## Introduction

This report will examine the “PimaIndiansDiabetes” dataset, which records a series of characteristics of patients with or without diabetes. The report will outline the main characteristics of this dataset and its findings. The main aim of the dataset is to determine the predictability of a subject having diabetes, based on other characteristics such as blood pressure, body mass index etc. Analysis of the dataset was carried out in R, and associated graphs were also created in R. All R code used can be viewed in the attached file.

## Methods

The variables observed in this analysis included the following:

Number of times pregnant, plasma glucose concentration at 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function, age (years) and finally actual diagnosis of diabetes

The first 8 variables were all numerically valued depending on the results. The diagnosis of diabetes was assigned a value of either 1 for positive or 0 for negative to allow for ease of interpretation and graphing.

Some information was missing from the data set and given a value of 0, and hence this data was reassigned median values in order to standardise the analysis.

The GLM model was then fit using the binomial family, and predicted probabilities were calculated. A data frame was created of the predicted outcomes and a new column was added to the data frame to hold predicted outcomes (P for positive, or N for negative) based on the threshold of 0.5.

![Chart, line chart

Description automatically generated]()

![Chart, scatter chart

Description automatically generated]()

![Chart

Description automatically generated]()