**Report on Supervised and Unsupervised Machine Learning Algorithms**

Introduction:-

Machine learning playing a critical role in classifying, grouping and predicting the results. Machine learning algorithms come in different types based on their mathematical combination and working principles. Machine Learning has three types:-

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement Machine Learning

Supervised machine learning is in which data is with Label. Algorithms are trained on data with label and able to predict. Supervised Machine Learning has two types.

1. Classification
2. Regression

Classification problems are those in which predication is based on discrete values like he, she, yes, no, risk and non risk, etc. While in regression the predication in continuous like prices of house, .The classification and regression technique have different machine learning algorithms and performance evaluation criteria.

The unsupervised machine learning algorithms have only data without label. The unsupervised machine learning has one type that is Clustering.

Clustering:- Clustering is the unsupervised machine learning in which data point are grouped based on their similar properties. The clustering algorithms are different from classification and regression algorithms.

Reinforcement Machine Learning Algorithms:- In this type of machine learning, no data and label is considered. Algorithms learn from the environment and takes action according to state changes.

Algorithm used in our project:-

**Classification Algorithms:-**

* + 1. Decision Tree
    2. Multilayer Perceptron

**Regression Algorithm:-**

Logistic Regression

**Clustering Algorithms:-**

1. K-means clustering
2. DBSCAN Algorithm
3. Hierarchical Agglomerative Clustering

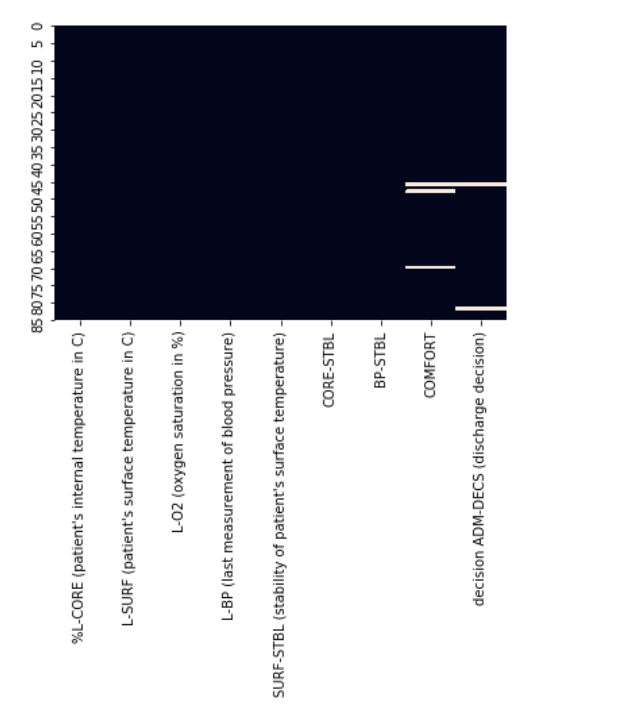
**Methodology:-**

**Dataset description**

The dataset consist of 9 attributes and 90 records. All the attributes are categorical types. The categorical features are difficult to deal with them. Accuracy of such kind of variable always low. We performed the statistical analysis of dataset that shows the top, frequent and unique values of each attribute.

**Missing Values:-**

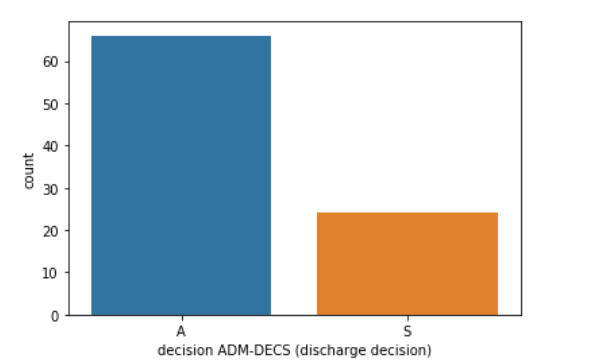
The dataset normal looks that it has no missing values but on exploratory analysis, it shows that some attributes contain the missing values. The missing values are handled by forward method. The missing values are represented graphically.



**The exploratory Data Analysis:-**

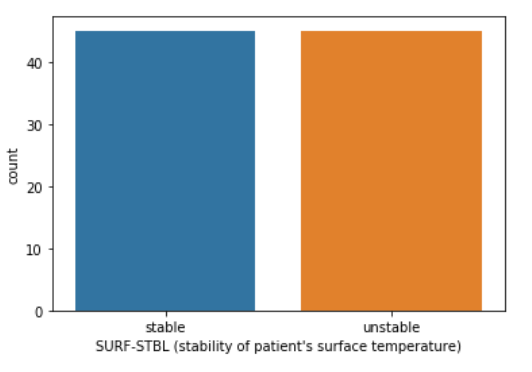
1. **Univariat Analysis:-**

The attribute “decision ADM-DECS (discharge decision)” contain two types of values A and S. The A occur 66 times and S value occurs 24 times in this attribute. This is class attribute in our dataset. The dataset values show that the dataset is imbalanced.

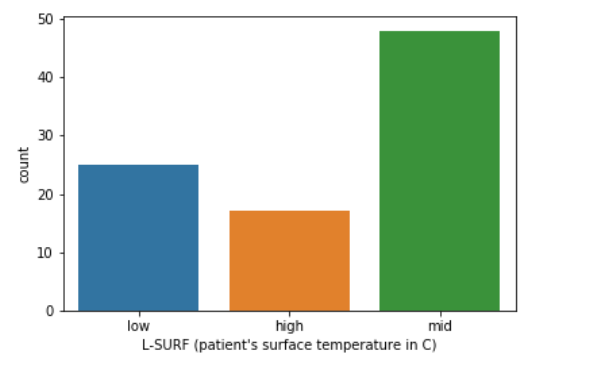
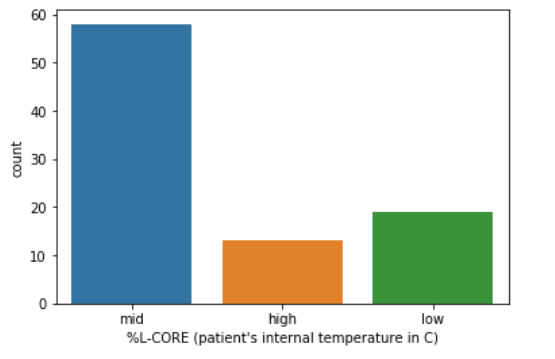
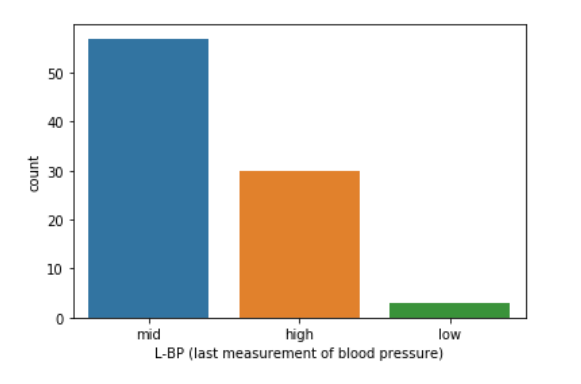
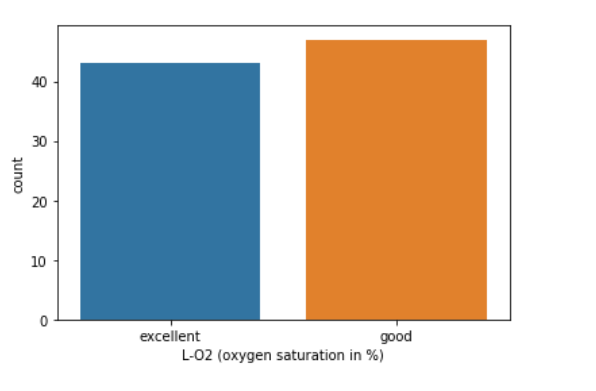
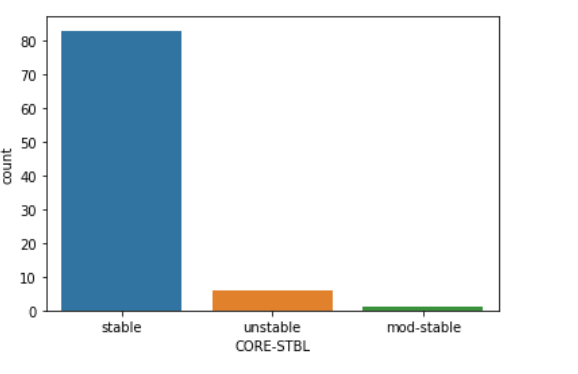
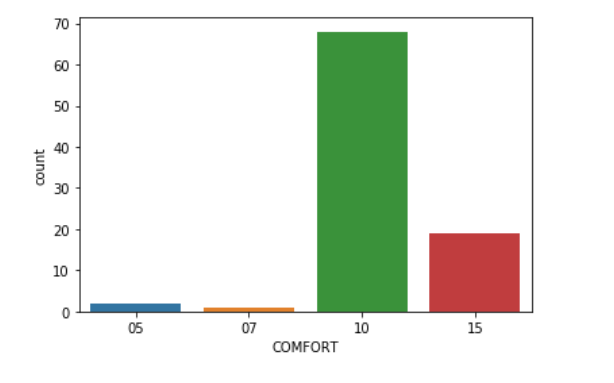
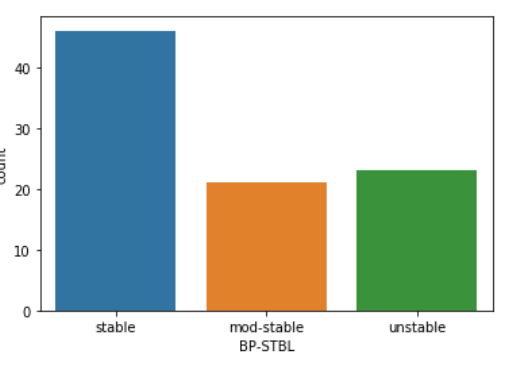


The other attribute of this dataset also imbalance except one attribute "SURF-STBL (stability of patient's surface temperature)". The visual representations of other attribute are shown below.

**Balanced Attribute:-**

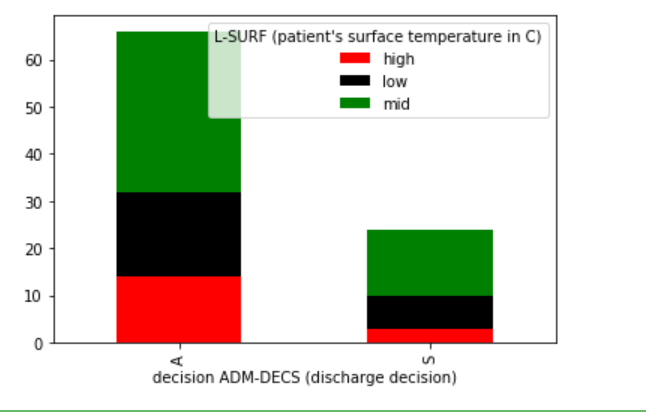
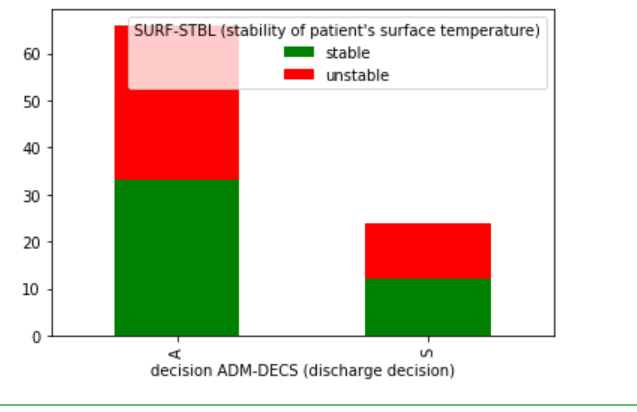
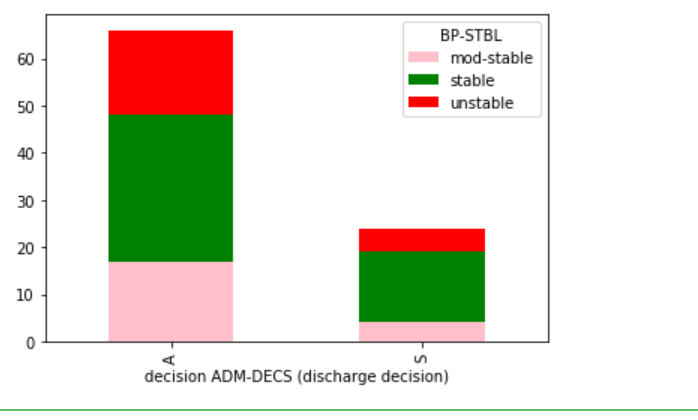
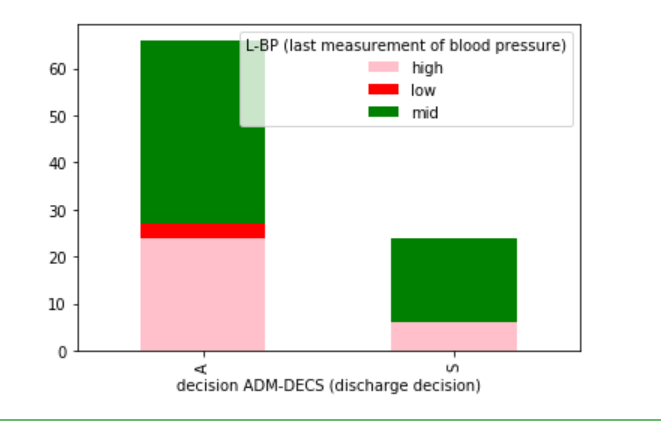


**Imbalanced Attributes:-**

****

1. **Bivariat Analysis**

The Bivariat Analysis shows the contribution of independent variable for prediction of depended variable. Each attribute contribution is shown through stacked graphs.



**Handling Categorical Features:-**

Machine Learning algorithm does not deal with categorical features. We have to convert them into the numeric form before passing to the machine learning algorithms. We used two methods to convert the categorical features into numeric forms.

1. Mapping Function
2. One hot Encoding Technique

**Mapping Function:-**

Mapping functions are used to convert those variables whose values are in two types. In our project, there are three variables that values are of two types. Their values are mapped into 0 and 1.

