**IMPLEMENTATION OF THE PROJECT:**

There are 6 steps involved in implementing this project. They are as follows:

1. **Data Description**

**The first step involves importing all the libraries required**

**Text, letter

Description automatically generated**

**Data description helps us understand the dataset.**

**Table

Description automatically generated**

**The above code reads the dataset which is a csv file and talks about its size.**

**Table

Description automatically generated**

**This shows any 5 random records from the dataset**

**From the above tables, it’s observed that there are certain cells of Insulin, Glucose, Blood Pressure, BMI, and Skin Thickness with 0 value in them.**

**So, it’s replaced with the mean values of each column instead.**

**Table

Description automatically generated with low confidence**

1. **Data Exploration**

Data exploration gives us insights about the various parameters in the dataset.

Correlation of the features and label can be visualized by using a heatmap. Here Outcome is the label, and the other parameters are the features

Table

Description automatically generated

Chart

Description automatically generated

From the heatmap, it is understood that Glucose, Age, BMI, and number of pregnancies have a greater correlation with Outcome.

1. **Data Preparation**

**The data is split into training data, testing data, and checking data (for verifying our model)**

**The dataframes are then split into labels and features, followed by converting them into Numpy arrays as that is the required format for the machine learning algorithm.**

Graphical user interface, text, application

Description automatically generated

**Normalization is the last step before applying the machine learning algorithm. Z score normalization was performed so that each variable has mean 0 and standard deviation of 1**

1. **Training and Evaluating the Machine Learning**

Logistic regression, K Nearest Neighbors, Support Vector Machine, Naive Bayes, Random Forest, and Decision Tree algorithms have been used by importing them from scikit learn library.

**Graphical user interface, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Logistic Regression and Support Vector Machine algorithms have the highest accuracy.**

1. **Saving the Model**

**Logistic Regression model is saved since it has the highest accuracy. Joblib present in scikit learn can be used for this. The file is saved as diabetesModel-joblib with .pkl extension. By importing the model and checking the accuracy we can verify if the model has been saved properly**

**Application

Description automatically generated with low confidence**

1. **Making Predictions with the Model**

**Predictions have been made by using data from the checkDF dataframe and also by putting random values.**

**Table

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**