NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SIP-2K18

ECE DEPARTMENT

CLINICAL DATA ANALYSIS REPORT

SUBMITTED BY:

S.PRANATHI -16EC238

BOKSAM MADHUMITHA-16EC210

STATEMENT:

Analysis of the laboratory records of the patients to understand the relationship between different types of test results and angiogram test. Our main aim is to avoid angiogram test, if we can predict the angiogram test result as negative.

HYPOTHESIS:

A data set containing 77 test results and angiogram report of 70 samples is given.

INSTALLATIONS:

- Python
- Pandas
- Sci-kit Learn
- Numpy
- Xlsxwriter
- Graphviz

The given file should be in CSV format to read it using PANDAS. Now this is called a dataframe. The function to read CSV file using pandas is

```
pd.read_csv (filename)
```

Using K-Fold cross validation, the given data set is divided into training and testing data. K-fold cross validation divides the data into K parts of each containing 1/k samples. The train data set contains K-1 parts where as test data contains remaining 1 part. The function for 5-Fold cross validation is

```
kf = KFold (n_splits=5, shuffle=True)

kf.split ()
```

RANDOM FOREST CLASSIFIER:

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

Now, based on the train data and corresponding outputs, predictions are made on the output of the test data using random forest classifier.

```
Model = RandomForestClassifier ()
model.fit (features_matrix, labels)
predicted_labels = model.predict(test_feature_matrix)
```

Comparing the actual outputs of test data with predicted values the accuracy score is calculated. Taking the mean of all 5 folds will give the overall accuracy.

```
accuracy_score (test_labels, predicted_labels)
```

This accuracy is to be found using single feature at a time of all the 70 samples. Observing those accuracy values, features for which highest accuracies obtained are taken.
Accuracy is also found using two features at a time of all the 70 samples and the features for which highest accuracies obtained are taken. Of all these features top ten are listed below:

- IHD
- PWV
- AUGMENTATION_PRESSURE
- EXERT_ANGINA
- AUGMENTATION_INDEX
- TMT+
- UNSTABLE_ANGINA
- MAP-C2_INDEX
- ANT_WALL_MI
- REFLECTION_MANITUDE

Using the function,

```
feature_importance (data)
```

the features which are of highest importance are noted. Some of the most important features obtained are:

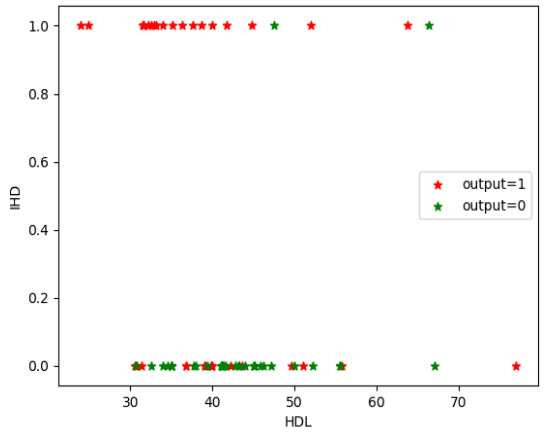
- VIT_D
- CALCIUM
- IHD
- HDL
- PWV

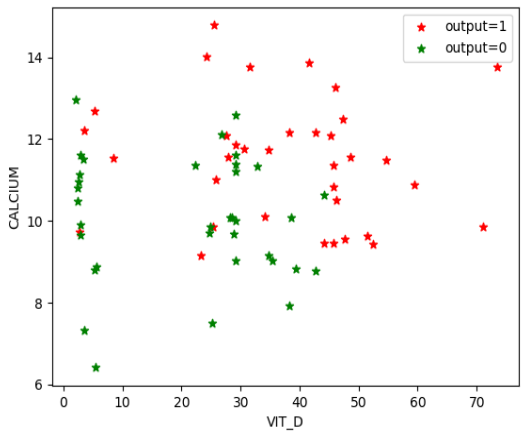
- FGF23
- FGF23.1
- INSULIN
- TC
- RR_MIN

Now scatter plots are plotted for all the combination of two features. These are plotted to know if we can obtain a limit which separates the data with angiogram report as positive and the data with angiogram report as negative. The function to plot scatter plots is

plt.scatter (s1, r1, marker='*', c='r', label='output=1')

But when the scatter plots are plotted for all the above, only for few cases the limit can be found. They are:



