AN END-TO-END DIABETES PREDICTION MODEL

Priyadarshi Chatterjee

Introduction:-

Diabetes is a chronic disease that affects a large population approx. 425 million people worldwide each year and can lead to severe complications if not diagnosed and managed in a timely manner. With the rise of machine learning techniques, developing predictive models for diabetes diagnosis has become increasingly popular. In this project, we have built an end-to-end diabetes prediction model using the Prima India diabetes dataset, which consists of various medical and lifestyle features. Our model uses both logistic regression and neural network models, including Single-Layer Perceptrons (SLPs) and Multi-Layer Perceptrons (MLPs), for training and predicting. The goal of this project is to develop a predictive model that accurately predicts the likelihood of an individual developing diabetes based on their medical history and lifestyle factors. We employed feature engineering techniques to pre-process and transform the data to improve model performance, and evaluated the model's performance using standard metrics such as accuracy, precision, recall, and F1-score. The results obtained from this project can provide insights into the predictive power of machine learning models for diabetes diagnosis and assist healthcare professionals in making more accurate predictions and improving patient outcomes.

Application Area: -

The diabetes prediction model developed in this project has a wide range of potential applications in healthcare.

1. **Preventive care**: The model can be used to identify individuals who are at high risk of developing diabetes. This can lead to timely interventions such as lifestyle modifications, medications, and regular monitoring, which can prevent or delay the onset of diabetes and its complications. By identifying individuals at high risk, healthcare providers can tailor their interventions and improve their chances of success. This can ultimately lead to better health outcomes and reduced healthcare costs.
2. **Clinical decision-making**: Healthcare providers can use the model's predictions to guide their clinical decisions. For example, if the model predicts a high likelihood of an individual developing diabetes, healthcare providers can recommend appropriate diagnostic tests, treatments, and follow-up care. By using the model's predictions, healthcare providers can make more informed decisions, leading to better patient outcomes.
3. **Population health management**: The model can be integrated into health information systems to facilitate population health management. By identifying high-risk individuals, healthcare providers can prioritize preventive interventions and allocate resources more efficiently. Additionally, the model can be used to monitor the effectiveness of interventions and assess the impact of public health campaigns aimed at reducing the prevalence of diabetes. By using the model to manage

population health, healthcare providers can improve the health of entire populations and reduce healthcare costs.

Machine Learning Algorithms Used:-

1. Logistic Regression: -

- The model uses logistic regression, which is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome.
- The model is trained on the Prima Indian Diabetes dataset, which contains various features such as glucose level, blood pressure, and BMI, that are used to predict the presence of diabetes.
- The model uses GridSearchCV and 5-fold cross-validation to perform hyperparameter tuning, which helps to optimize the model's performance by selecting the best values for the hyperparameters.
- The model achieves an accuracy of 70% on the testing set, which indicates that it correctly predicts the presence or absence of diabetes for 70% of the samples in the testing set.
- In addition to accuracy, other evaluation metrics such as precision, recall, and F1-score can be used to assess the model's performance. These metrics can help to identify areas where the model may be performing poorly, such as incorrectly predicting diabetes for individuals who do not have the disease.

2. SLP (Single Layer Perceptron)

- SLP is a simple neural network architecture that consists of only one layer with a single neuron. It can learn linear decision boundaries between the input features and the output classes, making it a computationally efficient and interpretable model.
- SLP uses the sigmoid activation function to map the input values to a probability value between 0 and 1, representing the likelihood of a patient having diabetes.
- SLP is well-suited for the Pima Indian Diabetes dataset due to its small number of input features and binary output. The model can learn the linear decision boundaries and make accurate predictions.
- By hyperparameter tuning using Grid Search CV and 5-fold CV the SLP model can be optimized to achieve the best performance on the Pima Indian Diabetes dataset. The accuracy of the SLP model can be further evaluated using additional evaluation metrics such as precision, recall, and F1 score.

