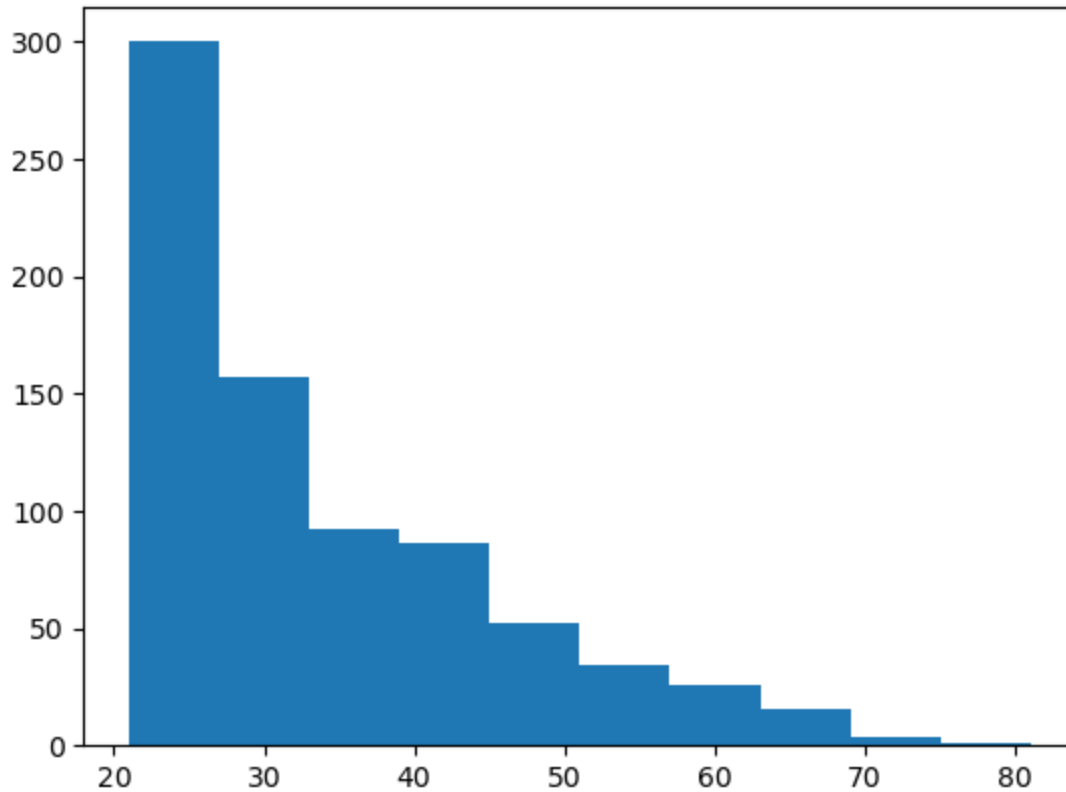
```
Out[12]: {'missing': 0,  
          'min': 0.0,  
          'max': 67.1,  
          'mean': 31.99,  
          'var': 62.16,  
          'std': 7.88,  
          'range': 67.1,  
          'q1': 27.3,  
          'q2': 32.0,  
          'q3': 36.6,  
          'skewness': -0.43,  
          'kurtosis': 3.29}
```

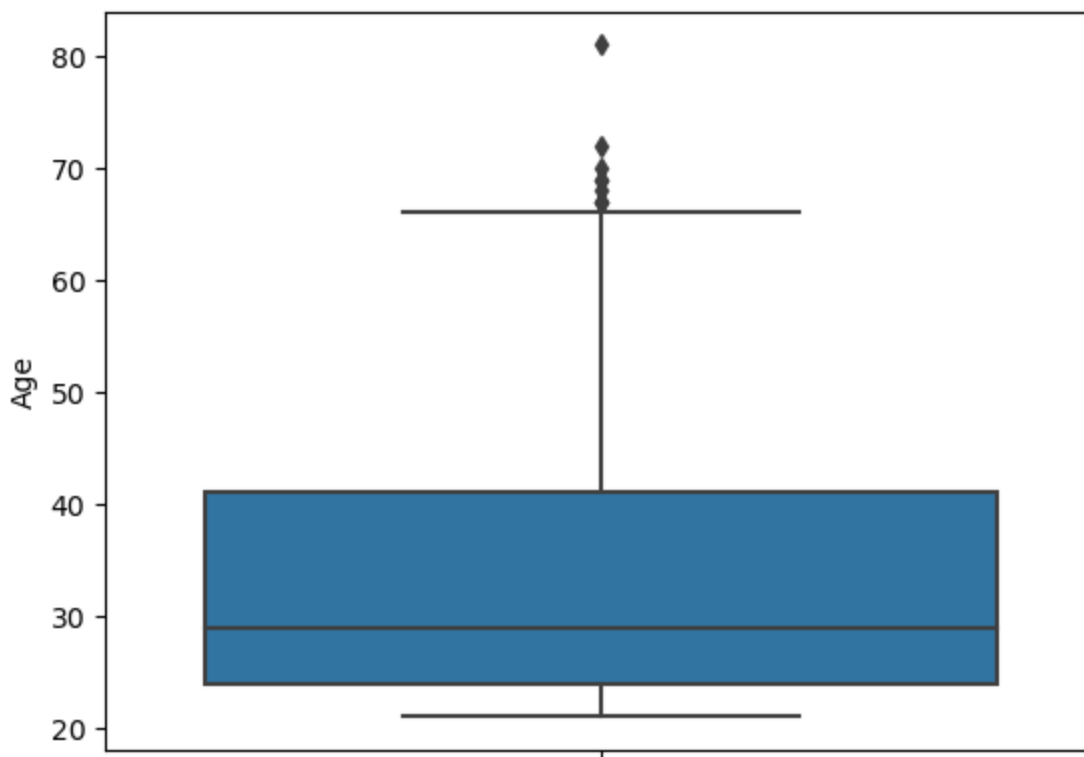
```
In [13]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
        #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
univariate_num(diab, "DiabetesPedigreeFunction") #has outliers
```



```
Out[13]: {'missing': 0,  
          'min': 0.08,  
          'max': 2.42,  
          'mean': 0.47,  
          'var': 0.11,  
          'std': 0.33,  
          'range': 2.34,  
          'q1': 0.24,  
          'q2': 0.37,  
          'q3': 0.63,  
          'skewness': 1.92,  
          'kurtosis': 5.59}
```

```
In [17]: #['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
          #'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
univariate_num(diab, "Age") # has outliers
```





```
Out[17]: {'missing': 0,
          'min': 21,
          'max': 81,
          'mean': 33.24,
          'var': 138.3,
          'std': 11.76,
          'range': 60,
          'q1': 24.0,
          'q2': 29.0,
          'q3': 41.0,
          'skewness': 1.13,
          'kurtosis': 0.64}
```