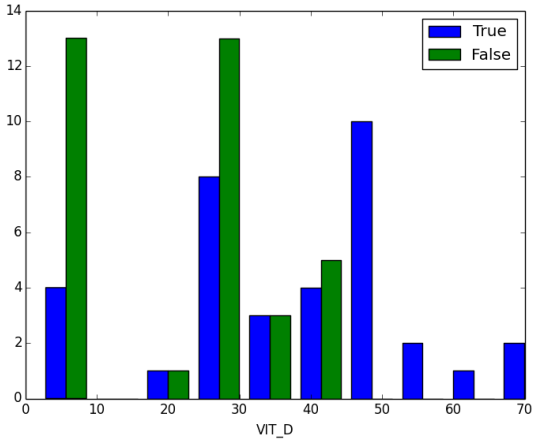


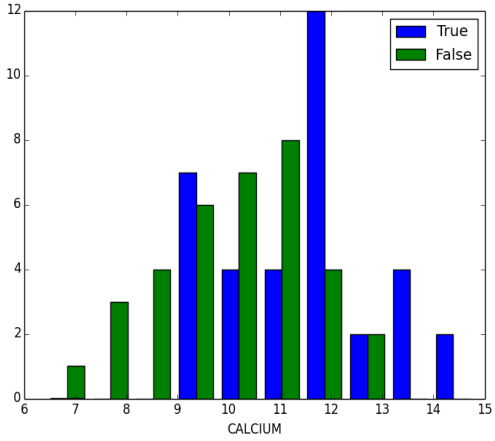
In the above plot HDL and IHD, if IHD=1 and HDL<50 we can say that angiogram report is Positive. Also in the next plot, if VIT_D >45 Angiogram report is positive.

To observe each individual feature, we have to plot Histogram. The function to plot histogram is

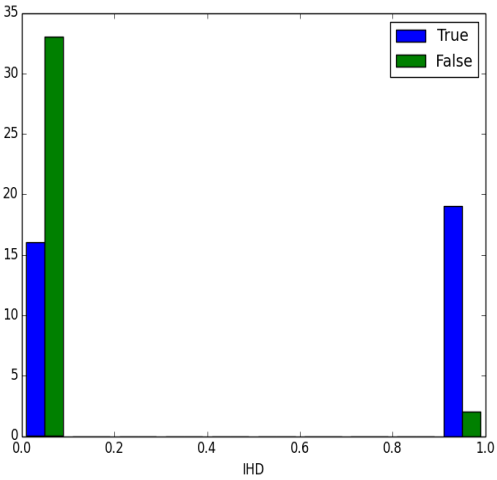
Plt.hist ((x, y))

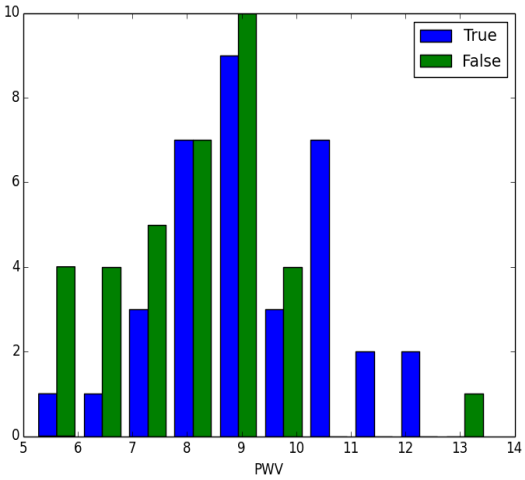
Here, BLUE: Angiogram report is positive
GREEN: Angiogram report is negative





From the above plots, it is clear that for VIT_D >45, angiogram report is positive and for CALCIUM<9, angiogram report is negative.





From the above plots, it is clear that for IHD=1, angiogram report is mostly positive and for 10<PWV<13, angiogram report is positive.

The features are classified into two types based on whether the Angiogram report is positive or negative.

ANGIOGRAM TEST -NEGATIVE	ANGIOGRAM TEST -POSITIVE
CALCIUM < 9	VIT_D >45
INSULIN > 50	IHD=1
PWV < 9	FGF23.1 <500
RR_MIN<20	

Considering all the plots, accuracies and importance of all features, some of the best features are selected and decision trees are made.

DECISION TREE:

A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. The function to find decision tree is:

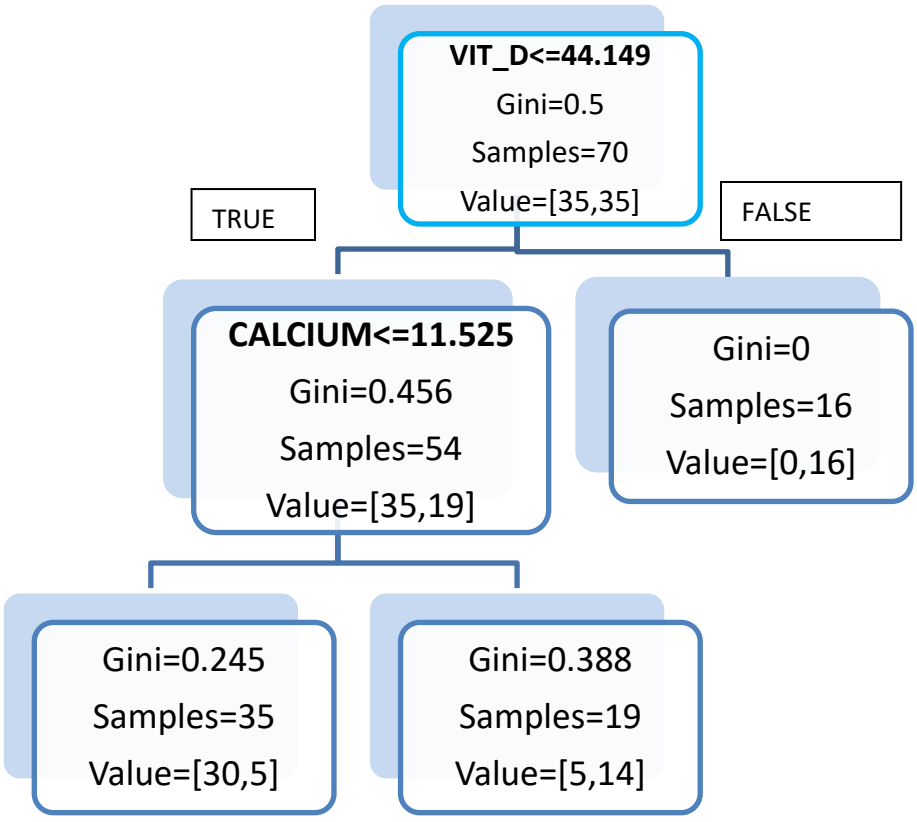
```
clf = tree.DecisionTreeClassifier (max_depth=3)

clf = clf.fit(Z,y)

tree.export_graphviz(clf,out_file=dot_data,max_depth=3,feature_names=Z
.columns, filled=True, rounded=True)

graph = pydot.graph_from_dot_data(dot_data.getvalue()
```

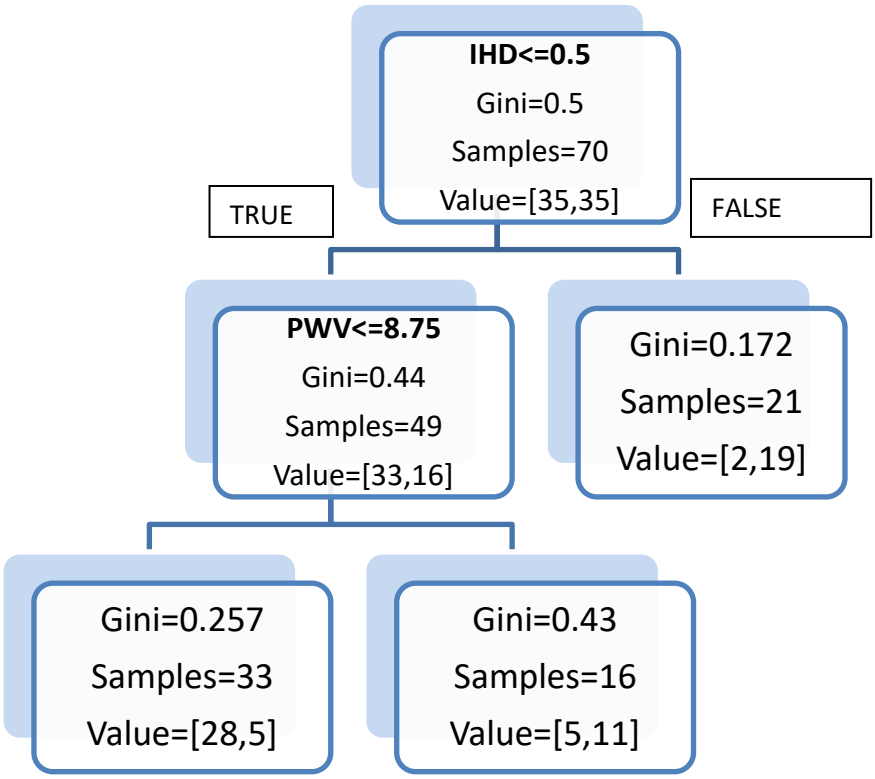
Decision trees are plotted using the best features. The best decision trees with depth=2 are:



For the above decision tree:

ANGIOGRAM TEST -NEGATIVE	**ANGIOGRAM TEST -POSITIVE**
VIT_D <= 45 and CALCIUM <=11	VIT_D > 45
	VIT_D <=45 and CALCIUM >= 11

Above conditions are satisfied by 60 samples among 70.



