DEEP LEARNING

We will be using the UCL pima Diabetes dataset

The attributes of the dataset are as follows-:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

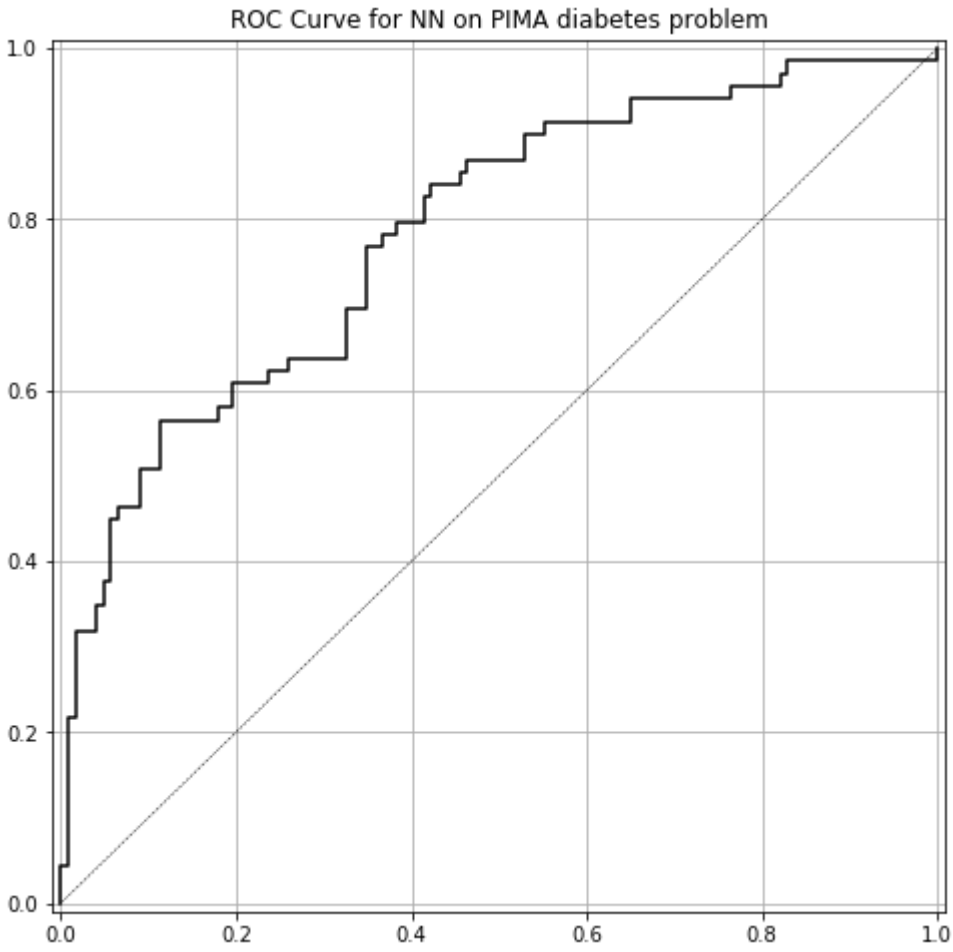
This dataset shows whethe a patient has diabetes or not. So we will be training our model to predict using the other attributes whether a person has diabtets or not.

From our initial data exploration we see that about 35% patients have diabetes and 65% do not. This means we can get an accuracy of 65% without any model - just declare that no one has diabetes. We will calculate the ROC-AUC score to evaluate performance of our model, and also look at the accuracy as well to see if we improved upon the 65% accuracy.

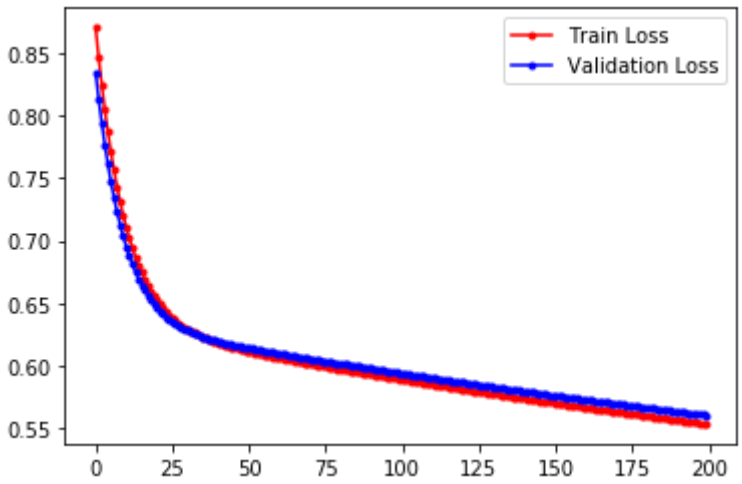
The data is first split into training and test sets with 25% of the data in test sets. The dataset is the normalised using Standard Scaler.