## missing value treatment

```
In [19]: diab.isnull().sum() # no missing value
```

```
Out[19]: Pregnancies      0
          Glucose          0
          BloodPressure    0
          SkinThickness     0
          Insulin           0
          BMI               0
          DiabetesPedigreeFunction  0
          Age              0
          Outcome           0
          dtype: int64
```

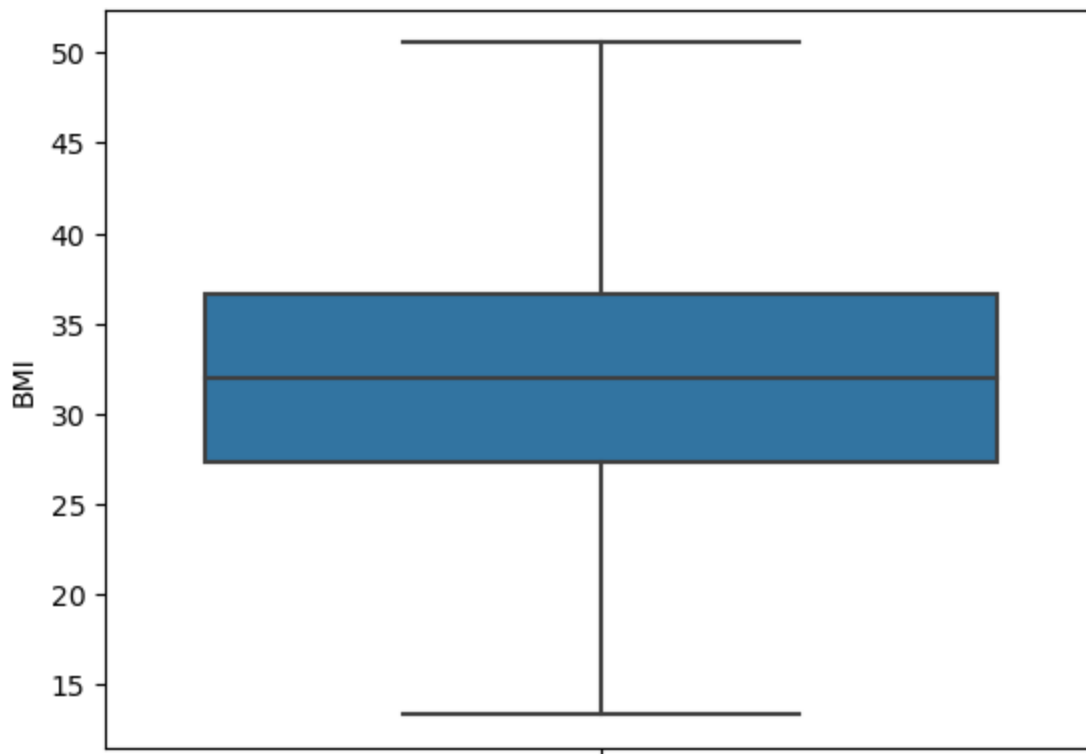
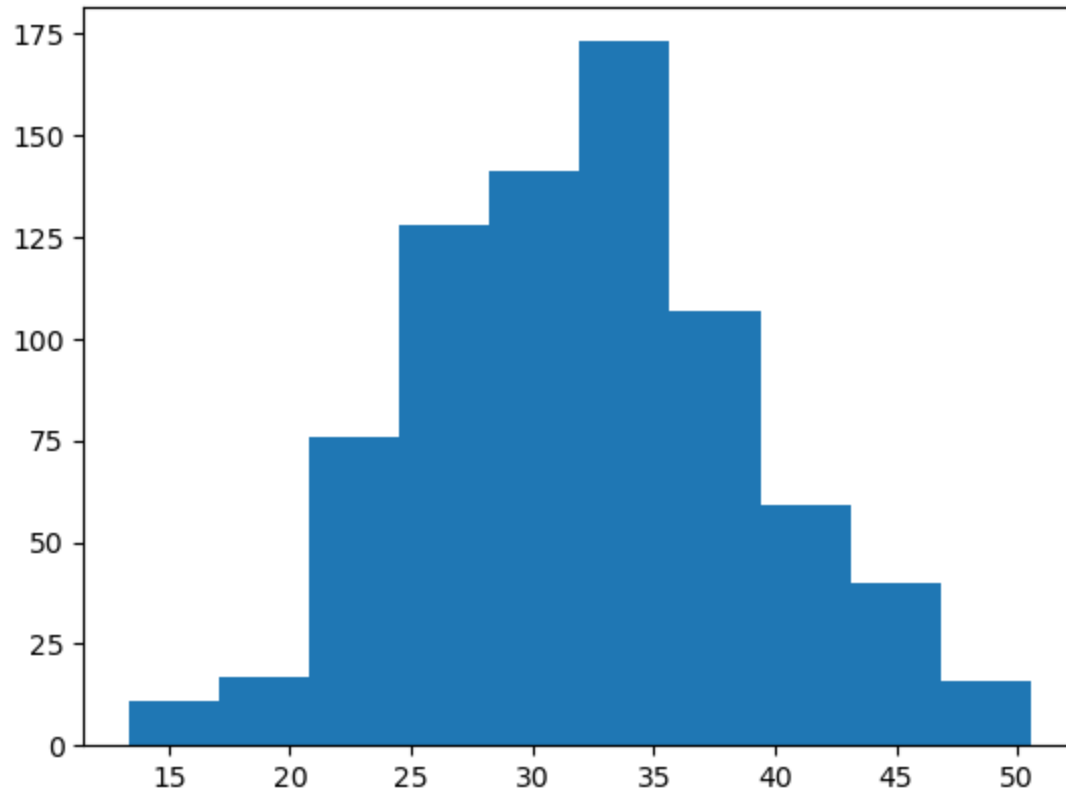
## outliers treatment

```
In [20]: var="BMI"

          q1=diab[var].quantile(.25)
          q3=diab[var].quantile(.75)
          lower_cap=q1-1.5*(q3-q1)
```

```
diab[var]=np.where(diab[var]>=upper_cap, upper_cap, diab[var] )  
diab[var]=np.where(diab[var]<=lower_cap, lower_cap, diab[var])
```

```
In [22]: univariate_num(diab, "BMI")
```



```
Out[22]: {'missing': 0,  
         'min': 13.35,  
         'max': 50.55,  
         'mean': 32.13,  
         'var': 49.7,  
         'std': 7.05,  
         'range': 37.2,  
         'q1': 27.3,  
         'q2': 32.0,  
         'q3': 36.6,  
         'skewness': 0.14,  
         'kurtosis': 0.05}
```

```
In [23]: diab.columns
```

```
Out[23]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',  
              'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],  
             dtype='object')
```

```
In [25]: var="Pregnancies"
```

```
q1=diab[var].quantile(.25)  
q3=diab[var].quantile(.75)  
lower_cap=q1-1.5*(q3-q1)  
upper_cap=q3+1.5*(q3-q1)  
  
diab[var]=np.where(diab[var]>=upper_cap, upper_cap,diab[var] )  
diab[var]=np.where(diab[var]<=lower_cap, lower_cap, diab[var])
```

```
In [26]: var="BloodPressure"
```

```
q1=diab[var].quantile(.25)  
q3=diab[var].quantile(.75)  
lower_cap=q1-1.5*(q3-q1)  
upper_cap=q3+1.5*(q3-q1)  
  
diab[var]=np.where(diab[var]>=upper_cap, upper_cap,diab[var] )  
diab[var]=np.where(diab[var]<=lower_cap, lower_cap, diab[var])
```

```
In [27]: var="Insulin"
```

```
q1=diab[var].quantile(.25)  
q3=diab[var].quantile(.75)  
lower_cap=q1-1.5*(q3-q1)  
upper_cap=q3+1.5*(q3-q1)  
  
diab[var]=np.where(diab[var]>=upper_cap, upper_cap,diab[var] )  
diab[var]=np.where(diab[var]<=lower_cap, lower_cap, diab[var])
```

```
In [28]: var="DiabetesPedigreeFunction"
```

```
q1=diab[var].quantile(.25)  
q3=diab[var].quantile(.75)  
lower_cap=q1-1.5*(q3-q1)  
upper_cap=q3+1.5*(q3-q1)  
  
diab[var]=np.where(diab[var]>=upper_cap, upper_cap,diab[var] )  
diab[var]=np.where(diab[var]<=lower_cap, lower_cap, diab[var])
```

```
In [30]: var="Age"
```

