TECHNOLUTION ASSIGNMENT-DOCUMENT

1.Introduction to the problem

The dataset shared herewith is from the informatics department of a hospital which has patient level vitals for each patient they have on ICU beds and the corresponding alerts based on the status of their experiencing a code-blue event (viz. "Coded" = "high-risk":1 or 'low-risk' :0).

Now, predict whether code blue event occurs or not based on given patient vitals.

2.ALGORITHMS USED

To solve this problem we could use classification based algorithms as the outcome is a classification problem where code blue event occurred or no(It is YES or NO type).

There are many classification algorithms . Inorder to solve this I used LogisticRegression, KNN Classification ,DecisionTree based algorithms.

LogisticRegression: Is one of classification algorithm which is used to determine/predict output for discrete values with help of sigmod function.

SIGMOD FUNCTION: F(x)=1/1+e^(-x)

F(x)=wanted probability

X=input of the equation (eg: mx+b)

KNN-Classification: In this algorithm we calculate and set “K” value using elbow function and then calculate distances from each point using methods

EUCLIDEAN DISTANCE: : d=((x1-y1)^2+(x2-y2)^2))^1/2

MANHATTAN DISTANCE: d=|x1-y1|+|x2-y2|+…..+|xn-yn|

Decision Tree:In this algorithm we making decisions based on splitting nodes into multiple sub nodes.

Split uses following algorithms:

Information Gain: (1-entropy) here we consider node which gives highest gain.

Gini Impurity: (1-Gini) we calculate gini impurity of each split and select node with low gini impurity.

3. EXPERIMENTAL ENVIRONMENT :

I’ve used Jupyter notebook from anaconda which is a free and open-source distribution of python,R programming languages for scientific computing.In jupyter note book

I have installed all required libraries which are imported as of our model requirement.

Libraries like pandas ,numpy are imported as they are used for data extraction and data manipulation .

Matplotlib package is a visualization tool used to plot graphs for understanding the relation between various attributes.

Scikit-learn package is installed in order to import wanted libraries of various pre implemented algorithm methods.

4. Brief introduction to dataset used

Brief introduction to dataset used At first step we assign attributes dependent and independent variables .

vitals_datetime Time at which the patient got admitted to the hospital

heart_rate Patient vital

respiration_over_impedence Patient vital

spirometry_oxygen_saturation Patient vital

pulse Patient vital

blood_pressure_systolic Patient vital

blood_pressure_diastolic Patient vital

blood_pressure_average Patient vital

patient_id Unique-id of the patient

machine_id Unique machine id (ICU bed ID)

Coded Actual alerts [high-risk/low-risk]

Then, we divide dataset into training and testing ..where train data is used to build the model and make predictions on the test data.

For further proceeding we choose **coded** as target variable.

After this we predict Accuracy which gives correctly predicted observations of total observations. DATA SET PREPARATION

1. Imputing missing values.
   We can fill the missing values with help of computing the mean.

2. Removing categorical variables.

Sklearn cannot compute categorical variables like(machine_id,patient_id,vitals_datetime) we need to assign numbers to each variable. We can use dummy encoding to achieve this.

DATASET IS CLEANED AND MANIPULATED (i.e imputed categorical variables and missing values)

5. Performance metrics:

As above algorithms are classification based algorithms, we can generate classification report which contains following metrices:

Accuracy: It denotes how efficiently our model performing.(Model is good if accuracy is greater than70%)

Precision: Tells us amount of meaningful information present among retrieved information.

Recall: Tells us amount of meaningful information retrieved

F1score: its is also measurement of accuracy.

LOGISTIC REGRESSION

Accuracy:0.93

KNN-Classification:

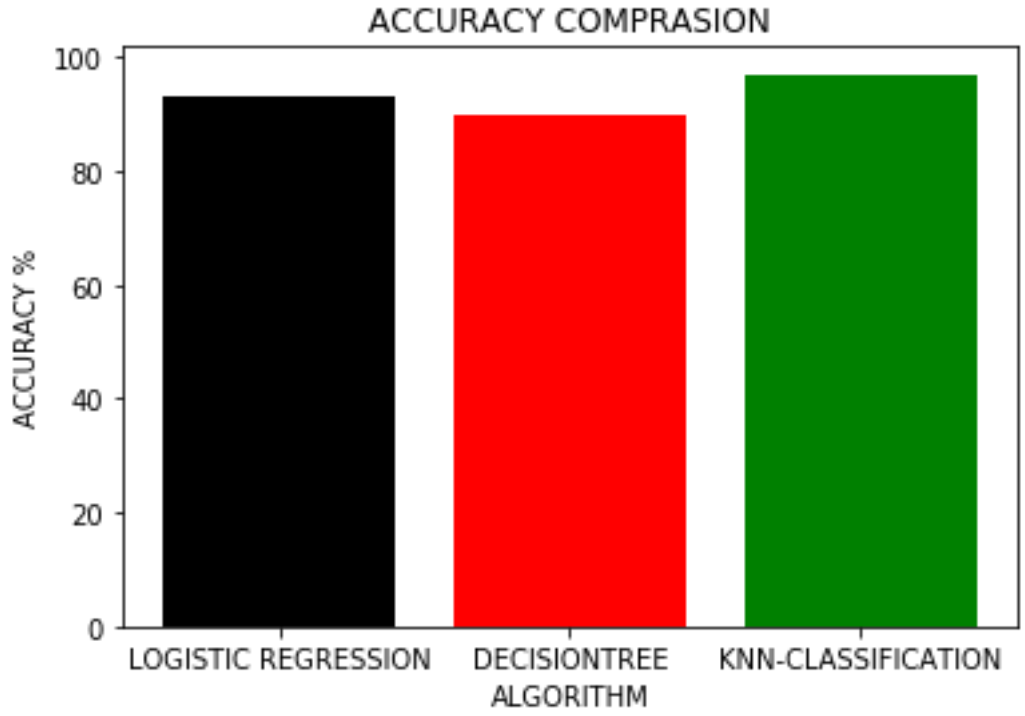
Accuracy:0.97

Decision Tree:

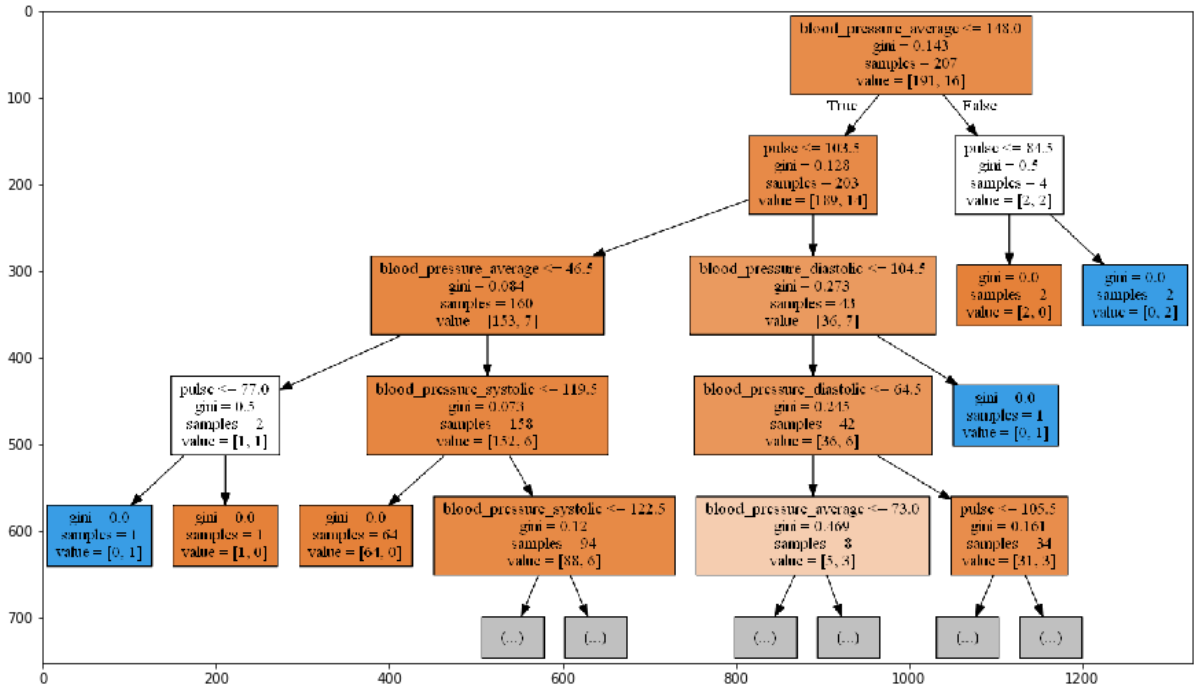
Accuracy:0.90

6.Comparison charts and observations

ACCURACY COMPRASION



DECISION TREE where max_depth = 4



Observations:

1. From above comparison charts we can notice that KNN-Classification predicts model better when compared to Logistic and DecisionTree.
2. As we see based on given patient vitals on testdata there is solid conclusion that we could predict if patient will undergo code blue event or not .
3. In decision tree as we change max_depth the tree grows eventually. Sometimes decision tree can lead to overfitting.

   Final comparison
4. **KNN CLASSIFICATION>LOGISTICREGRESSION>DECISIONTREE.** According to above results.

SUBMITTED BY,

VINAY Y