The First Deep Learning Model is going to be a single hidden layer network in which we have 12 nodes. The model will be compiled using Stochastic Gradient Descent with a learning rate of 0.03 and we will run it for 200 epochs. The test set is used as the validation dataset. We will have 121 trainable parameters in this model. After this model is ru for 200 epochs we get training loss of 0.5540 and test loss of 0.5608, whereas the accuracy for training set was 0.7135 and for test set was 0.7344.

The ROC curve for this Model is as follows-:



And the Loss for training and test set will be following along the epochs-:

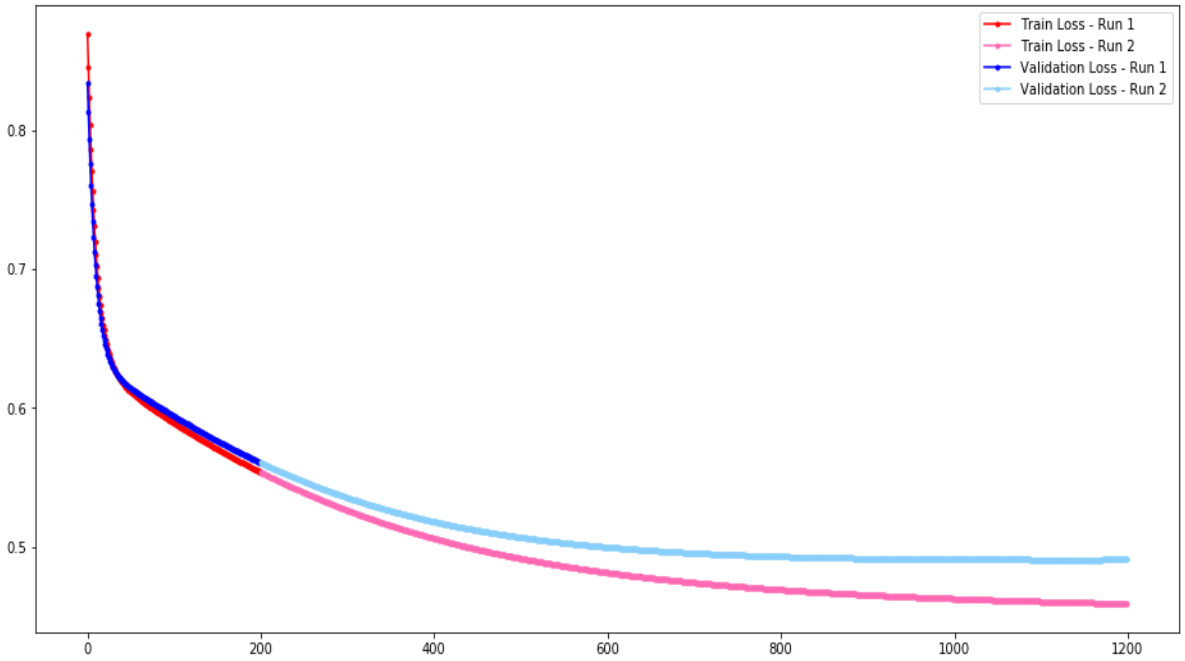


We see that the loss function is still going down for both the training and test sets.

So our next model will be to run it for more epochs-:

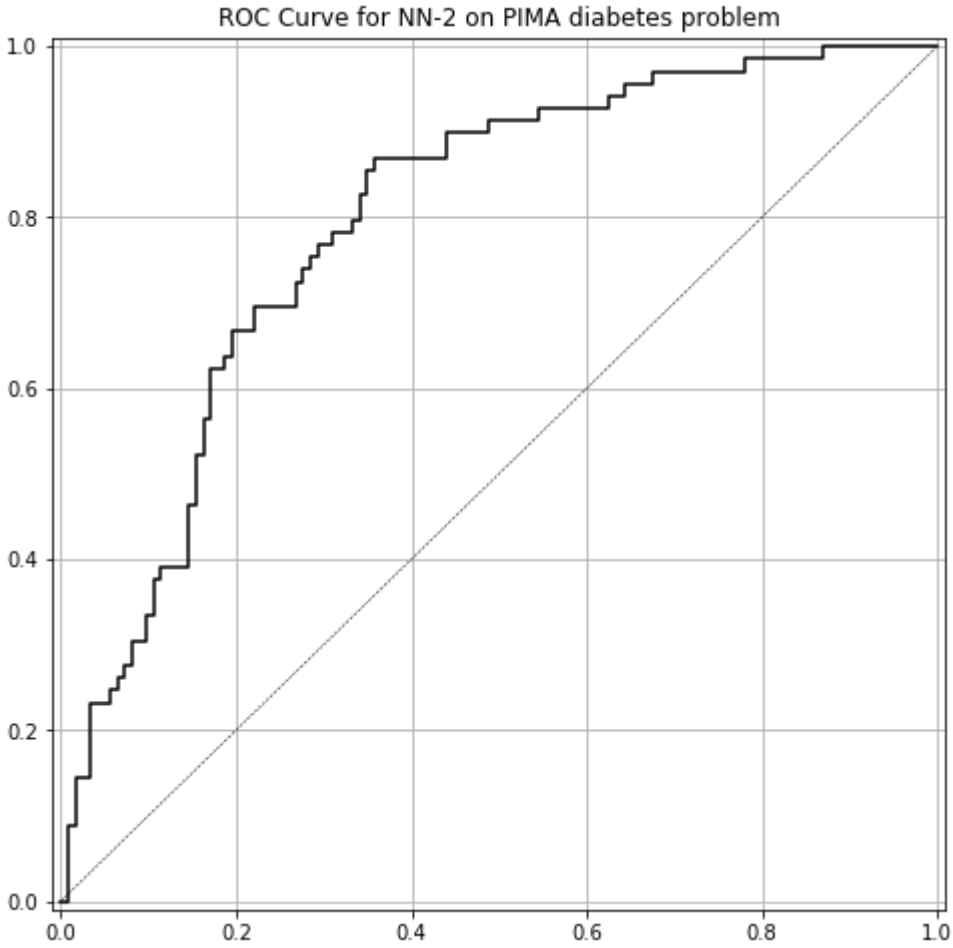
The Seocnd Deep Learning Model is going to be a single hidden layer network in which we have 12 nodes. The model will be compiled using Stochastic Gradient Descent with a learning rate of 0.03 and we will run it for 200 epochs. The test set is used as the validation dataset. We will have 121 trainable parameters in this model. After this model is run for 1200 epochs we get training loss of 0.4590 and test loss of 0.4908, whereas the accuracy for training set was 0.7734 and for test set was 0.7656.