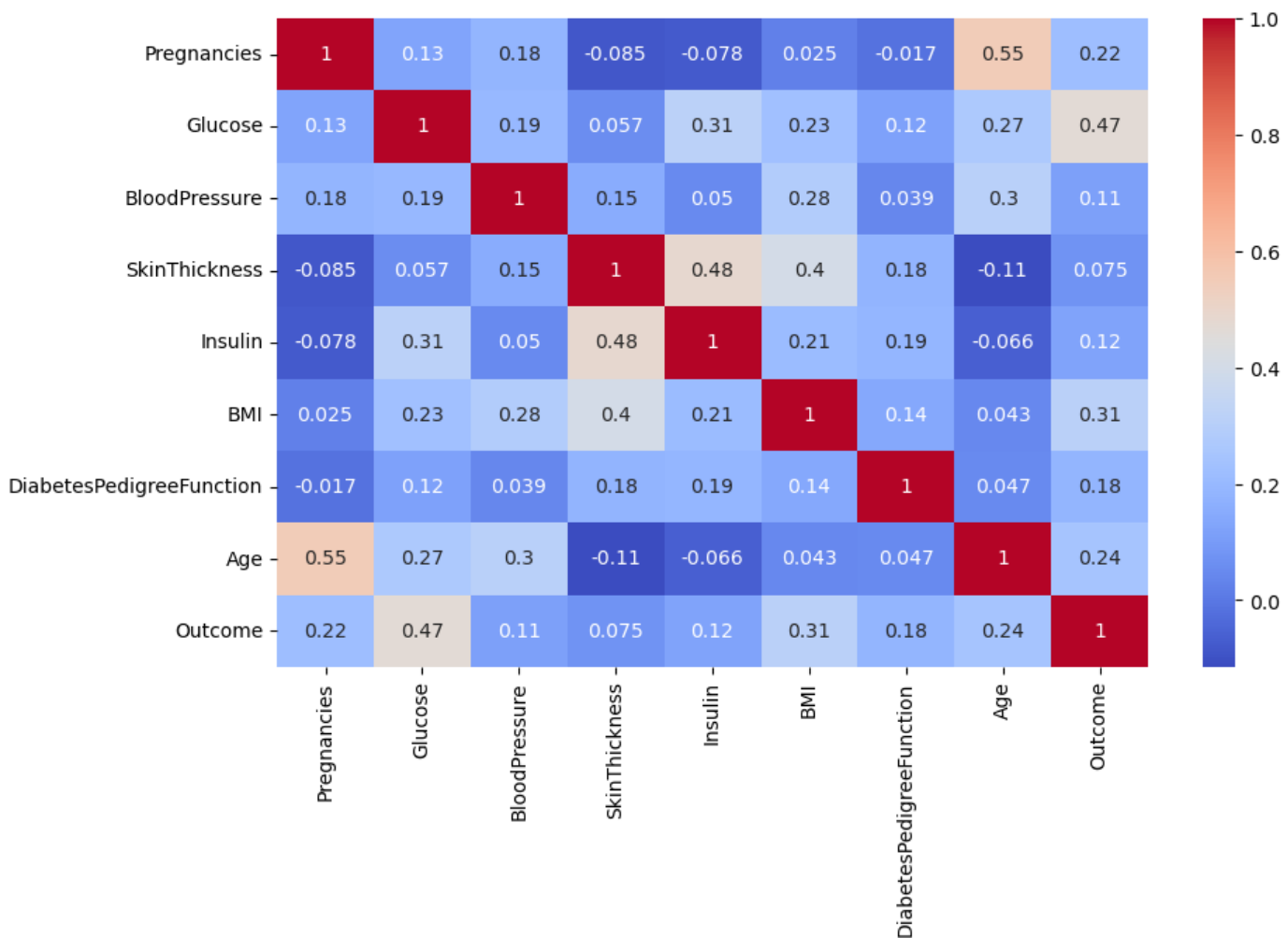
```
q1=diab[var].quantile(.25)  
q3=diab[var].quantile(.75)
```

```
lower_cap=q1-1.5*(q3-q1)
upper_cap=q3+1.5*(q3-q1)
```

```
diab[var]=np.where(diab[var]>=upper_cap, upper_cap,diab[var] )
diab[var]=np.where(diab[var]<=lower_cap, lower_cap, diab[var])
```

## multicollinearity

```
In [37]: cr=diab.corr()
plt.figure(figsize=(10,6))
sns.heatmap(cr,annot=True,cmap="coolwarm")
plt.show()
```



## model development

```
In [52]: diab1=pd.get_dummies(diab, drop_first=True)
```

```
In [55]: x=diab1.drop(columns=["Outcome"])
y=diab1["Outcome"]
```

```
In [57]: x_train, x_test, y_train, y_test=train_test_split(x,y, test_size=.3, random_state=0)
```

```
In [56]: from sklearn.tree import DecisionTreeClassifier
```

```
In [58]: dt=DecisionTreeClassifier()
```

