Comparing Accuracy of various Machine Learning models on Pima Indians Diabetes Dataset

Keshav Patel Keval (200110055) Anmol Saraf (200070007) Prateek Garg (20D070060)

Vadapalli Arvind Narasimha (200070087) Shrey Ganatra (20D070074)

All the team members are from the Department of Electrical Engineering, IIT Bombay

***Abstract*—We compare the accuracy of the following three machine learning models: Serial Vector Classifier, k-Nearest Neighbours, Artificial Neural Network on Pima Indians Diabetes Database.**

I. Introduction

Diabetes mellitus, commonly known as diabetes, is a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time. The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014 and it is estimated by 2040, the world's diabetic patients will reach 642 million, with a disproportionate fraction of the increase coming from low to middle income countries. Diabetes can cause blindness, kidney failure, heart attacks, stroke and lower limb amputation. Diabetes can be treated and its consequences avoided or delayed with diet, physical activity, medication and regular screening for he disease. This makes predicting Diabetes in humans one of the most important problems in 21st century healthcare. Many factors can be used to predict the onset of diabetes in humans, one of them being the genetic factor. The Diabetes pedigree function gives us a measure of the hereditary risk that one might have with the onset of Diabetes. However, apart from the genetic influence other factors like number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, BMI, and age are also important indicators of Diabetes, we plan to take all these factors into account in our Machine Learning Models.

II. Background and Prior Work

With the onset of Machine Learning quite a lot if work has been done in predicting diabetes in humans. The use of various models like Decision Tree Classifier, Random Forest Classifier, Artificial Neural Networks among others has made the diabetes predictions more accurate than before. In this project we compare the accuracy of the three models that we apriori feel are the most appropriate and accurate.

III. Data and Methodology

The project uses factors like number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, BMI age as well as the Diabetes Pedigree function to predict diabetes in patients. The output of each of the three Machine Learning Models is a simple +1 or -1, with +1 for a positive prediction of Diabetes in the patient. We test the accuracy of the Models using F1 score and by producing the Confusion Matrix. The dataset which was used for this project was taken from kaggle and the link is given below:

- Pima Indians Diabetes Database: Predict the onset of diabetes based on diagnostic measures https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

The above dataset had 9 columns including the outcome. None of the columns had any missing data. We performed some Data Analysis to get some insight into the data before we begin training the Machine Learning Models.
The correlation of the various features with the outcome is given in the bar chart below:

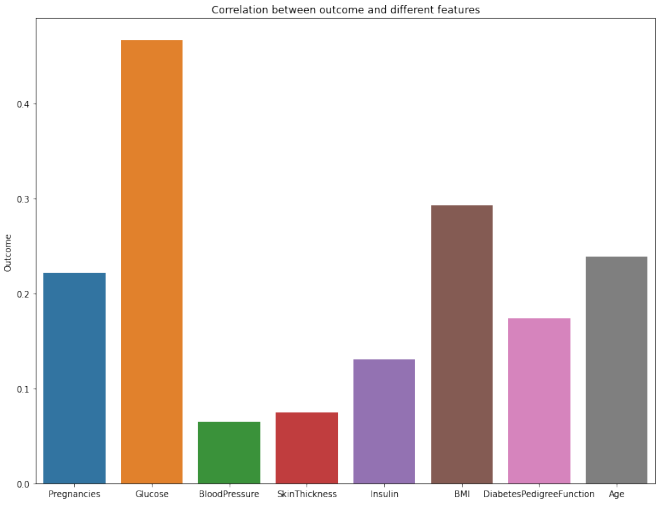

Fig. 1. Correlation of the Features with the outcome

We made a scatter plot between the two features that had the highest correlation with the outcome:Glucose and BMI.(Fig. 2.)

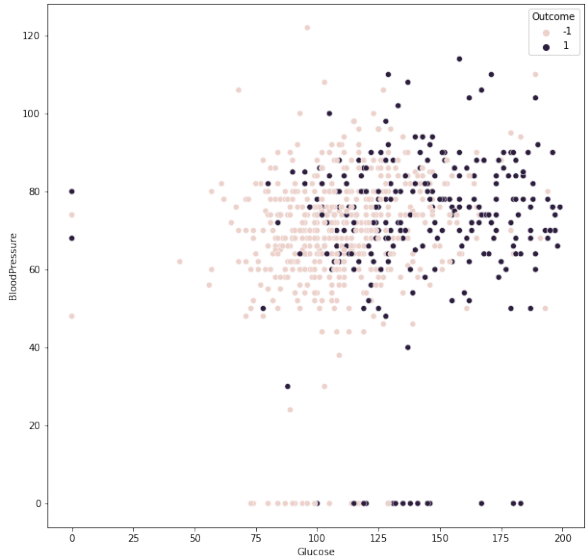

Fig. 2. Scatter plot between Glucose and BMI

We also made a scatter plot between Skin thickness and BMI.(Fig. 3.).

