CS-419
Course Project

Predicting diabetes in Pima Indian women

EDA

(Exploratory Data Analysis)

Libraries Used

Numpy

Pandas

Seaborn

Matplotlib

Scikit-learn

Random

Features Considered

Glucose
Blood pressure
Insulin
Skin thickness
BMI (Body Mass Index)
Diabetes Pedigree Function
Age
Number of pregnancies

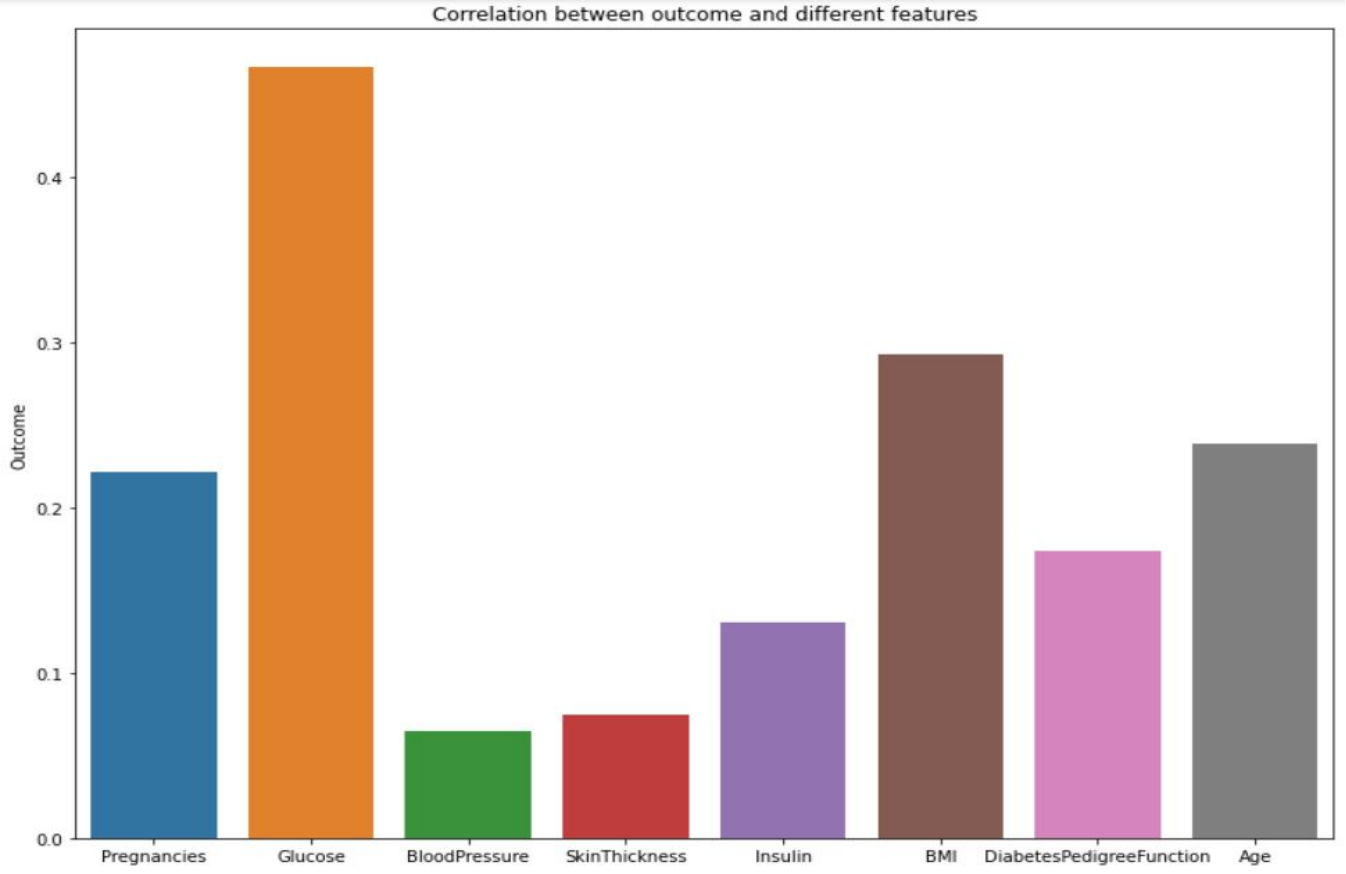
Statistics of the features & outcome

	mean	std	min	25%	50%	75%	max
Pregnancies	3.845052	3.369578	0.000000	1.000000	3.000000	6.000000	17.000000
Glucose	120.894531	31.972618	0.000000	99.000000	117.000000	140.250000	199.000000
BloodPressure	69.105469	19.355807	0.000000	62.000000	72.000000	80.000000	122.000000
SkinThickness	20.536458	15.952218	0.000000	0.000000	23.000000	32.000000	99.000000
Insulin	79.799479	115.244002	0.000000	0.000000	30.500000	127.250000	846.000000
BMI	31.992578	7.884160	0.000000	27.300000	32.000000	36.600000	67.100000
DiabetesPedigreeFunction	0.471876	0.331329	0.078000	0.243750	0.372500	0.626250	2.420000
Age	33.240885	11.760232	21.000000	24.000000	29.000000	41.000000	81.000000
Outcome	-0.302083	0.953903	-1.000000	-1.000000	-1.000000	1.000000	1.000000

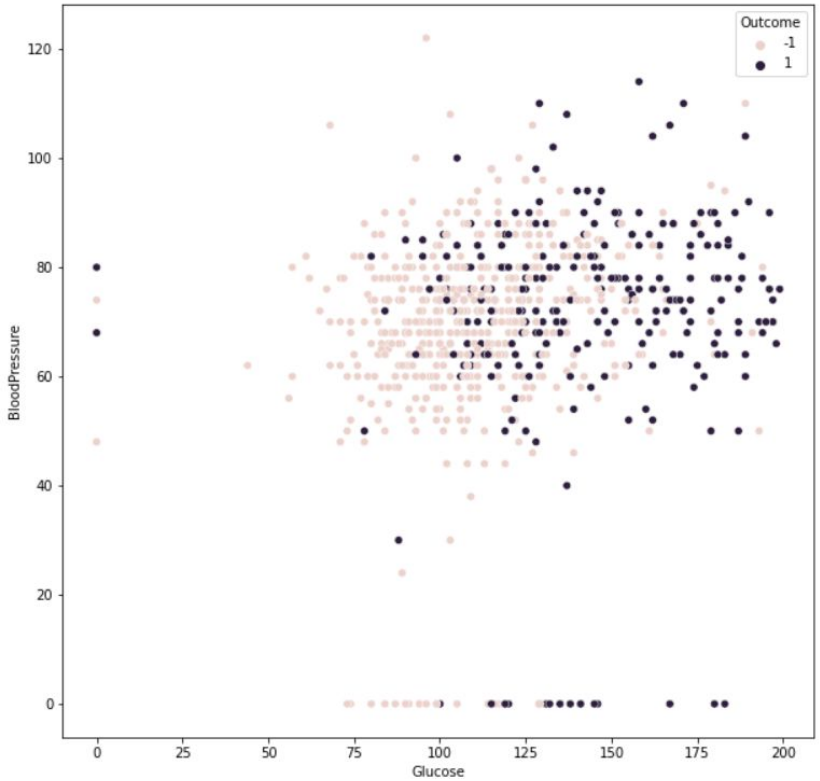
Correlation heatmap of features & outcome label

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

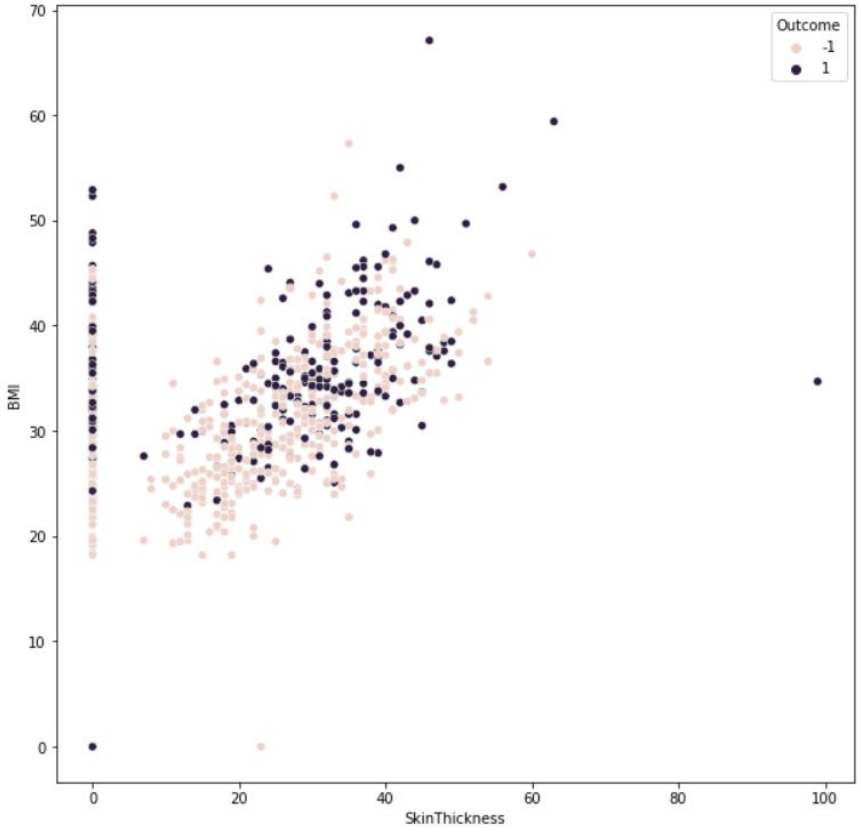
Correlation between outcome & features



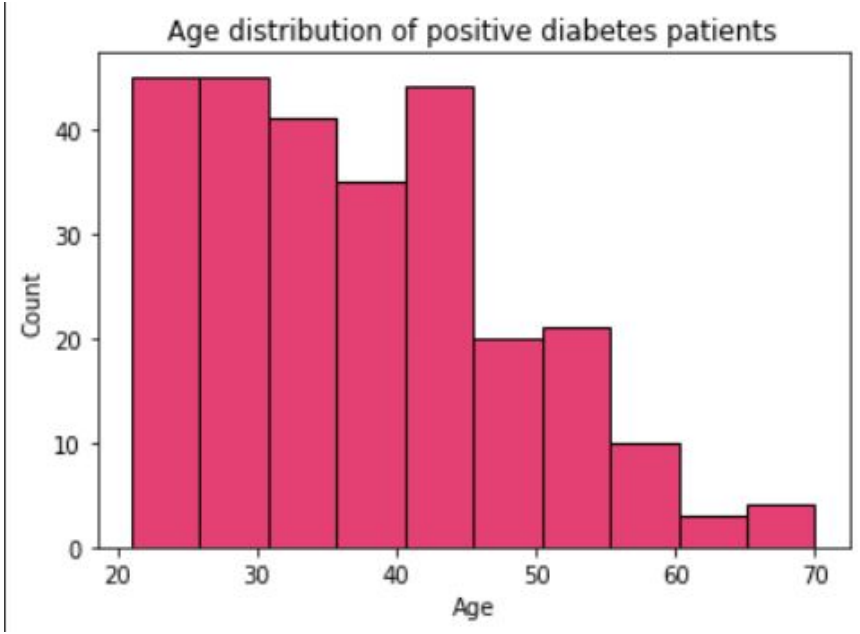
Scatter plots of features with outcome label

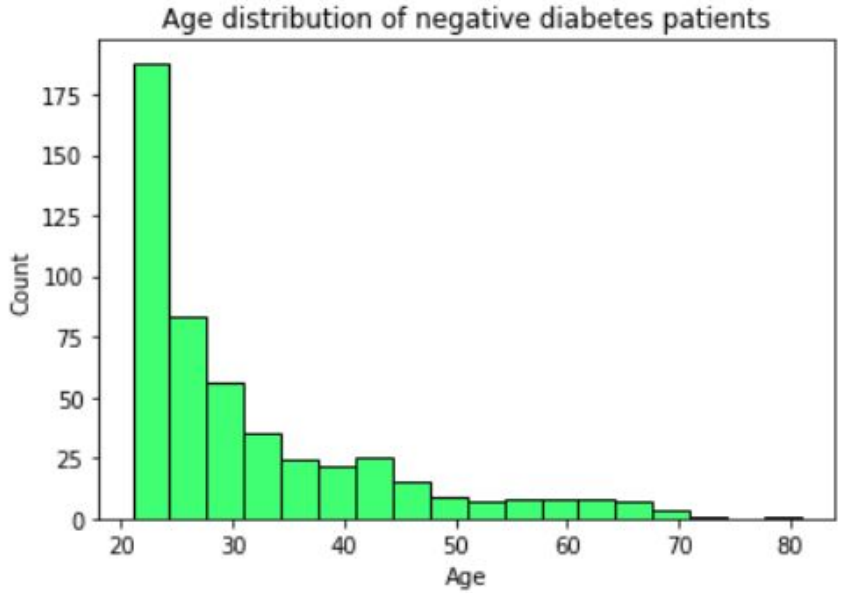


Scatter plots of features with outcome label



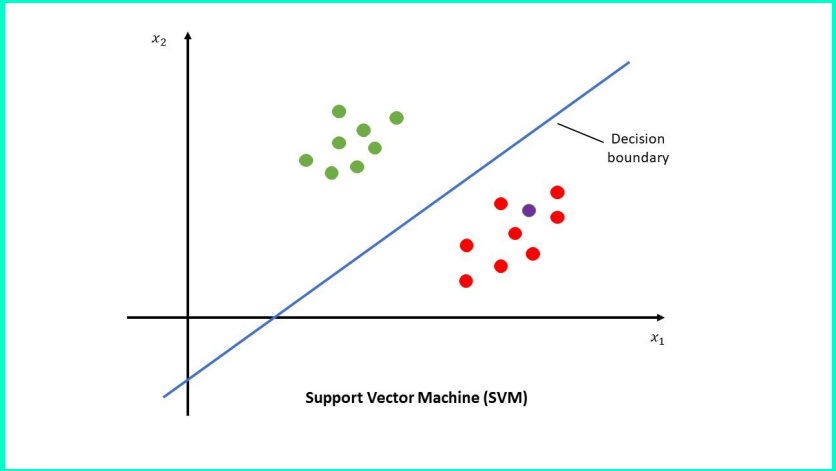
Age distributions of positive and negative patients





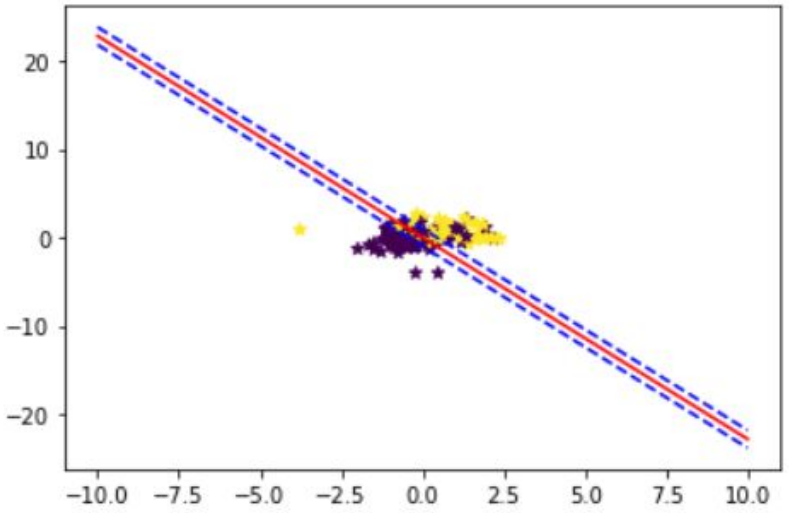
SVC

(Support Vector Classifier)



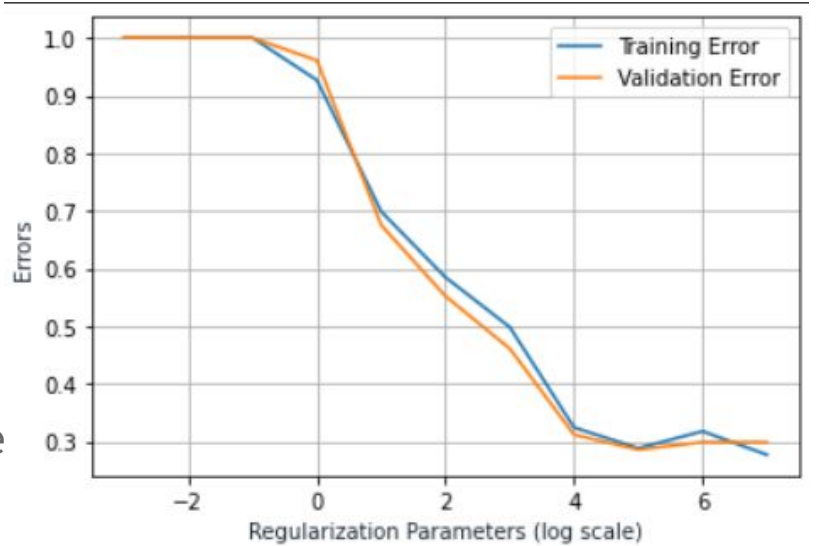
SVC

Support Vector Classifier is a machine learning algorithm that is used for classification problems where each data point is plotted in n-dimensional space and its value being the point's value, is then classified by finding a hyper-plane that differentiates the two classes very well. Since the SVC will be in n dimensions, we can only visually understand them in two dimensions. Therefore, the two most correlated features of the dataset namely Glucose and BMI are classified as shown.



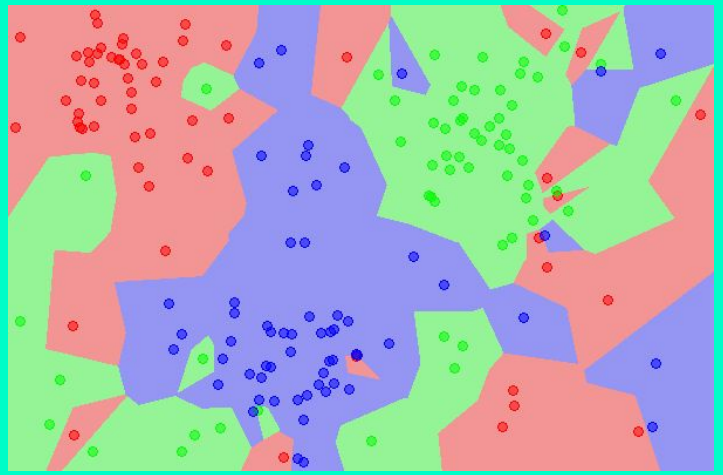
SVC

Furthermore, we check the training and validation errors for orders of regularization parameters and find that increasing the regularization parameter to higher orders decreases the error, which can be seen as shown.



KNN

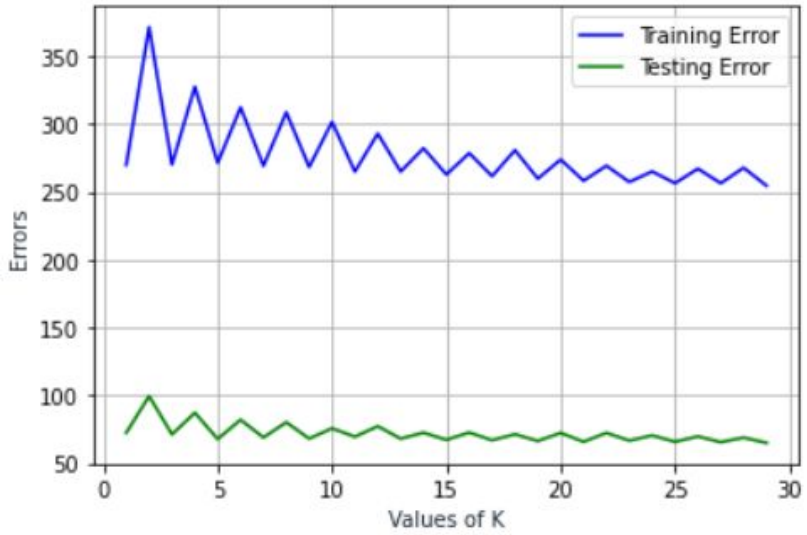
(K-Nearest Neighbours)



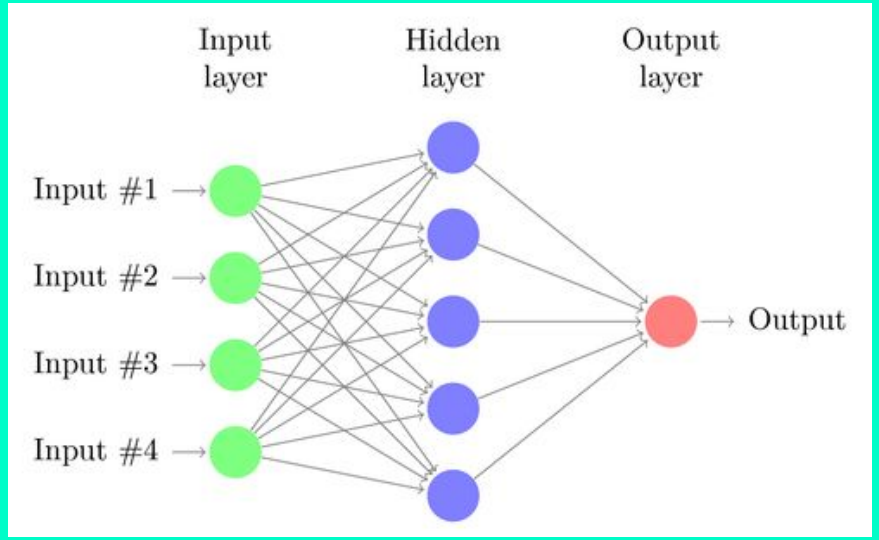
KNN

The K-Nearest Neighbours is a supervised an algorithm which groups together data points which are close as one class. One shortcoming of this algorithm is that this slows down with increasing data size.

Furthermore, upon training and parameter hyper tuning we find that the best value of k is 19 and the validation error from it is 66.4285. Since, this is very high K-Nearest Neighbours is certainly not a good model for this dataset.



Neural Network

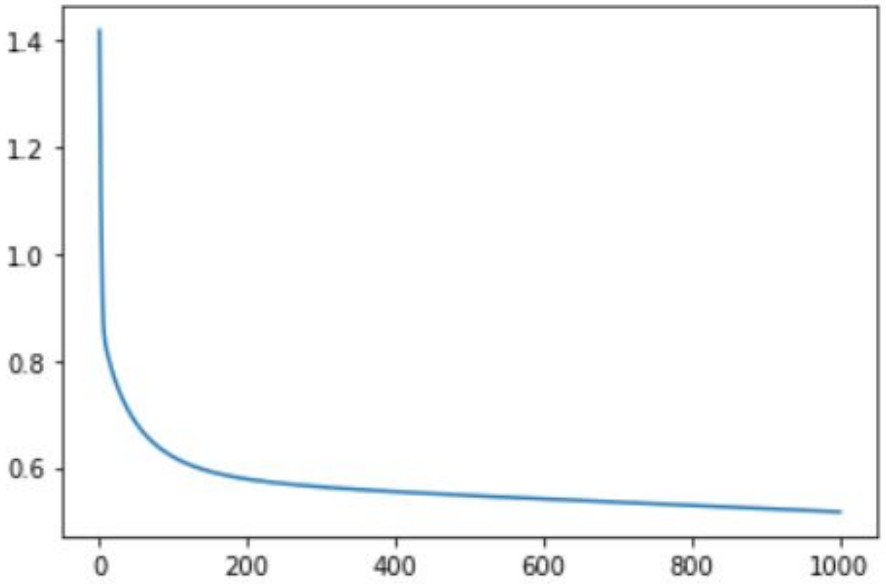


Neural Network

A Neural Network is used to recognize underlying relations in a set of data through a process that mimics the human brain. Thus, we use a 2 hidden layers of size 4 and 2 respectively in the neural network to classify the problem we have before us. The initial layer is of 8 neurons(number of features) and the output layer is of 1 neuron(number of outputs). The activation function used for the layers is ReLU. The final activation layer used for the output of the classification model is the sigmoid function. Finally, for the loss function we have used the Binary Cross Entropy Loss function.

Neural Network

After running the training dataset on the Neural Network for 1000 epochs with a learning rate of 0.0001. We get decreasing Binary Cross Entropy error terminating to about 0.5167. This is a respectable error and can not be decreased significantly further as the training error vs epoch graph is tending to constant value. Therefore, for a better model we might have to change the hidden layers.



Conclusion

Through this project we have seen the positive impact that Machine Learning Models can have in the medical field. Not only can we use common identifiers to diagnose diabetes but we could also potentially use CNNs to used X-rays and CT-scans to diagnose other diseases, saving countless lives. In the future we could also rely on Machine Learning models to predict the structure and composition of the drugs involved in treating these diseases. Nevertheless there is still a lot of development that need to take place before we can employ these models in the field.