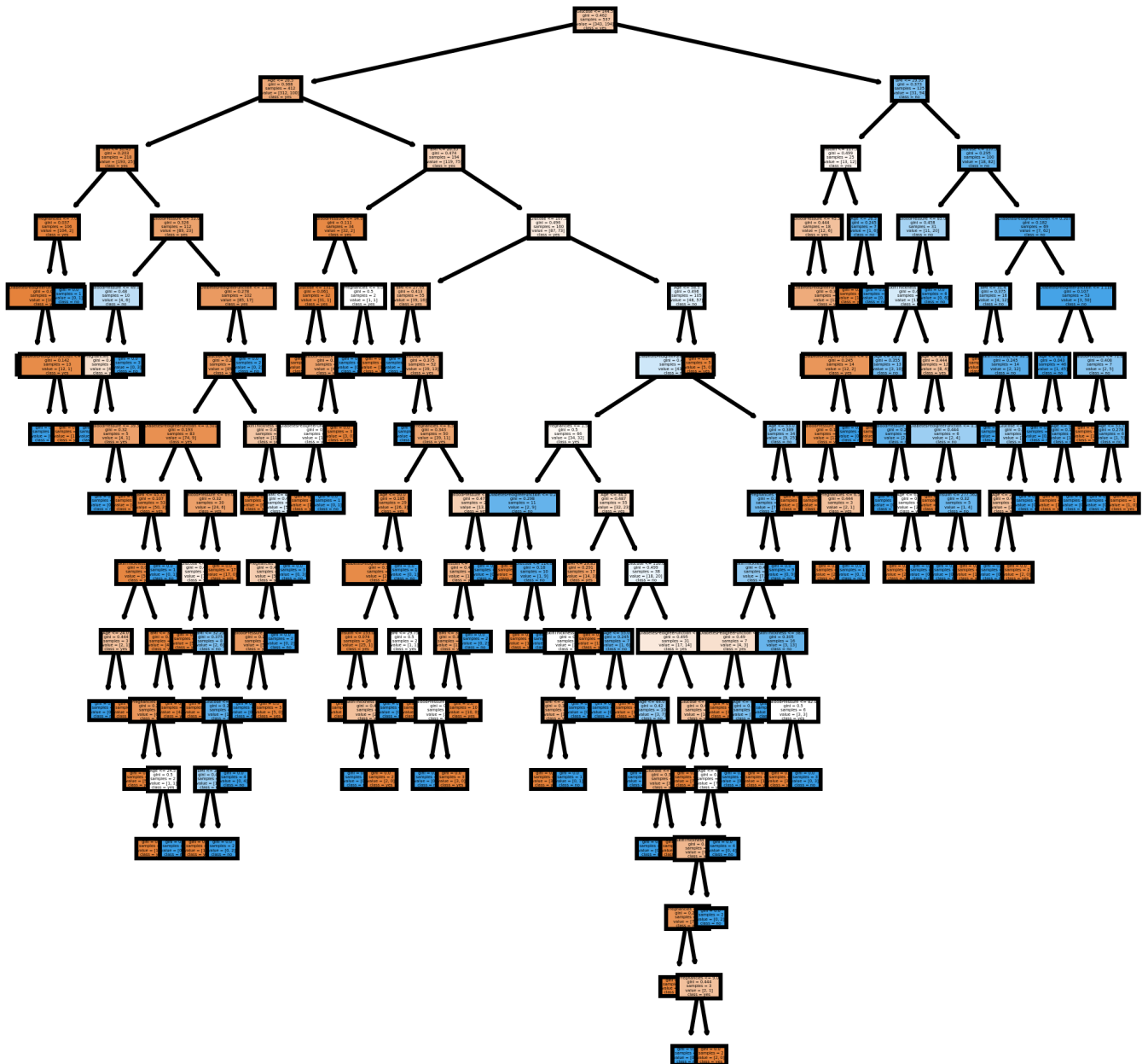
```
print("Train Accuracy :", round(dt.score(x_train, y_train),3))
print("Test Accuracy :", round(dt.score(x_test, y_test),3))
```

Train Accuracy : 1.0  
Test Accuracy : 0.749

```
In [59]: from sklearn.tree import plot_tree
fn=x_train.columns
cn=["yes", "no"]
```

```
# Setting dpi = 300 to make image clearer than default
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (5,5), dpi=500)
```

```
dt_plot=plot_tree(dt,
    feature_names = fn,
    class_names=cn,
    filled = True);
```





# Grid search

```
In [60]: param={"criterion":["gini", "entropy"],
               "max_depth":[5, 7, 9, 11],
               "min_samples_split":[20, 50, 100, 120, 150],
               "min_samples_leaf":[10, 20, 50, 80, 100]}

dtg=DecisionTreeClassifier(random_state=0)

gd=GridSearchCV(estimator=dtg, param_grid=param, cv=10, n_jobs=-1, verbose=2)
gd.fit(x_train, y_train)
```

Fitting 10 folds for each of 200 candidates, totalling 2000 fits

```
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=120; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=120; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=10, min_samples_split=150; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=20; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=20; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=120; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=120; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=150; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=20, min_samples_split=150; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=50, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=50, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=50, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=50, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=80, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=80, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=80, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=80, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=80, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=80, min_samples_split=50; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=100, min_samples_split=100; total time= 0.0s
[CV] END criterion=gini, max_depth=5, min_samples_leaf=100, min_samples_split=100; total
```



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js







Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

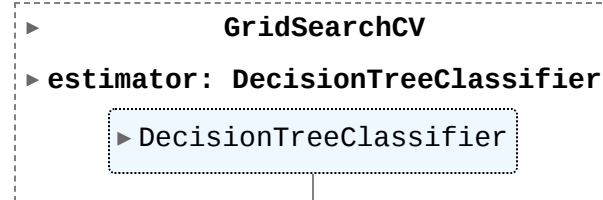
Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js





```
l time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=10, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=120; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=120; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=120; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_sam
```

Out[60]:



```

ples_split=120; total time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=120; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=120; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=120; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=120; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=5, min_samples_leaf=100, min_samples_split=150; to
tal time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=20; tota
l time= 0.0s
[CV] END criterion=entropy, max_depth=7, min_samples_leaf=10, min_samples_split=50; tota
l time= 0.0s

```

In [61]: `gd.best_params_`

Out[61]: `{'criterion': 'entropy',  
'max_depth': 5,  
'min_samples_leaf': 10,  
'min_samples_split': 100}`

