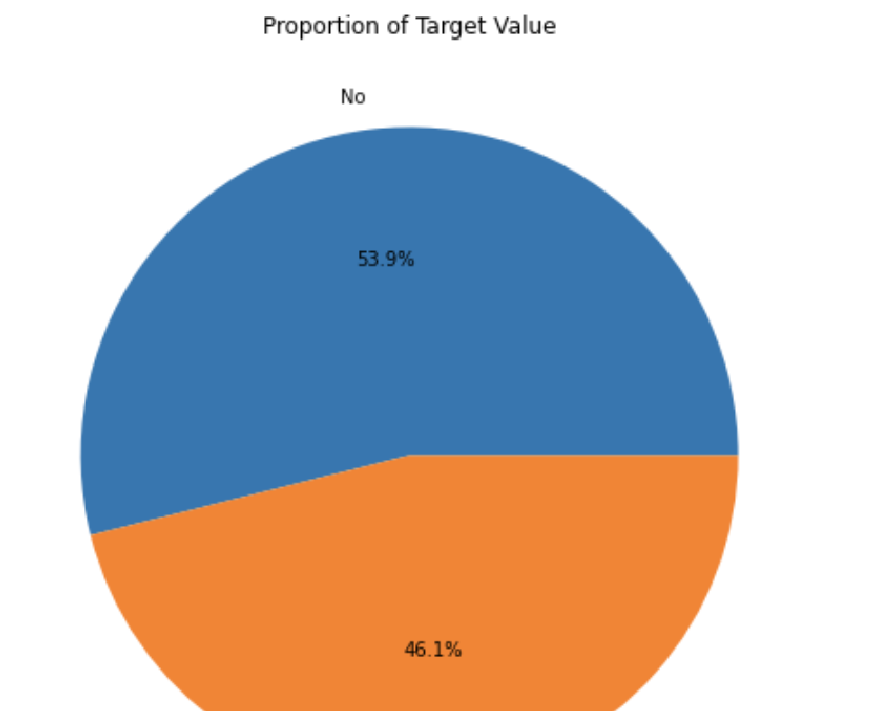
**INTRODUCTION:**

A hospital readmission is when a patient is discharged and gets readmitted in a certain period of time. The quality of the hospital is nowadays determined by the readmission rates. Because of this reason the center for Medicare and Medicaid services has started the hospital readmission reduction program that will help in improving the quality of the hospitals and the care of patients.In 2011, American hospitals spent over $41 billion on diabetic patients who got readmitted within 30 days of discharge. The primary objective of this project is to determine the factors that help to detect the what factors are leading to the readmission of the patients which will directly impact the quality of the hospital and care given to their patients.

**DATASET DESCRIPTION:**

The data is taken from one of the best websites for Data Science dataset called UCI Repository: [UCI Repository link](https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008) .

The data set represents 10 years (1999-2008) of clinical care at 130 US based hospitals. It is a large dataset with 55 attributes that are the variables and 101767 instances that are the rows. It contains data of different hospitals of the patients being treated and admitted there for a disease called diabetes.

Of the total 101767 instances, 53.9% is the negative classes, that is the patients were not re-admitted in the hospital after discharge and 46.1% were the positive class. There are 10 numerical columns and 1 target column that is the objective column.

**EXPLORATORY ANALYSIS AND VISUALIZATION:**

Some of variables in the dataset of the patient admitted are:

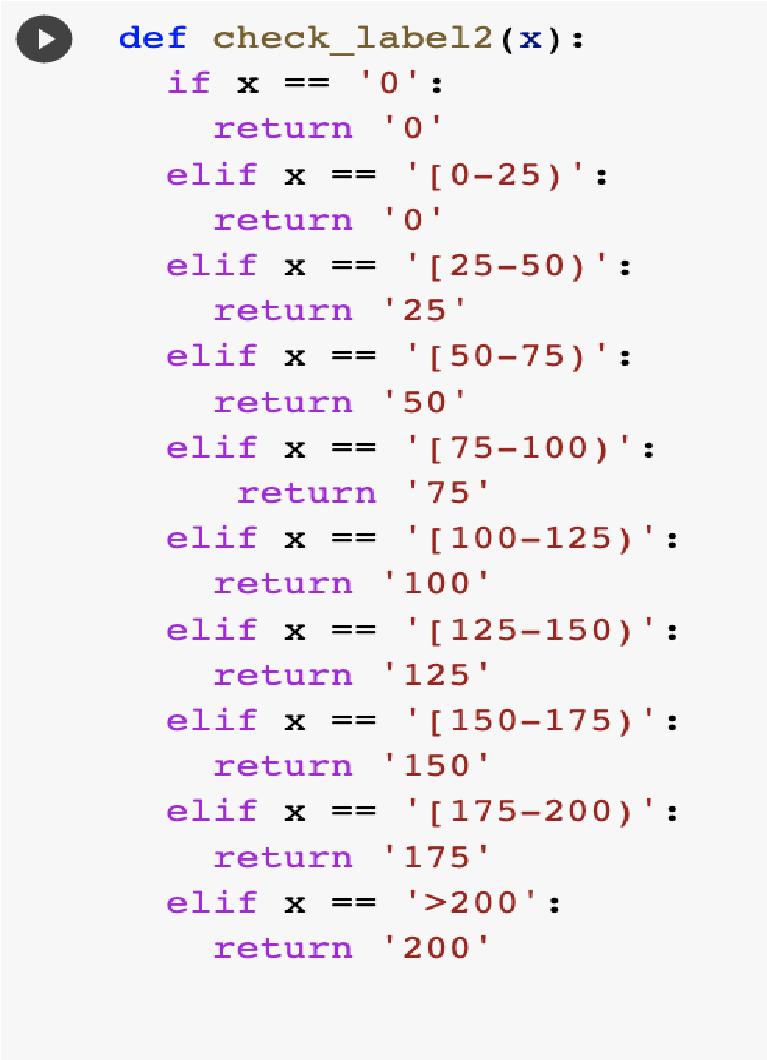
Age, Gender, Weight, Medical Speciality, Number of Lab procedures, Number of medications, Number of Emergency, Diabetes medication, Readmitted.