- The SLP model was trained on the Pima Indian Diabetes dataset, and hyperparameter tuning was performed using Grid Search CV to optimize the model's hyperparameters. **The optimal set of hyperparameters was identified, which resulted in a test accuracy of 74%.**
- The SLP model is an effective and computationally efficient model for the Pima Indian Diabetes dataset. However, it is important to note that other advanced neural network architectures such as MLP or CNN can also be used to improve the model's performance on the dataset.
- Additionally, it is essential to evaluate the model's performance using other evaluation metrics such as precision, recall, and F1 score to gain a better understanding of the model's performance.

3. MLP (Multi Layer Perceptron)

- The Pima Indian Diabetes dataset is a binary classification problem, and MLP is a popular neural network architecture that is widely used for such problems.
- In contrast to the Single Layer Perceptron (SLP), MLP contains multiple layers of neurons, making it capable of learning complex decision boundaries between the input features and the output classes.
- For the Pima Indian Diabetes dataset, an MLP model was trained with three layers. The input layer has 8 neurons, one for each input feature, and two hidden layers with 12 and 8 neurons, respectively. The output layer contains one neuron that uses the sigmoid activation function to predict the likelihood of a patient having diabetes.
- Hyperparameter tuning using Grid Search CV was performed to optimize the model's hyperparameters, such as learning rate and batch size. **The optimal set of hyperparameters was identified, which resulted in a test accuracy of 75%.**
- Overall, MLP is a powerful neural network architecture that can learn complex decision boundaries for the Pima Indian Diabetes dataset. By performing hyperparameter tuning using Grid Search CV, the model's hyperparameters can be optimized to achieve the best performance on the dataset
- Additionally, it is essential to evaluate the model's performance using other evaluation metrics such as precision, recall, and F1 score to gain a better understanding of the model's performance.

Data Description: -

- The Pima Indian Diabetes dataset is a popular dataset used for binary classification tasks, and it contains information on female patients of Pima Indian heritage who are at least 21 years old.
- The dataset consists of 768 observations and 8 input features, including age, BMI, blood pressure, skin thickness, insulin level, pregnancy history, glucose level, and diabetes pedigree function.
- There are no missing data in the dataset and the correlation of the target variable (outcome) with the input parameters is given below:-

Outcome	1.000000
Glucose	0.466581
BMI	0.292695
Age	0.238356
Pregnancies	0.221898
DiabetesPedigreeFunction	0.173844
Insulin	0.130548
SkinThickness	0.074752
BloodPressure	0.065068

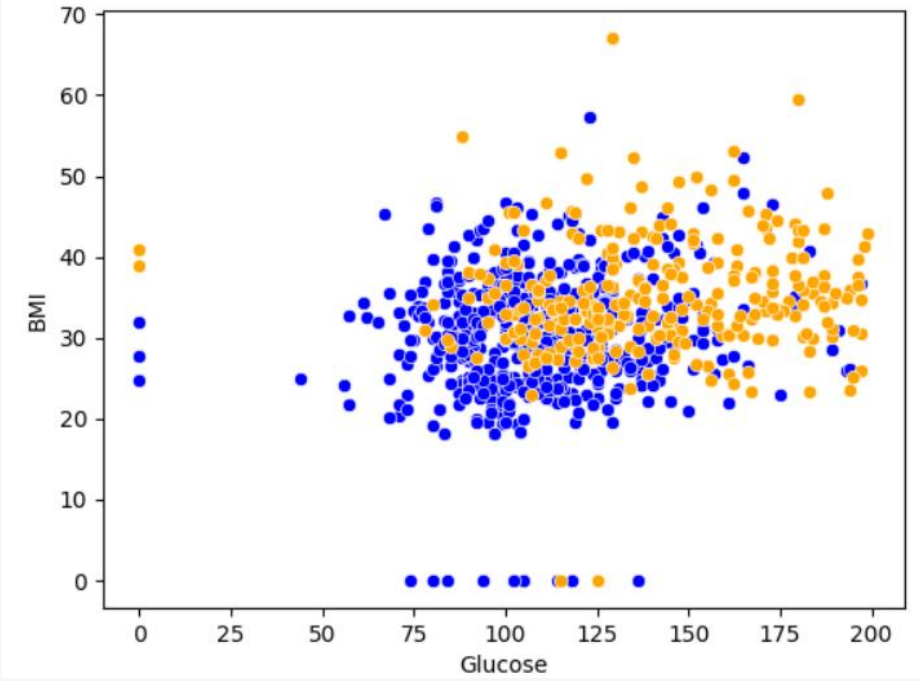
- Here we can observe that the factors leading to diabetes in order of decreasing importance are Glucose>BMI>Age>Pregnancies>DiabetesPedigreeFunction>Insulin

