**PGP AI - Building a model to predict Diabetes**

DESCRIPTION

**Problem:**

The given dataset lists the glucose level readings of several pregnant women taken either during a survey examination or routine medical care. It specifies if the 2-hour post-load plasma glucose was at least 200 mg/dl. Analyze the dataset to:

1. Find the features of the dataset,
2. Find the response label of the dataset,
3. Create a model  to predict the diabetes outcome,
4. Use training and testing datasets to train the model, and
5. Check the accuracy of the model.

Graphical user interface, text

Description automatically generated

The features are linked with the dataset.

Graphical user interface, application

Description automatically generated with medium confidence

Here the independent variables are pregnant, glucose, bp, skin, insulin, bmi, pedigree and age. The response variable is the “label”. This denotes whether a person has diabetes or no. Label value of 1 indicates – tested positive for diabetes. Based on the independent variables, we need to predict the outcome. So, this falls under classification problem. We will be using Logistic Regression for this problem.

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

First, split the data into train and test data.

Graphical user interface, text, application

Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

The accuracy of the model is as follows:

Graphical user interface, text, application

Description automatically generated

The model does a decent job of predicting diabetes, though there is some error. The below confusion matrix for the test data. It summarizes the actual and predicted values. It tells us about True Positives, True Negatives, False Positives and False Negatives.

Graphical user interface, text, application

Description automatically generated

Chart, treemap chart

Description automatically generated

The model has predicted 109 + 40 = 149 values correctly. That is the actual and predicted values are the same. However, the model has predicted 29 + 14 = 43 values wrongly. The total test data are 192. The accuracy is (149 / 192) \* 100 which is approximately 77 %