**One Hot Encoding Technique:-**

One hot encoding technique is used to convert those variable whose values are more than 2. In our project, there are five variables that contain more than two types of values. We used get dummy method to convert them and create extra columns for them.

We do need to normalize the attribute of dataset. As dataset attribute were categorical and after conversion they get the value of 1 and 0. So they are already normalized.

**Training and Test split**

We split the dataset into training and test part. The test size is 25% and chose” decision ADM-DECS (discharge decision)” as class variable and other variable are independent variables.

**Machine Learning Algrothms:-**

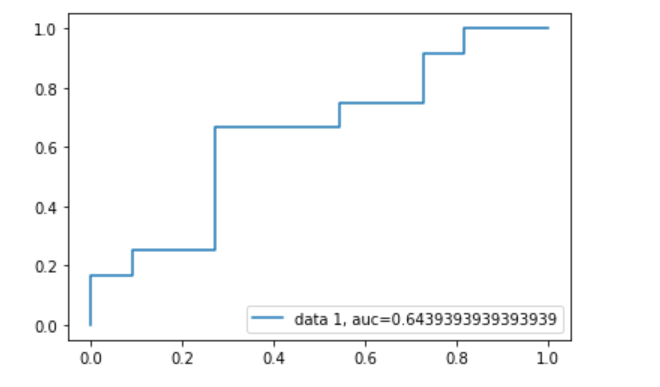
**1-Logistic Regression:-**

Logistic Regression is type of supervised machine learning algorithm that deals the regression problems. This is probabilistic model and its value lie between 0 and 1. It threshold values is 0.5. The Logistic regression take the independent variable and make comparison of their with respect to threshold and predict the depended variable. The formula of logistic regression is 1/1+e(-z)

**Performance Evaluation:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| 0.52 | 0.83 | 0.48 | 0.61 |

ROC Curve of Logistic Regression:-



1. **Multi\_Layered Perceptron:-**

MLP is class a of feedforward neural network. It used supervised machine learning algorithm for training called backpropagation. MLP basically composed three layers one is input , second hiddent and third one is output layer. In our case, we used three hidden layer in which one layer contain 150 nodes, second contain 100 nodes and third contain only 50 nodes. The numbers of maximum iteration are 300. The relu action function used in our case. We used adam optimizer as parameter.

**Performance Evaluation:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| 0.52 | 1.0 | 0.52 | 0.69 |

1. **Decision Tree Model:-**

Decison tree is supervised machine learning algorithm that is used both for classification and regression. It is most power algorithm for classification task. To stop the overfitting, pruning method is used in decision tree.

**Performance Evaluation:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **0.43** | **0.48** | **0.83** | **0.61** |

**Comparison of Performance Evaluation:-**

The above mentioned three supervised machine learning comparison are made based on F1-Score as this is most important metric of performance evaluation. F1-Score of MLP is larger than others because MLP is deep neural network algorithm and used backpropagation for making prediction and good results.

**Unsupervised Machine Learning Algorithms:-**

In this category a data is without label and passed to the clustering algorithm. Clustering algorithm group the data point based on their similarities.

**K-Means Clustering:-**

It is simplest and popular unsupervised machine learning algorithms and it identified the number of centriods and assigned the data point to the nearest data cluster. The algorithm of K-means clustering work as follows;

* Initial the k random points
* Assign the point to closed mean and update coordinates of means
* Repeat this process and increase the number of K

**Performance Evaluation:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| 0.37 | 1.0 | 0.37 | 0.54 |

**DBSCAN**

DBSCAN stand for Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning.

Based on a set of points (let’s think in a bidimensional space as exemplified in the figure), DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

**Performance Evaluation:-**

The data point are small and does no come under DBSCAN. So the number of cluster are 1, so it do not predict anything as the number of features are categorical and on converting the data type, they just remain in only in one cluster.

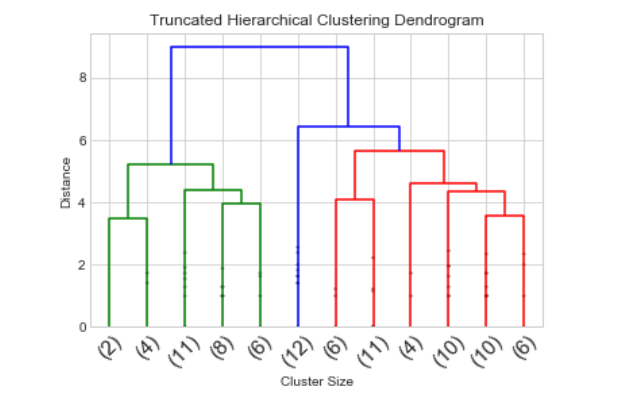
**Hierarchal Agglomerative algorithms**

In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

The basic algorithm of Agglomerative is straight forward.

* Compute the proximity matrix
* Let each data point be a cluster
* Repeat: Merge the two closest clusters and update the proximity matrix
* Until only a single cluster remains

The diagram of Hierarchal agglomerative cluster is given below:-



**Performance Evaluation:-**

The accuracy of this algorithm is 0.47. we used different mechanism for the computation of accuracy by changing linkage types. However maximum accuracy was achieved linkage with ward.

**Performance Evaluation Comparision of Unsupervised ML Algorithms:-**

In unsupervised ml, accuracy is considered most import metric. So according to accuracy Hierarchal agglomerative clustering algorithm is very with respect to linkage of ward.