- The mean age of non-diabetic patients is about 30 years and the mean age of diabetic patient is 35 years, which shows that's age is not a strong determinant when compared to Glucose and BMI
- We have also computed the mean Glucose level for each range of ages which would help to find the decision boundary for each age ranges

Age Range	
20-29	114.175060
30-39	126.178344
40-49	124.884956
50-59	141.148148
60-69	137.560000
70-79	119.000000

- The relation between the two most important parameters in the dataset Glucose and BMI is given below:-

```
<Axes: xlabel='Glucose', ylabel='BMI'>
```



Results:-

- When using Logistic Regression to train the model we got the below results for the performance metrics: -

```
Accuracy: 0.7142857142857143
Precision: 0.6086956521739131
Recall: 0.5185185185185185
F1-score: 0.5599999999999999
```

- When hyperparameter tuned using GridSearchCV the measures remained same with no improvement,
- Then once implementing the SLP model by using Binary Cross-entropy as the loss function and "Adam" as the optimizer; without and with GridSearchCV we got the below results

```
5/5 [==============================] - 0s 3ms/step - loss: 0.4946 - accuracy: 0.7143
Accuracy: 71.43
```

```
5/5 [==============================] - 0s 3ms/step - loss: 0.5056 - accuracy: 0.7403
Accuracy: 74.03
```

Therefore using SLP with GridSearchCV we got an accuracy measure of 74 percent on the testing data

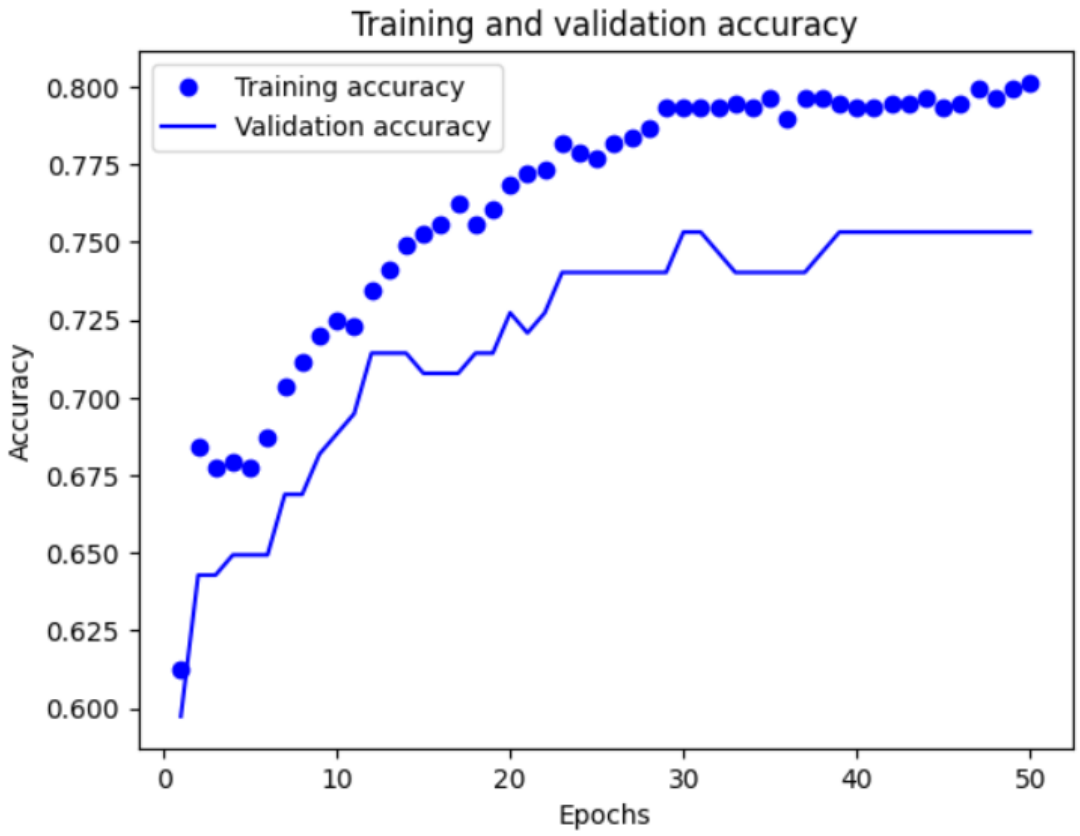
- Then finally implementing MLP model which was trained with three layers. The input layer has 8 neurons, one for each input feature, and two hidden layers with 12 and 8 neurons, respectively. The output layer contains one neuron that uses the sigmoid activation function to predict the likelihood of a patient having diabetes, we got the following result.

```
5/5 [==============================] - 0s 3ms/step - loss: 0.4852 - accuracy: 0.7532
Test loss: 0.4851553738117218
Test accuracy: 0.7532467246055603
```

This is the best result we got i.e we got improved testing accuracy score and also reduced test loss. Therefore, we will be using this MLP with GridSearchCV model for our further analysis.

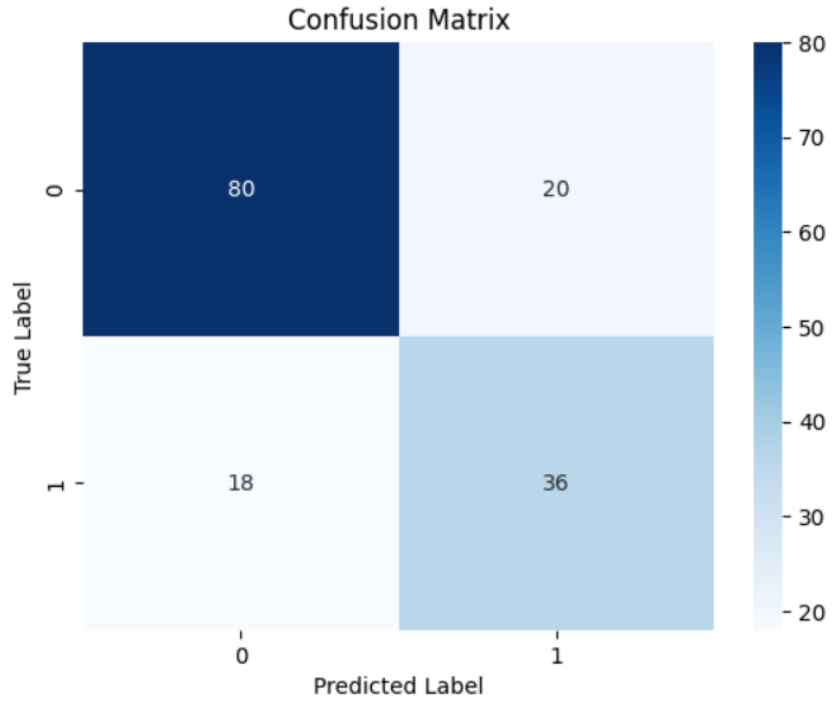
Analysis:-

- When the plot of training and validation model accuracy is plotted over the epochs we get the below plot



We find that the accuracy measure flattens at around 45-50 epochs, that's when we are getting the maximum accuracy

- The confusion matrix of the given model looks like this:-



- The class wise classification report for the model is given below: -

```
5/5 [==============================] - 0s 3ms/step
              precision    recall  f1-score   support

           0       0.82      0.80      0.81       100
           1       0.64      0.67      0.65        54

    accuracy                           0.75       154
   macro avg       0.73      0.73      0.73       154
weighted avg       0.76      0.75      0.75       154
```

For the model using MLP with GridSearchCV, we are getting recall score of Class 0 as 0.80 that is, it is predicting 80 percent of the non-diabetic cases correctly whereas for class 1 ie diabetic cases it having 67 percent recall score.

As a model deployed in medical diagnosis, the problem of False Negatives is more important than the other ones. So, we must focus on increasing the recall score for class 1