Graphic Era (Deemed to be University)

MACHINE LEARNING IN HEALTHCARE

HEART DISEASE PREDICTION

BY PRAKHAR BHANDARI

CARDIOVASCULAR DISEASES (CVDS)

REPORT BY

Prakhar Bhandari
Sec: ML
University Roll No. - 2015258

+91 8126368858
prakhar.luke@email.com
2019-2023 | BTech CSE

WHAT IS CVDS ?

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Four out of 5CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.

FEATURE SELECTION

USING THE BORUTA FEATURE SELECTION

Working of boruta algorithm:

1. Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
2. Then, it trains a random forest classifier on the extended data set and applies a feature importance measure to evaluate the importance of each feature where higher means more important.
3. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.
4. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest runs.

MODEL DEVELOPMENT

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.metrics import classification_report
from sklearn.metrics import recall_score,precision_score,cla
```

```
# search for optimun parameters using gridsearch
params = {'penalty':['l1','l2'],
          'C':[0.01,0.1,1,10,100],
          'class_weight':['balanced',None]}
logistic_clf = GridSearchCV(LogisticRegression(),param_grid=
```

```
#train the classifier
logistic_clf.fit(X_train,y_train)

logistic_clf.best_params_
```

```
{'C': 10, 'class_weight': None, 'penalty': 'l2'}
```

RESULTS

In [94]:
```
my_data = pd.read_csv('../input/my-data/my_data.csv')
my_data
```

Out[94]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysB
1	20	4	0	0	0	0	0	0	20	120	98

In [95]:
```
my_data=my_data[top_features]
my_data
```

Out[95]:

	age	totChol	sysBP	diaBP	BMI	heartRate	glucose
1	4	120	98	60	22	56	92

In [96]:
```
prediction = logistic_clf.predict(my_data)
```

In [97]:
```
print("YOU REALLY ARE HEALTY !") if prediction[0] == 0 else print("YOU'RE AT RISK ! GET LIFE INSURANCE R
```

```
YOU'RE AT RISK ! GET LIFE INSURANCE READY
```

So according to our model (logistic regression)

There's risk of CVD for me :(

Fortunately for me current model is trained on random set of data , thus the predictons might be incorrent. Also to get even more accurate pridiction we can use different models as well (didn't used them in first place cause I've yet to masted them)

NOTE : current model is only 60% accurate, predictions might be wrong, do consult doctor for real reports.