In [62]: `gd.best_estimator_`

Out[62]: ▼ DecisionTreeClassifier  
DecisionTreeClassifier(criterion='entropy', max\_depth=5, min\_samples\_leaf=10, min\_samples\_split=100, random\_state=0)

In [63]: dt\_new=DecisionTreeClassifier(max\_depth=5, min\_samples\_leaf=10, min\_samples\_split=100, r  
dt\_new.fit(x\_train, y\_train)

Out[63]: ▼ DecisionTreeClassifier  
DecisionTreeClassifier(max\_depth=5, min\_samples\_leaf=10, min\_samples\_split=100, random\_state=0)

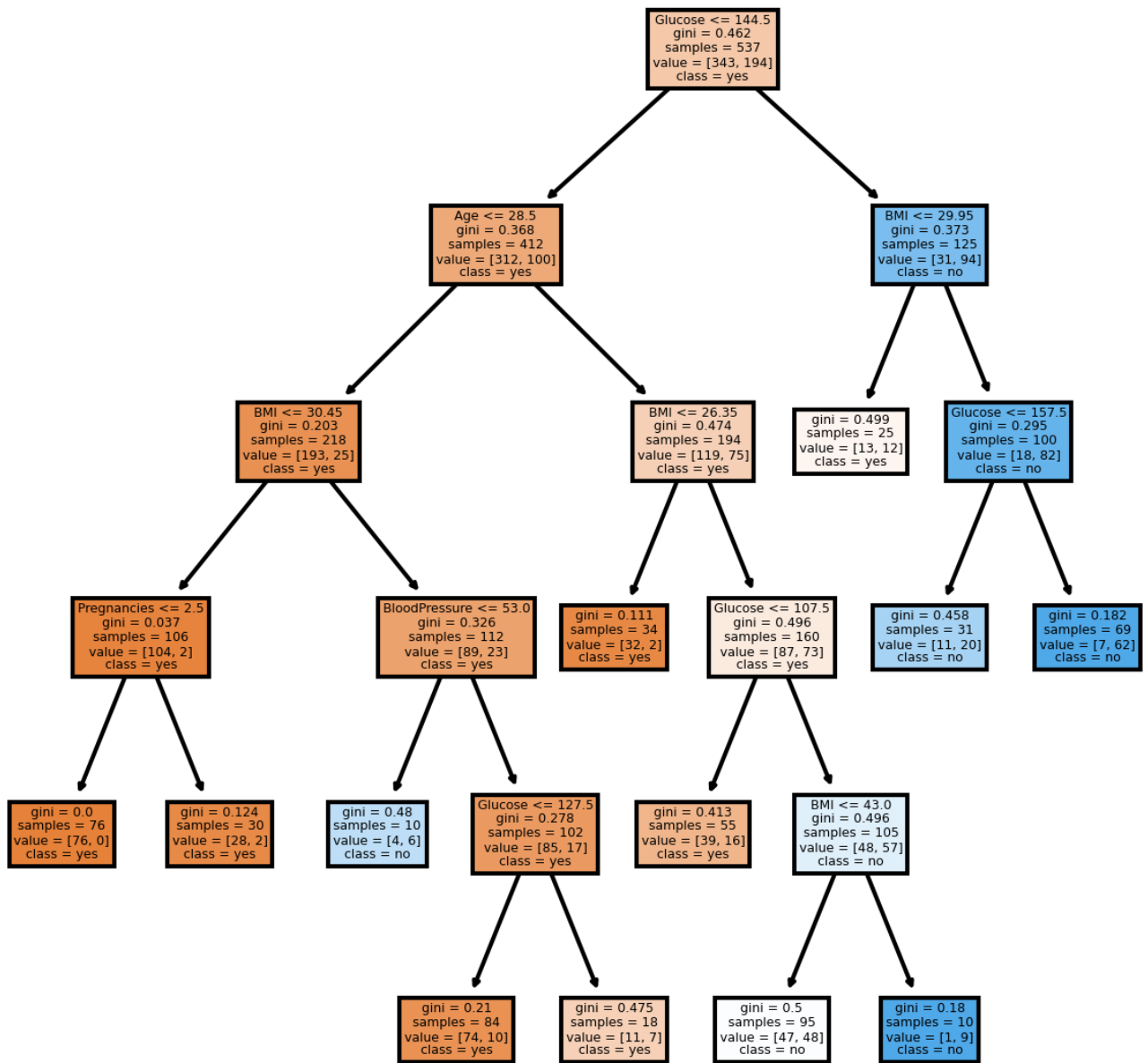
In [64]: dt\_new.score(x\_test, y\_test)

Out[64]: 0.7489177489177489

In [66]: dt\_new.score(x\_train, y\_train)

Out[66]: 0.7783985102420856

In [67]: from sklearn.tree import plot\_tree  
fn=x.columns  
cn=["yes", "no"]  
  
fig, axes=plt.subplots(nrows=1, ncols=1, figsize=(5,5), dpi=300)  
  
dt\_plot=plot\_tree(dt\_new, feature\_names=fn, class\_names=cn, filled=True)



```
In [68]: featImp=pd.DataFrame({"Variables":x_train.columns,
                                "imp":dt_new.feature_importances_}).sort_values(by=["imp"], ascending=False)
featImp
```

Out[68]:

	Variables	imp
1	Glucose	0.603729
5	BMI	0.204751
7	Age	0.154712
2	BloodPressure	0.034859
0	Pregnancies	0.001949
3	SkinThickness	0.000000
4	Insulin	0.000000
6	DiabetesPedigreeFunction	0.000000

```
In [69]: dt1=DecisionTreeClassifier(max_depth=5, min_samples_leaf=10, min_samples_split=100, rand  
  
x_train1=x_train[["Glucose", "BMI", "Age", "Pregnancies"]]  
x_test1=x_test[["Glucose", "BMI", "Age", "Pregnancies"]]  
  
dt1.fit(x_train1, y_train)
```

```
Out[69]: ▼ DecisionTreeClassifier  
DecisionTreeClassifier(max_depth=5, min_samples_leaf=10, min_samples_split=100,  
random_state=0)
```

```
In [70]: dt1.score(x_train1, y_train)
```

```
Out[70]: 0.7746741154562383
```

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

```

[CV] END criterion=entropy, max_depth=11, min_samples_leaf=50, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=50, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=50, min_samples_split=100; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=50, min_samples_split=100; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=50, min_samples_split=120; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=50, min_samples_split=120; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=80, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=80, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=80, min_samples_split=100; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=80, min_samples_split=100; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=80, min_samples_split=150; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=80, min_samples_split=150; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=100, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=100, min_samples_split=20; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=100, min_samples_split=100; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=100, min_samples_split=100; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=100, min_samples_split=120; total time= 0.0s
[CV] END criterion=entropy, max_depth=11, min_samples_leaf=100, min_samples_split=120; total time= 0.0s

```

```
In [72]: dt1.score(x_test1,y_test)
```

```
Out[72]: 0.7489177489177489
```

```
In [73]: from sklearn.tree import export_text

rules=export_text(dt1, feature_names=x_train1.columns.to_list())
print(rules)
```

```

| --- Glucose <= 144.50
| | --- Age <= 28.50
| | | --- BMI <= 30.45
| | | | --- Pregnancies <= 2.50
| | | | | --- class: 0
| | | | --- Pregnancies > 2.50
| | | | | --- class: 0
| | | --- BMI > 30.45
| | | | --- Glucose <= 127.50
| | | | | --- class: 0
| | | | --- Glucose > 127.50
| | | | | --- class: 0
| | --- Age > 28.50
| | | --- BMI <= 26.35
| | | | --- class: 0
| | | --- BMI > 26.35
| | | | --- Glucose <= 107.50
| | | | | --- class: 0
| | | | --- Glucose > 107.50
| | | | | --- BMI <= 43.00
| | | | | | --- class: 1
| | | | | --- BMI > 43.00
| | | | | | --- class: 1
| --- Glucose > 144.50
| | --- BMI <= 29.95
| | | --- class: 0
| | --- BMI > 29.95
| | | --- Glucose <= 157.50
| | | | --- class: 1
| | | --- Glucose > 157.50
| | | | --- class: 1

```

```

In [74]: pred_train=dt1.predict(x_train1)
         pred_test=dt1.predict(x_test1)

```

```

In [75]: prob_train1=dt1.predict_proba(x_train1)[:,-1]
         prob_test1=dt1.predict_proba(x_test1)[:,-1]

```

```

In [76]: def class_metrics(act, pred, probs):
         ac1=metrics.accuracy_score(act, pred)
         rc1=metrics.recall_score(act, pred)
         pc1=metrics.precision_score(act, pred)
         f1=metrics.f1_score(act, pred)
         auc1=metrics.roc_auc_score(act, pred)
         result={"Accuracy":ac1, "Recall":rc1, "Precision":pc1, "F1 Score":f1, "AUC":auc1}

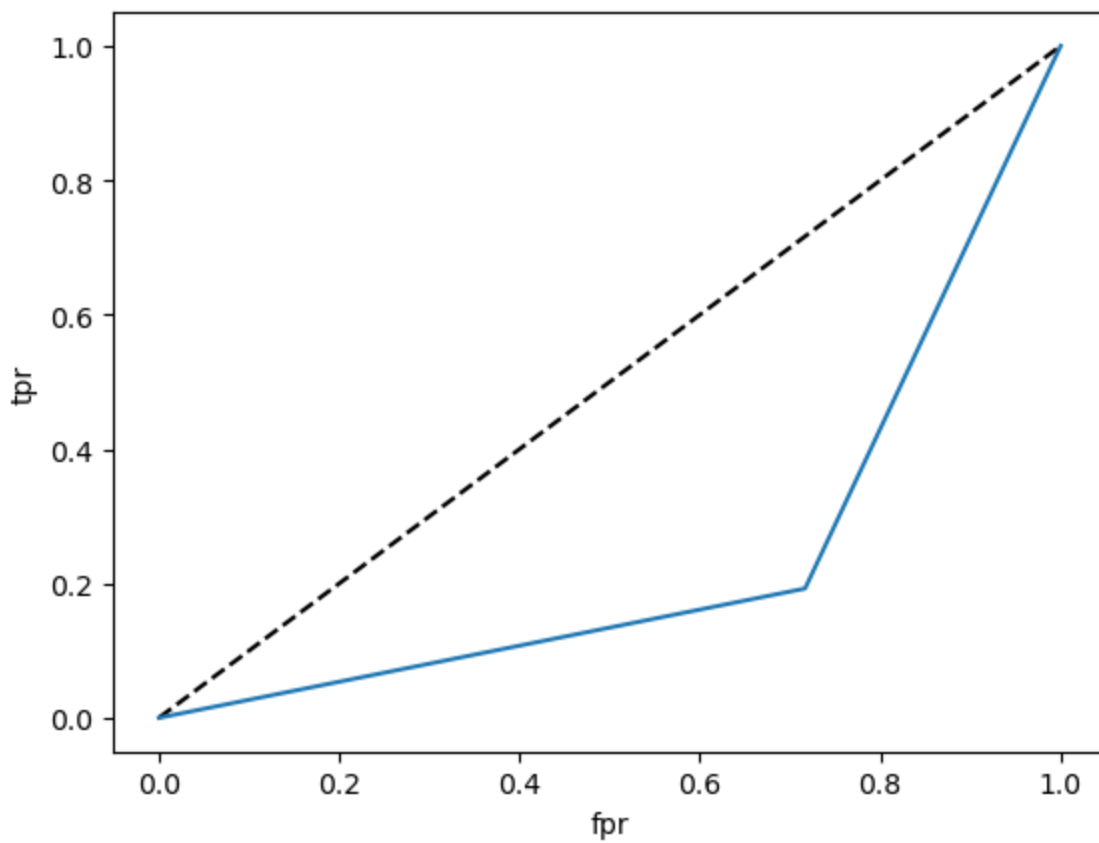
         tpr, fpr, threshold=metrics.roc_curve(act, pred)
         plt.plot([0,1],[0,1],"k--")
         plt.plot(fpr,tpr)
         plt.xlabel("fpr")
         plt.ylabel("tpr")
         plt.show()
         return result