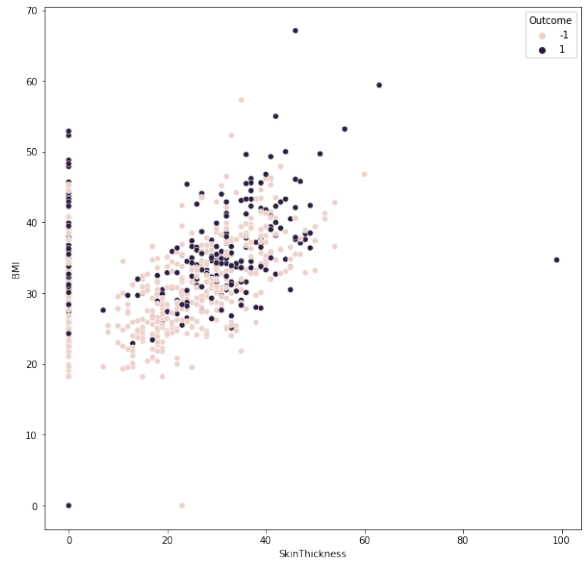

Fig. 3. Total Affected over the years for different disaster subgroup

The last part of our Data Analysis was to plot the distribution of positively and negatively diagnosed Diabetes patients in various age brackets.The results of these are given in Fig. 4 and Fig. 5.

This revealed that most of the positively diagnosed Diabetes patients in our dataset are in the 20 to 40 age group.This is in agreement with the fact that most diabetes patients see the onset of diabetes by the age of 45.(American Diabetes Association).

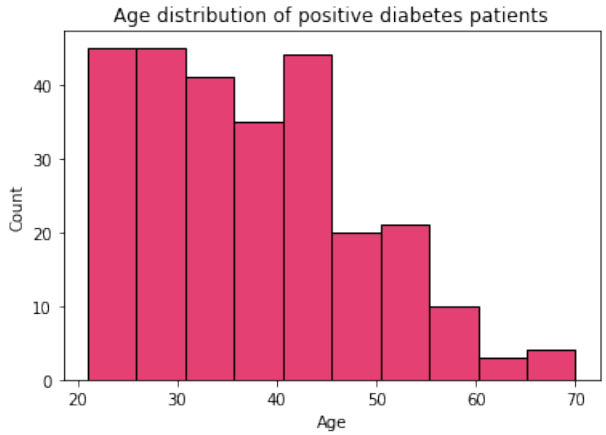

Fig. 4. Total Affected over the years for different disaster subgroup

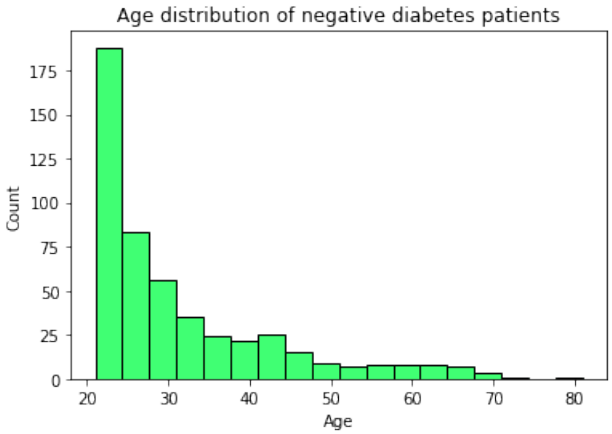

Fig. 5. Total Affected over the years for different disaster subgroup

IV. Experiments and Results

A. Support Vector Classifier

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 below,

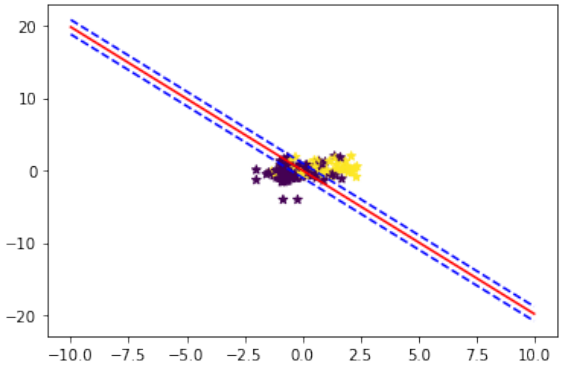

Fig. 6. Validation error vs Epoch OR Validation error vs learning rate graph for SVC

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 below,

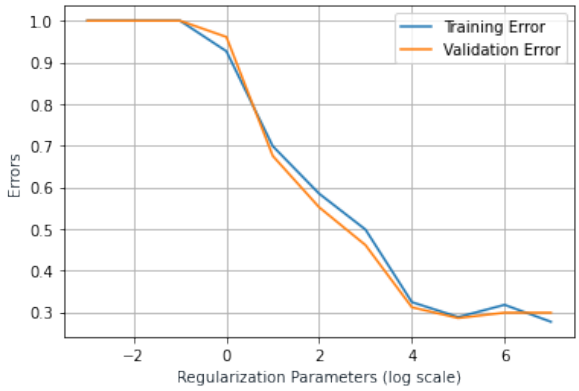

Fig. 7. Regularization Parameter vs error

B. K- Nearest Neighbours

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.

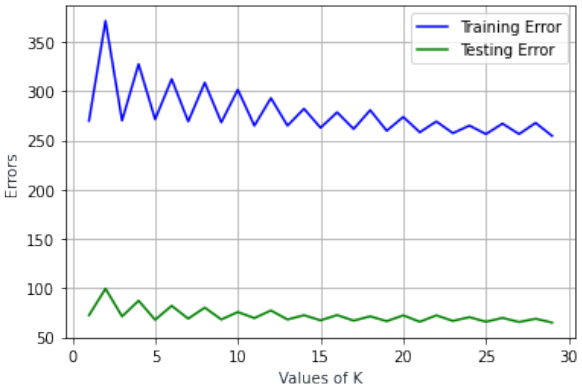

Fig. 8. Train Test analysis for different k

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.

C. Artificial 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(x) = $\max(0, x)$. The final activation layer used for the output of the classification model is the sigmoid function, $S(x) = \frac{1}{1+e^{-x}}$. Finally, for the loss function we have used the Binary Cross Entropy Loss function,

$$-\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i))$$

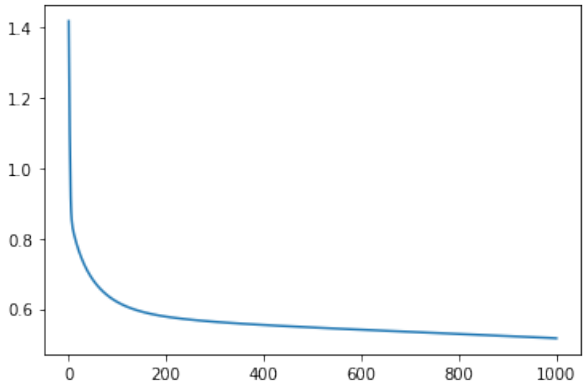

Fig. 9. training error vs Epoch

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.

V. Learning, Conclusions and Future Works

Conclusion of the project would be that the best machine learning model for predicting diabetes in Pima Indians is the Support Vector Classifier, which had trainging error as low as 0.2785 and validation error as low as 0.3571, while the second best model was the Neural Network with 4,2 neurons long hidden layers. The worst model was of K-Nearest Neighbours which was giving very high error of around 66.

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.

Contribution of Team Members

Reading the data and the Exploratory Data Analysis was done by Vadapalli Arvind Narasimha. The Serial Vector classifier was implemented by Shrey Ganatra. The k-Nearest Neighbours classifier was implemented by Prateek Garg. The Artificial Neural Network was implemented Anmol Saraf. The conclusion as well as this report were done by Keshav Patel Keval.

Acknowledgements

We would like to thank our professor Prof. Abir De for teaching us this course, without which we would not have been able to complete this project. We would also like to thank all the TA's involved with this course, who helped us a lot in doing the various assignments, and completing this course.

References

1) https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
2) https://www.who.int/news-room/fact-sheets/detail/diabetes
3) https://www.medicalnewstoday.com/articles/317375#average-age-of-onset-for-diabetes