For the above decision tree:

ANGIOGRAM TEST -NEGATIVE	**ANGIOGRAM TEST -POSITIVE**
IHD=0 and PWV <=9	IHD=1
	IHD=0 and PWV > 9
	IHD=1 and HDL < 50

Above conditions are satisfied by 58 samples among 70.

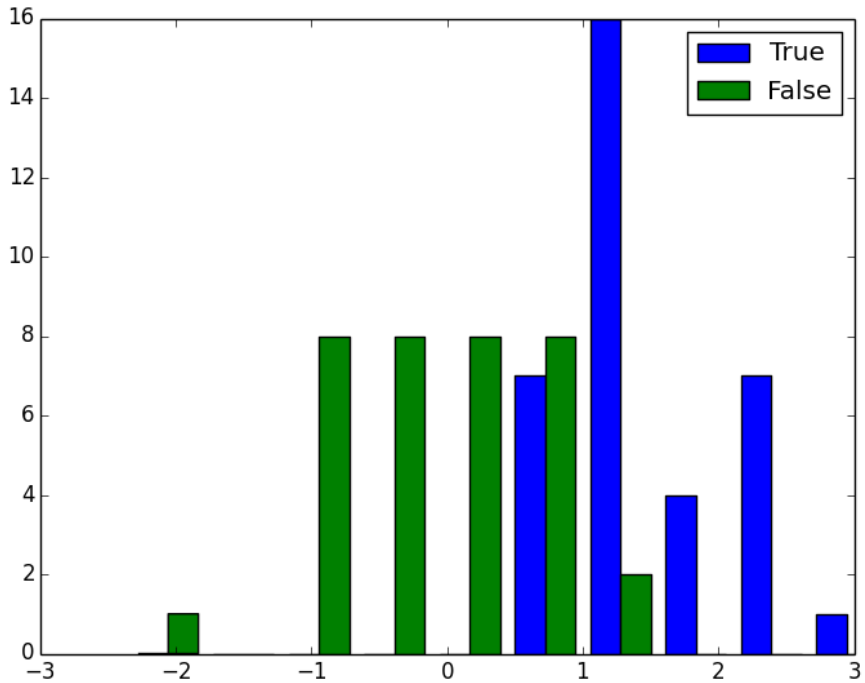
The above results are obtained using random forest classifier. To verify them anther classifier called Support vector Machine (SVM) is used.

SVM:

SVM is a supervised learning method that looks at data and sorts it into one of two categories. Linear SVC is one of the classifier in SVM. The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.

After getting the hyperplane, if one feature is given to the classifier, the perpendicular distance from the feature to the hyperplane will give us the weight of that feature. Similarly, the weights of all 77 features are found. The features with highest weights are taken and aggregate weight for each sample is found.

Histograms are plotted for 70 samples using aggregate weights. The plots in which Positive and negative classes separated are considered so that we can find the hyperplane. The plot obtained is:



This plot is obtained when we give the following features to the classifier: VIT_D, CALCIUM, IHD, HDL, PWV, INSULIN, MAP-C2_INDEX, RR_MIN, AUGMENTATION_PREESSURE, REFLECTION_MANITUDE, TC, FGF23, FGF23.1, UNSTABLE_ANGINA. In this plot, as we cannot separate the classes completely, we cannot find the better hyperplane.

To find the relation between features**, Correlation coefficients** are found. Negative correlation coefficient between two features indicates that they are inversely related to each other where as positive correlation coefficient indicates that they are directly related to each other.

FEATURE 1	**FEATURE 2**	**CORRELATION COEFFICIENT**
RECENT_WORSENING_ANGINA	TEMP	-0.979
HEART_RATE	VOLUME_STROKE	-0.774
HYPERTENSION	PAST_TREAT	-0.705
LV_DYSFUNCTION	DCM	0.701
Cpp	AUGMENTATION_PRESSURE	0.709
BP_H	Csys	0.716
BP_L	Cdia	0.723
CARDIAC_OUTPUT	CARDIAC_INDEX	0.736
WEIGHT	BMI	0.802
BP_L	BP_H	0.815
Csys	Cdia	0.857
TC	LDL	0.858
Cdia	MAP	0.888

Csys	MAP	0.893
VLDL	TG	0.989

RESULT:

From the random forest classifier, we conclude that these features will affect the angiogram report as follows:

ANGIOGRAM TEST -NEGATIVE	**ANGIOGRAM TEST -POSITIVE**
VIT_D <= 45 and CALCIUM <=11	VIT_D > 45
IHD=0 and PWV <=9	VIT_D <=45 and CALCIUM >= 11
	IHD=1
	IHD=0 and PWV > 9
	IHD=1 and HDL < 50

---------- THANK YOU ----------