Here is the loss function after running it for more epochs conitnuing from the previous model-:

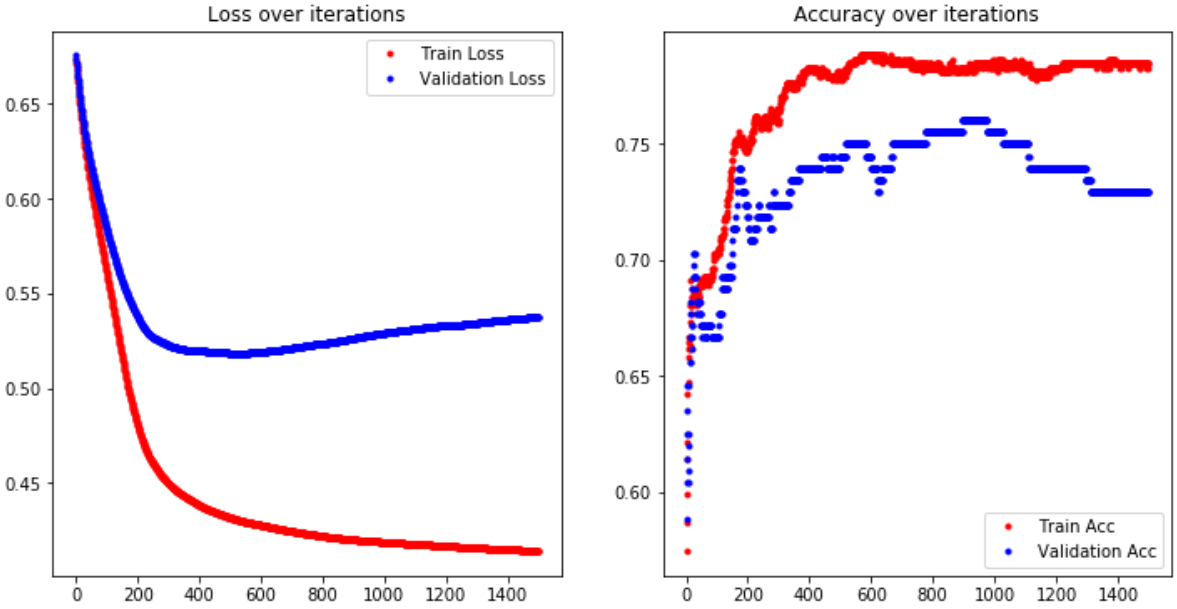


Our next and final model will have 2 hidden layers with 6 nodes each.We will be using relu as the activation function for hidden layers and sigmoid for the final layer and a learning rate of 0.003 we will run it for 1500 epochs. Stochastic Gradient Descent will be used as the optimizer.After running this model we get trainnig loss of 0.4142 and testing loss of 0.5376, whereas the accuracy for this mdoel for training data is 0.7847 and testing data is 0.7292.

The roc curve for this model is-:



Where the loss and accuracy for this model is as follows-:



We see that the best model for this dataset is model 2 with 1200 epochs with a testing loss of 0.4908.

The last model was performing goof with the training set but for the test set the loss never got below 0.5.

We can further analyse this data using more complex deep learning models and we should try to minimize the loss function more and imporve the accuracy.