Data Cleaning , Preprocessing and Feature Engineering:

For data to be ready for the model implementation, it had to be cleaned and processed.

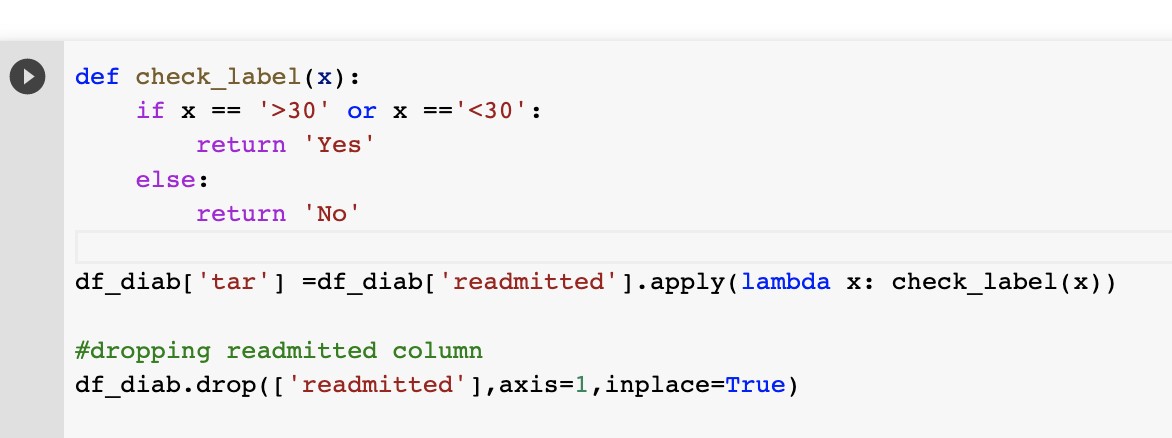
The columns having “?” as inputs were replaced by “0” for further analysis.

The age and weight columns were changed from range to numerical values and two new variables were added to the data that is “Age” and “has\_weight”.

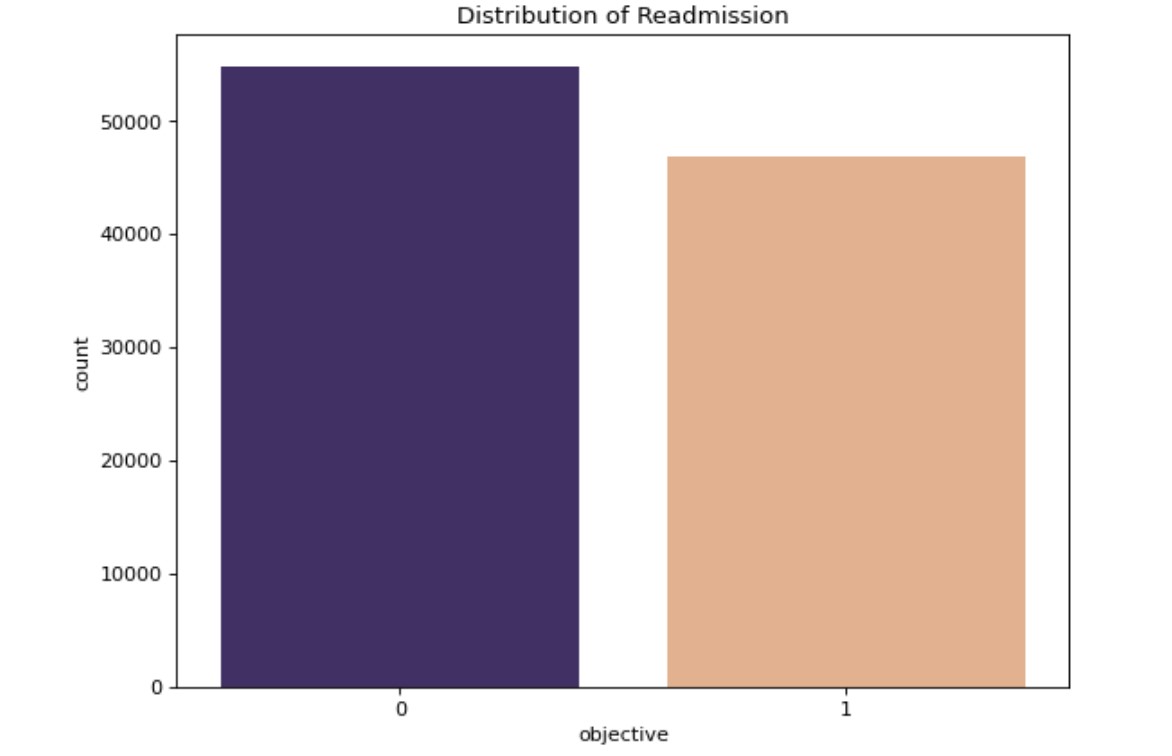
 Age 

Weight

The target column, that is the readmitted column had values “<30”, “>30” and “NO”. These values were mapped to ‘0’ or ‘1’ as “<30” and “>30” were mapped to 1 and “NO” was mapped to ‘0’.

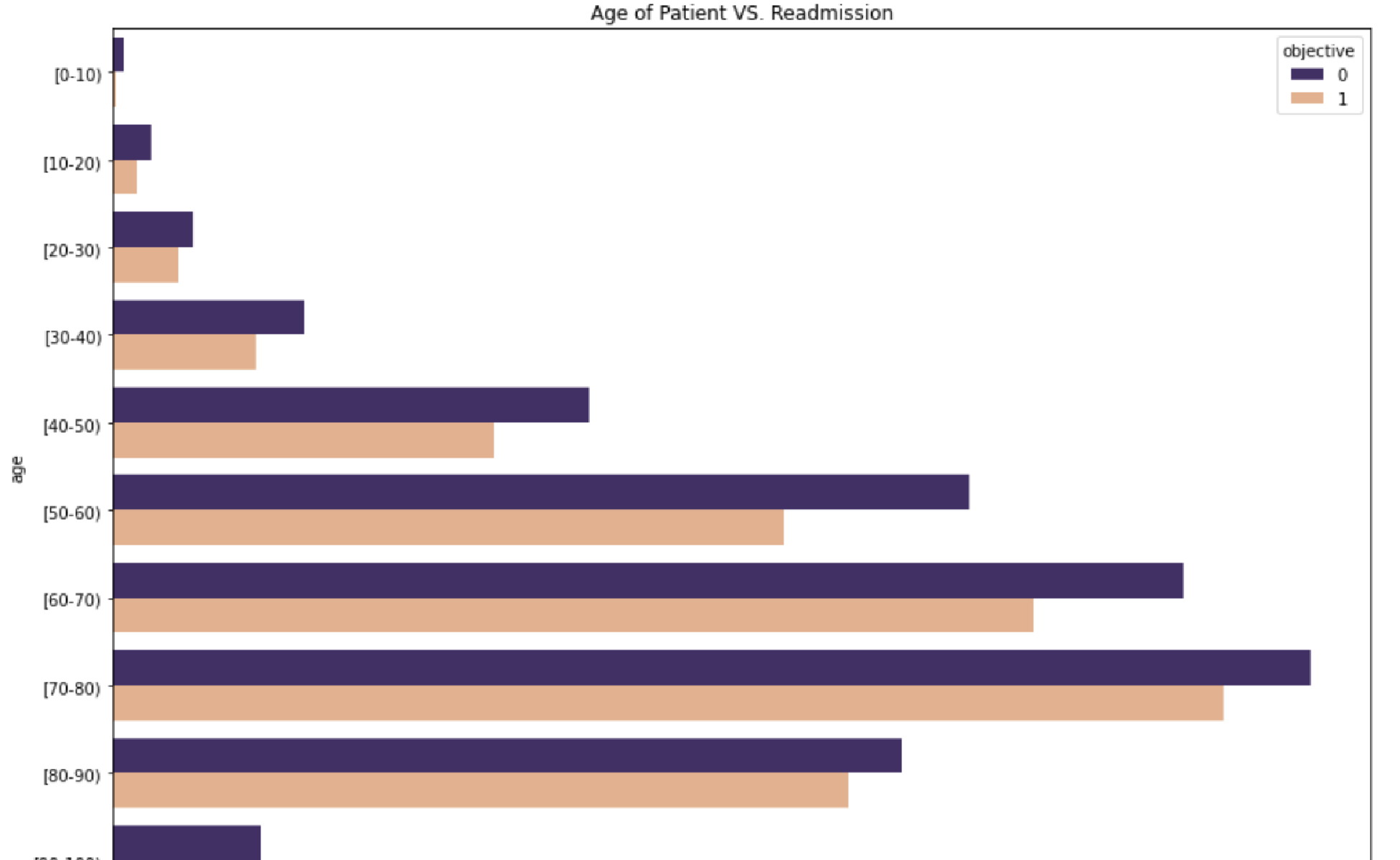


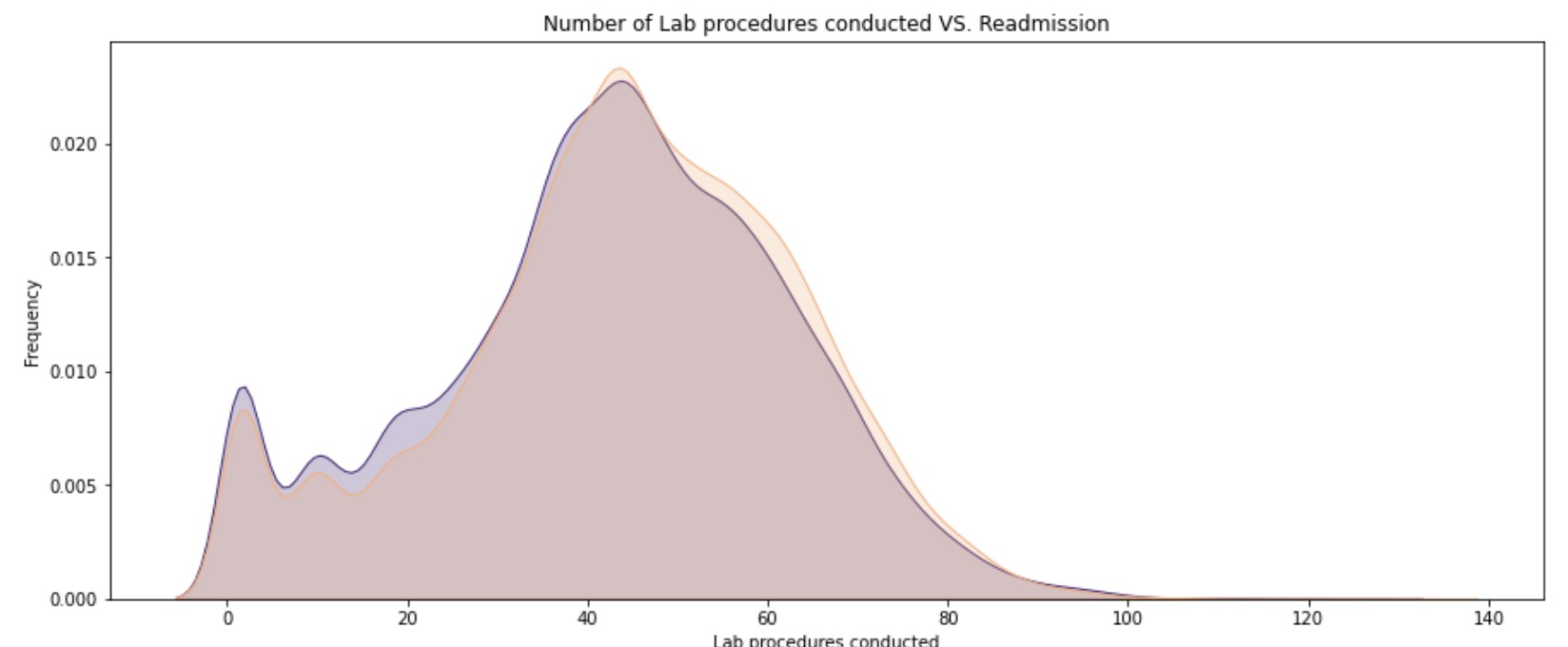
Data Visualization:



As told earlier the objective variable is imbalance, that is the number of readmitted patients is less than the ones not admitted. So, we have used the SMOTE function that is (synthetic minority oversampling technique) to balance the data set. It was done before putting the data in the model for prediction.

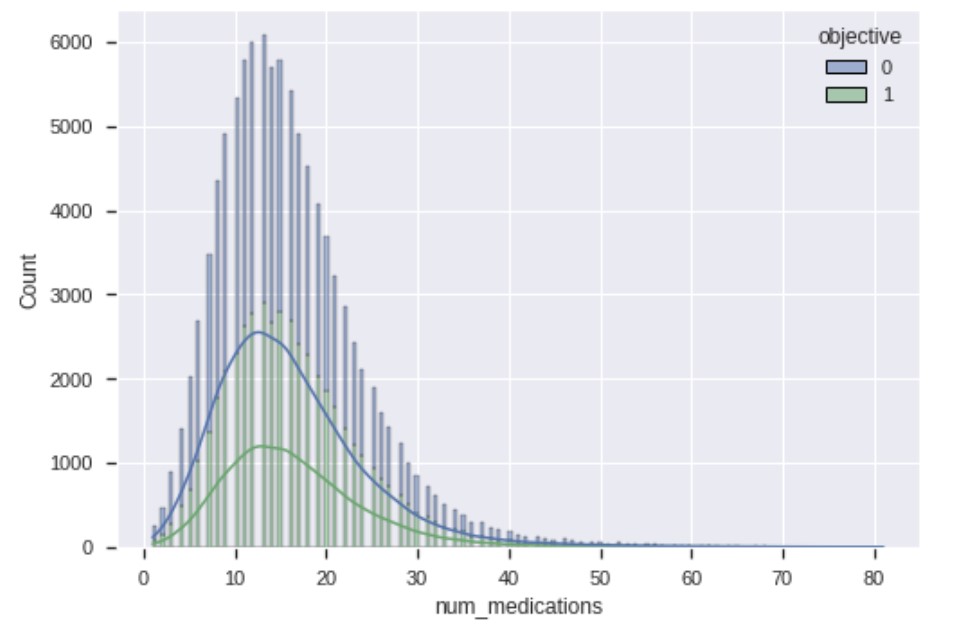
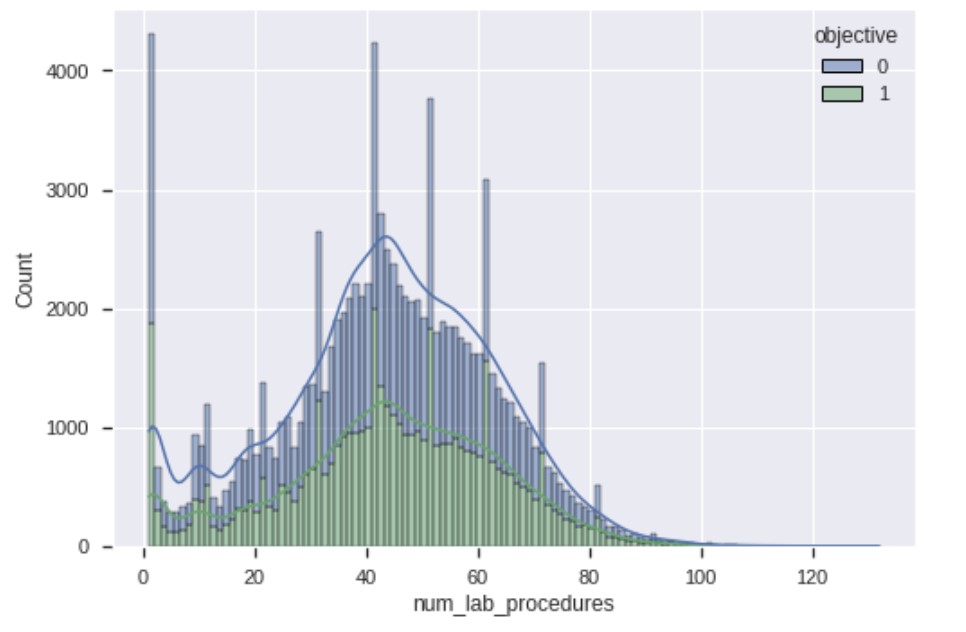
SMOTE function helps in balancing the dataset by randomly increasing the minority class and replacing them, it synthesizes new minority instances between existing minority instances.

We checked different variables that had an effect on readmission and found that both the patients admitted and not admitted had a similar pattern for the number of lab procedures.



Unexpectedly the patients that were given the medications for diabetes were admitted often and patients between the age of 60 and 80 were readmitted often.

The outliers were detected as follows:



**MODEL IMPLEMENTATION:**

The dataset was split into training and testing, 25% for testing and 75 for training.

We implemented the models for the classification:

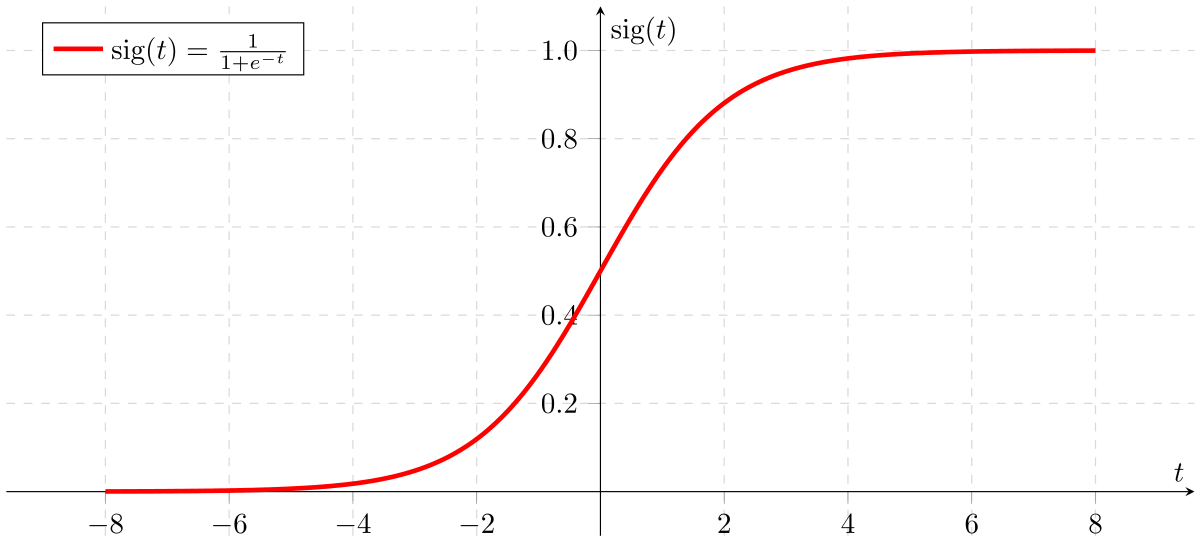
1. Logistic Regression
2. Gaussian Naive Bayes
3. Neural Network

**1. Logistic Regression:**

The main idea of logistic regression is to establish the relation between the input variables that are the dependent variables and the independent variables.

There are two types of logistic regression depending upon the number of classes in the target variable are:

Binary Classification: This type of classification is done when the predicted class is binary that is either ‘0’ or ‘1’. Some examples for such kind of data set are whether the loan will be sanctioned or not i.e. ‘yes’ or ‘ no’ class. To solve this problem, the sigmoid function is used.



Multiclass Classification: This classification is done when the predicted class has more than two outcomes classes, for example to predict whether a particular input will fall under which class. To solve this problem, the softmax function is used to calculate the probabilities and the highest probability obtained is used to classify the record.

For our dataset, the objective variable has binary classes:

|  |  |
| --- | --- |
| **Category** | **Encoded Values** |
| Readmitted | 1 |
| Not Readmitted | 0 |

Outcomes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Learning Rate** | **Tolerance** | **Accuracy** | **Recall** | **Precision** | **F1 Score** |
| 0.001 | 0.00001 | 0.78814 | 0.63392 | 0.88551 | 0.73889 |
| 0.01 | 0.001 | 0.73433 | 0.86490 | 0.66962 | 0.75484 |
| 0.1 | 0.01 | 0.72942 | 0.87119 | 0.66271 | 0.75278 |
| 0.0001 | 0.0001 | 0.79789 | 0.72367 | 0.82729 | 0.77202 |

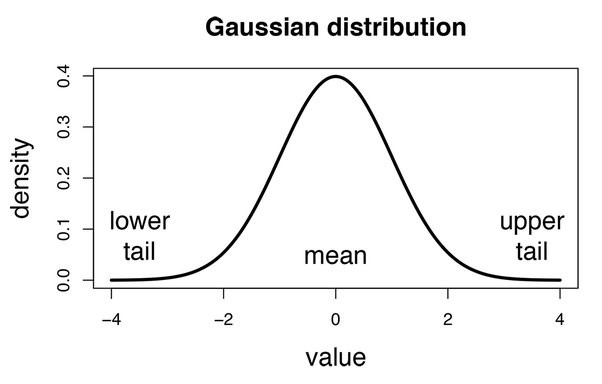
The best result compared to the baseline model is given when the learning rate and the tolerance is 0.0001 and 0.0001. The accuracy is 79.789% with recall, precision, and F1-Score to be 0.72367, 0.82729, and 0.77202 respectively.

**2. Gaussian Naive Bayes:**

Gaussian Naive Bayes is a simple probabilistic algorithm that is based on Bayes’ theorem with strong independence assumptions. Each class follows the gaussian curve according to the algorithm.

For each class mean, standard deviations were calculated. As soon as there is a new input, the probability has to be checked in the gaussian distribution and predicted to which class it will belong.

Gaussian distribution looks like:



Outcomes:

We found the metrics for different batch size of the data:

Batch Size 20%:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accuracy** |  |  | **Cl** | **ass** |
| **0** |  | **1** |
| 0.61757 | **Precision** | 0.58714 |  | 0.72108 |
| **Recall** | 0.87746 |  | 0.33925 |

Batch Size 25%:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accuracy** |  |  | **Cl** | **ass** |
| **0** |  | **1** |
| 0.61836 | **Precision** | 0.58710 |  | 0.72487 |
| **Recall** | 0.87909 |  | 0.34004 |

Batch Size 30%:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accuracy** |  |  | **Cl** | **ass** |
| **0** |  | **1** |
| 0.61603 | **Precision** | 0.58535 |  | 0.72025 |
| **Recall** | 0.87669 |  | 0.33827 |

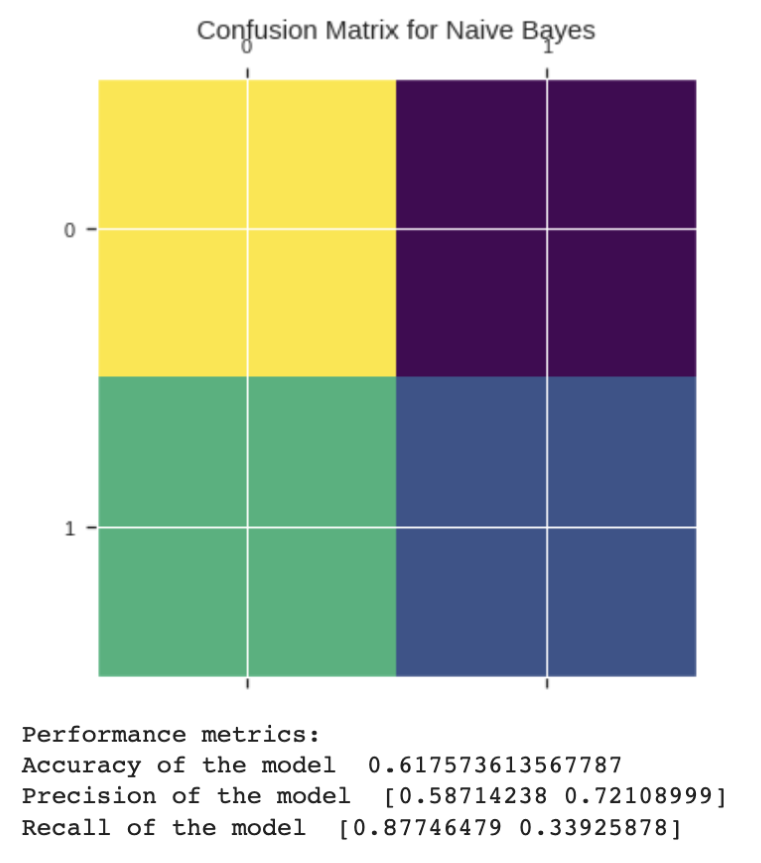
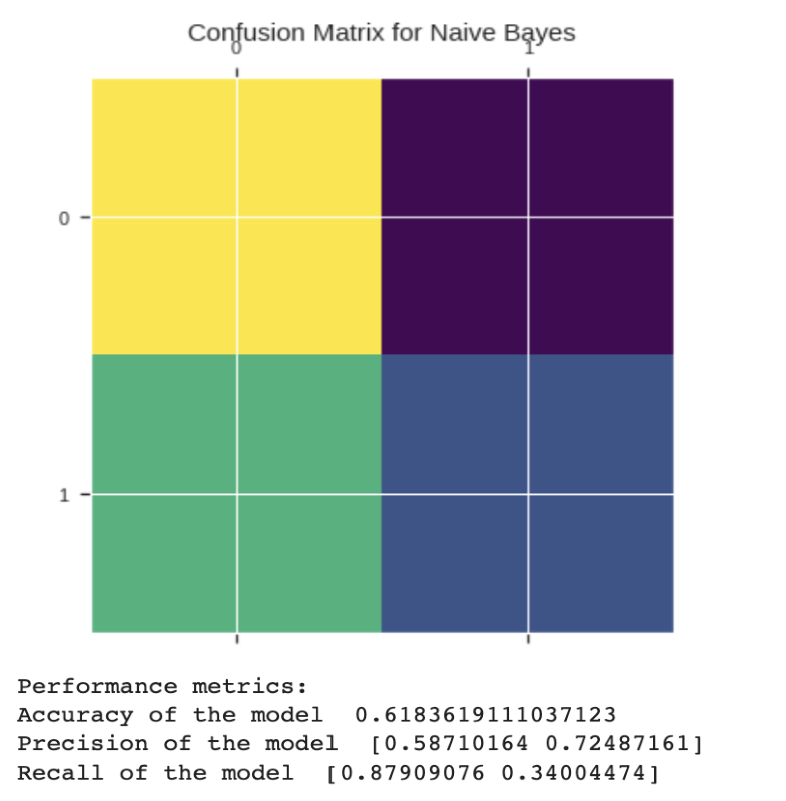
Batch Size 35%:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accuracy** |  |  | **Cl** | **ass** |
| **0** |  | **1** |
| 0.61562 | **Precision** | 0.58411 |  | 0.72207 |
| **Recall** | 0.87656 |  | 0.33943 |

Batch Size 40%:

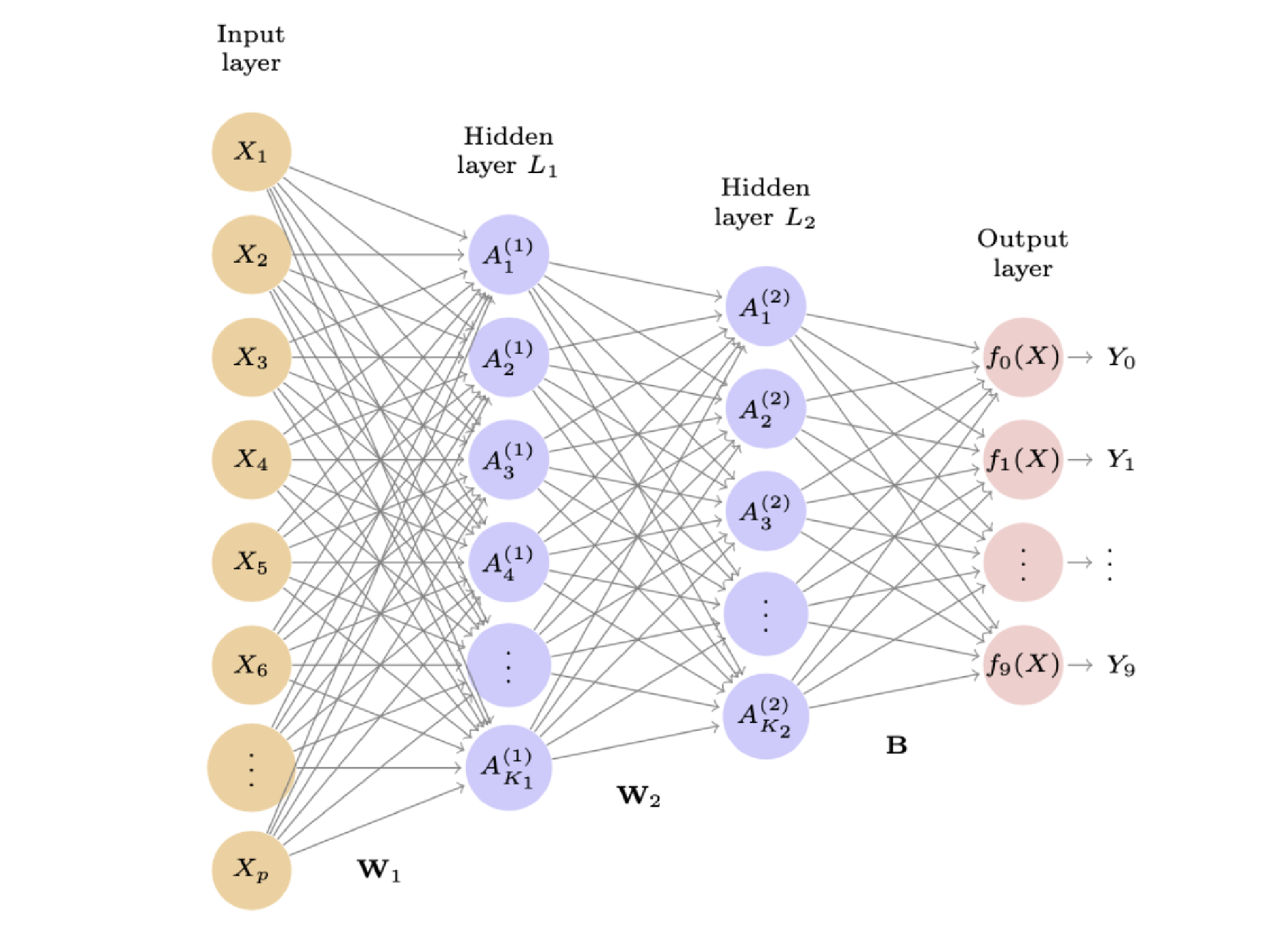
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accuracy** |  |  | **Cl** | **ass** |
| **0** |  | **1** |
| 0.61562 | **Precision** | 0.58318 |  | 0.72590 |
| **Recall** | 0.87854 |  | 0.33874 |

Confusion Matrices:



Batch Size: 20% Batch Size: 25%

**2. Neural Networks:**



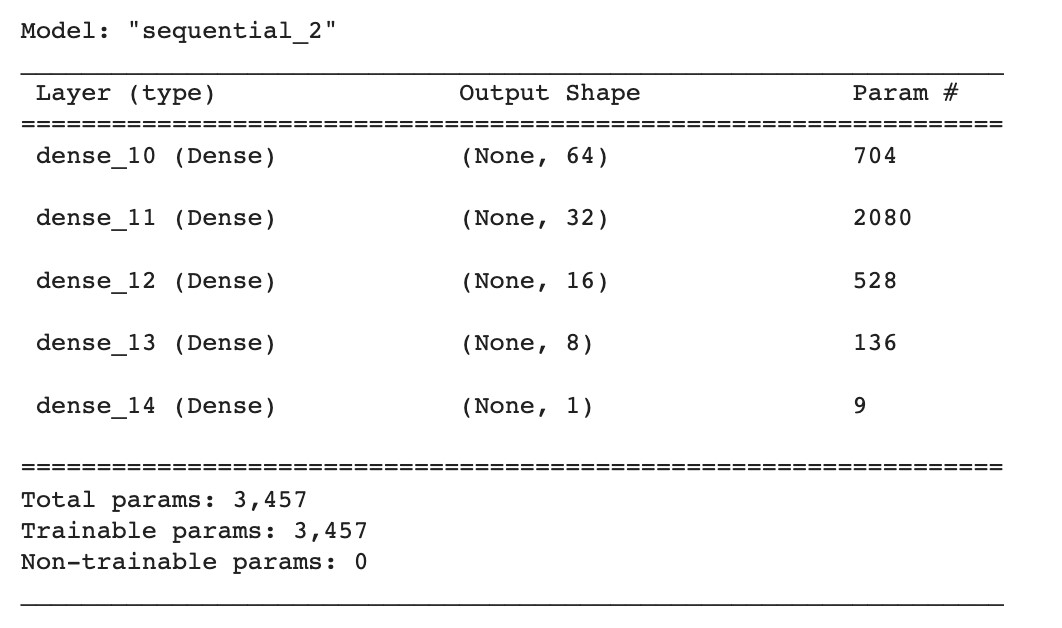
This is how fully connected neural networks look like with one input layer, two hidden layers and one output layer. Every node in the input layer is connected to the hidden layers and finally to the output layer.

We have used the sigmoid function as the activation function for the final output layer as our dataset is a classification problem, it predicts and gives a probability for the respective class that is ‘0’ or ‘1’.

