**Prostate Cancer Correlation in Python**

The following python program uses a data file [Prostate\_Cancer.csv](https://www.kaggle.com/imkrkannan/prostate-cancer-prediction-using-ml-algorithms/data)

The first step is to clean and prepare the data. Based on the heat map, the best features were selected for training

The testing results were within the expected range

import pandas as pd

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import classification\_report

from sklearn.metrics import roc\_curve

import matplotlib.pyplot as plt

# Read the data

df=pd.read\_csv("Prostate\_Cancer.csv")

df.describe()

# Data clean up

df.drop(columns=['id'], inplace=True, axis=1)

# Convert strings to int

df["diagnosis\_result"].replace({'B':1, 'M':0}, inplace=True)

corrD=df.corr()

corrD

# Show heatmap

sns.heatmap(corrD, square = True)



# Drop uncorrelated data

df.drop(columns=['fractal\_dimension','texture','perimeter'], inplace=True, axis=1)

y=df.diagnosis\_result

x=df.drop(columns=['diagnosis\_result'], inplace=False, axis=1)

x

# Split training and testing data

x\_train, x\_test, y\_train, y\_test=train\_test\_split(x,y,test\_size=0.2, random\_state=10)

x\_train.info()

# ML

logreg = LogisticRegression()

logreg.fit(x\_train,y\_train)

pred\_log = logreg.predict(x\_test)

# Verify results

class\_test=classification\_report(y\_test, pred\_log)

print(class\_test)

# Logistic Regression Classification

pred\_log\_proba = logreg.predict\_proba(x\_test)

pred\_log\_proba =pred\_log\_proba[:, 1]

fpr, tpr, thresholds = roc\_curve(y\_test,pred\_log\_proba)

plt.subplot(331)

plt.plot([0,1],[0,1],'k--')

plt.plot(fpr,tpr, label='ANN')

plt.xlabel('fpr')

plt.ylabel('tpr')

plt.title('ROC Curve Logistic Regression')

plt.grid(True)

