# Introduction

* The aim of the project is to implement the learnings of the course on a dataset. The dataset was taken from an online site which you can find here: [OpenML diabetes](https://www.openml.org/d/37). The project focuses on following areas of coding in data science:

1. Preprocessing the dataset
2. Checking the correlation of the features
3. Splitting the dataset into training and testing data
4. Implementing the models which include:
5. Decision tree
6. Linear discriminant analysis
7. Random forest
8. Gradient boosting
9. Checking the accuracy, Specificity and sensitivity.
10. Finding out which feature is most important.

* The dataset consists of data on female patients of Pima Indian heritage who have either been tested positive for diabetes or tested negative. The problem can be taken as classification to predict if the person would have diabetes or not.

# Implementation

* The data was first downloaded and imported in the code as a data frame. The data then was checked for null values. There were no null values however the data consisted of several null values in form of ‘0’.
* The data should be improved so to replace these rows containing zero the library imputer was used and ‘mean’ was kept as the strategy.
* The data was then used to find correlation between the features. Then it was split into test and train data for training the models.
* The models were trained using the data and the accuracy specificity and sensitivity of the model was taken into notice.
* The model then was used to find what feature is the most important one.
* Relation between recall and precision was check.

# Decision tree

* The decision tree classifier consists of a tree which makes decision based on the feature. The selection of the feature is done based on either Gini index or entropy.
* The decision tree model gave following outputs:

1. Accuracy: 0.718
2. Confusion matrix: [111 32 33 55]
3. True positive rate: 62.5%
4. True negative rate: 77.62%

# Linear discriminant analysis

* The liner discriminant analysis has
* The linear discriminant analysis had following outputs:

1. Accuracy: 0.744
2. The confusion matrix: [128 43 16 44]
3. true positive rate: 73.33 %
4. true negative rate: 74.85 %

# Random forest

* The Random Forest consists of multiple decision tree of with the majority is taken to predict the output. The data set is also feed random to these decision tree.
* The Random Forest had following outputs:

1. Accuracy: 0.766
2. The confusion matrix: [130 40 14 47]
3. true positive rate is: 77.05 %
4. true negative rate is: 76.47 %

# Gradient Boosting

* The Gradient Boosting had following outputs:

1. Accuracy: 0.731
2. The confusion matrix: [117 35 27 52]
3. true positive rate: 65.82 %
4. true negative rate: 76.97 %

# Conclusion

The accuracy was found to be greatest of random forest classifier. In addition to that the feature that was most important was Plasma glucose concentration 2 hours in an oral glucose tolerance test.