```

```

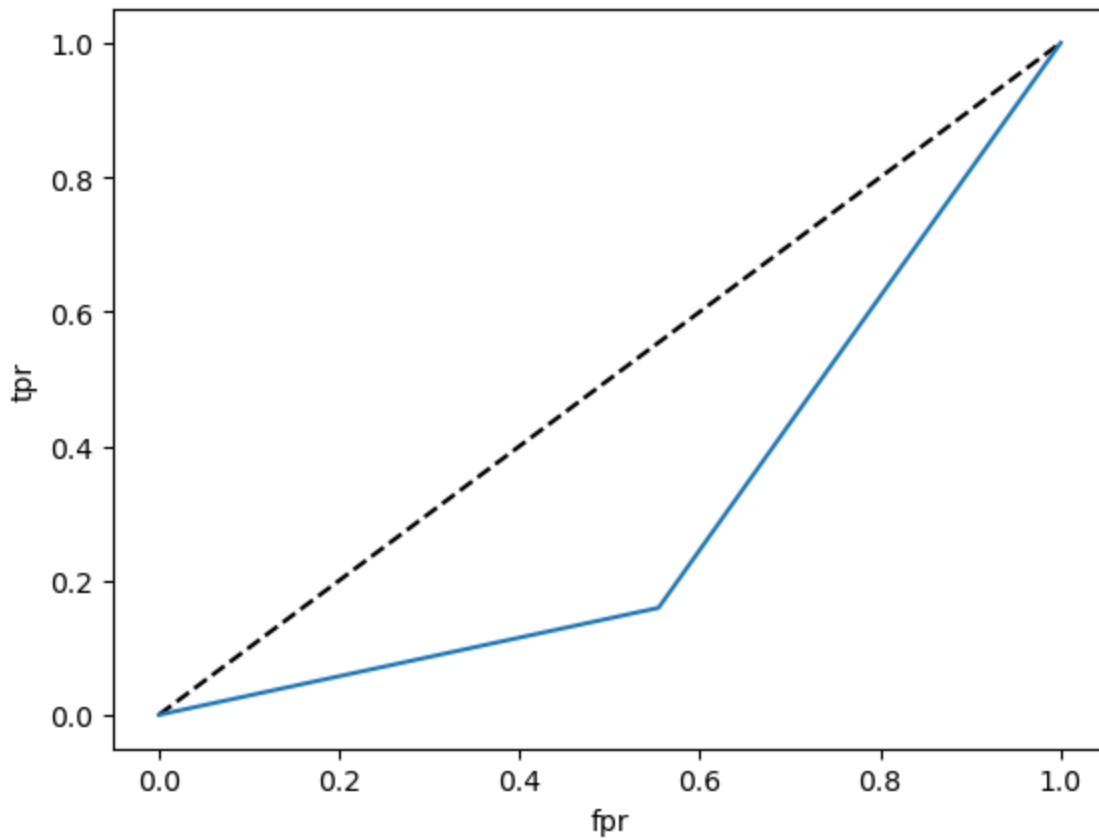
In [77]: class_metrics(y_train, pred_train, prob_train1)

```



```
Out[77]: {'Accuracy': 0.7746741154562383,  
          'Recall': 0.7164948453608248,  
          'Precision': 0.6780487804878049,  
          'F1 Score': 0.6967418546365914,  
          'AUC': 0.7620375101439693}
```

```
In [78]: class_metrics(y_test, pred_test, prob_test1)
```



```
Out[78]: {'Accuracy': 0.7489177489177489,  
          'Recall': 0.5540540540540541,  
          'Precision': 0.6212121212121212,  
          'F1 Score': 0.5857142857142857,  
          'AUC': 0.6974091926321225}
```

In [ ]: