# 1. Diabetes Prediction - Naive Bayes

November 28, 2022

## 1 Predict whether or not a patient has diabetes, based on certain diagnostic measurements

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- The datasets consists of several medical predictor variables and one target variable, Outcome.
- Predictor variables includes:
  - the number of pregnancies the patient has had,
  - their BMI,
  - insulin level,
  - age, and so on.

## 2 Import the required packages

[ ]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
```

## 3 Load and preprocess the data

[ ]:
```
df = pd.read_csv('../data/prima-indians-diabetes.csv')
```

[ ]:
```
df
```

## 4 Seperate features and label columns

[ ]:
```
X = df.drop("outcome",axis=1)
Y = df.outcome
```

## 5 Train Test Split

[ ]:
```
x_train, x_test, y_train, y_test = train_test_split(X, Y,
                                                    test_size=0.2,
                                                    random_state=5)
```

## 6 Train a classifer

- Build diabetes prediction model

[ ]:
```
# Define Model (create object of the class GaussianNB)
clf = GaussianNB()
```

[ ]:
```
# Train model using fit method
clf.fit(x_train, y_train)
```

## 7 Make predictions using test samples

[ ]:
```
# Make predictions on test data
predictions = clf.predict(x_test)
```

## 8 Evaluate the model

[ ]:
```
accuracy_score(predictions, y_test)
```