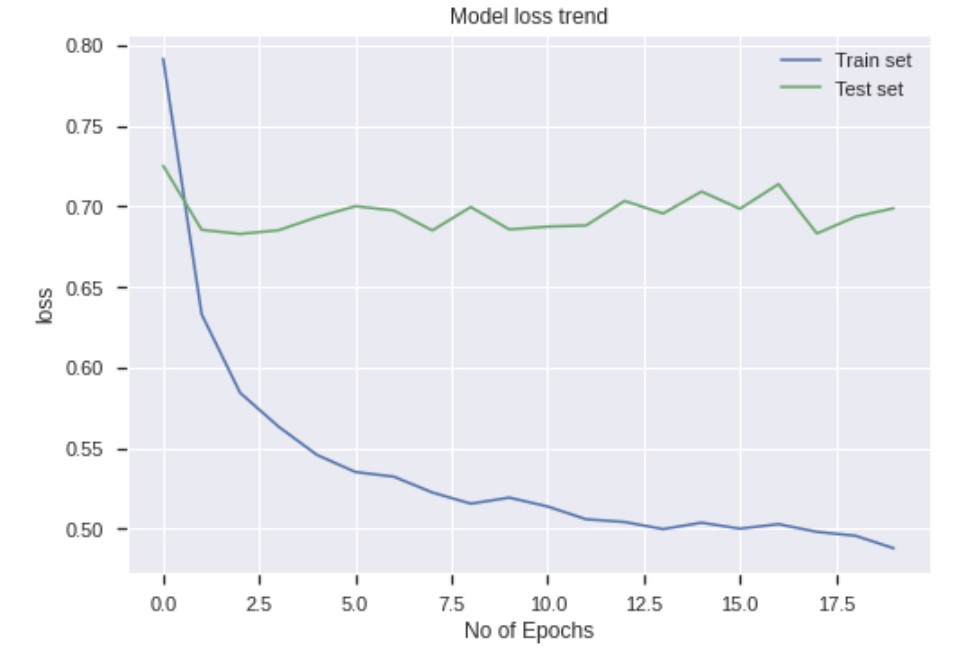
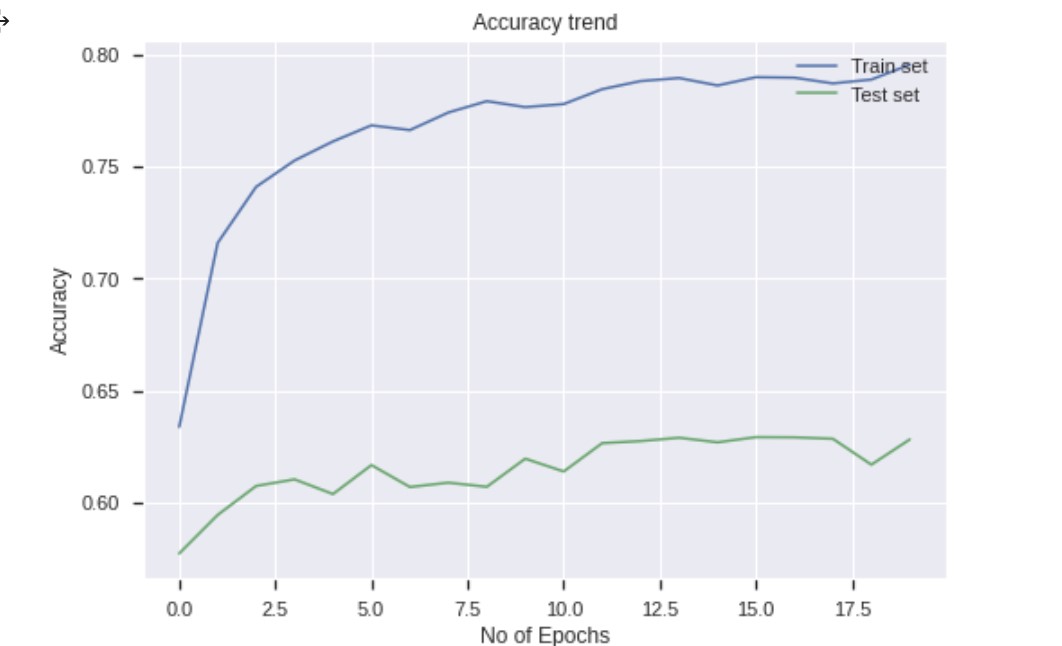
The input variables and the linear combination Z = W0 + W1X1 + W2X2 + ..... + WnXn is used to calculate the output values that are the predicted values. However, then the error and loss term are calculated and finally it is minimized.

Outcomes:

Summary of the model:



Bias-variance tradeoff: The loss of the model fluctuated a lot when number of epochs was 175 or higher that meant it was overfitting and it could have been reduced by reducing the number of epochs, below is the graph for accuracy and the loss that shows this trend.



After trying different values for the parameters that is epochs, batch size , the below results were obtained:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Batch Size** | **No of Epochs** | **Accuracy** | **Recall** | **Precision** |
| 512 | 100 | 0.58656 | 0.19896 | 0.83121 |
| 512 | 150 | 0.58219 | 0.17812 | 0.86223 |
| 256 | 100 | 0.58872 | 0.19244 | 0.86717 |
| 256 | 125 | 0.59251 | 0.21556 | 0.82764 |

**FINAL CONCLUSIONS:**

The dataset had to be cleaned and preprocessed a lot because of its categorical and missing values. The data was imbalanced because of which SMOTE function had to be used which was very helpful for the model for prediction. The total number of models implemented were three: binary class logistic regression, gaussian naive bayes where we assumed that the variables are following the gaussian distribution and neural networks. Out of all the three we found that for our dataset logistic regression model with tolerance and learning rate of 0.0001 performed really well with a good accuracy score as well as the precision and recall score,with them being close to each other too.

**References:**

1. https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospital s+for+years+1999-2008
2. Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo,

Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of

70,000 Clinical Database Patient Records,” BioMed Research

International, vol. 2014, Article ID 781670, 11 pages, 2014.