# Diabetes in Pima Native Americans

Andrew Min
The University of Texas at San Antonio, San Antonio TX, 78249

## Introduction

Machine learning is known to be able to predict many things, including whether a patient has a disease or not. This poster will go over three different statistical machine learning methods to determine how well they can predict diabetes in a patient.

## Data Structure

The data is from the National Institute of Diabetes and Digestive and Kidney Diseases. All patients are female whose age is at least 21 years, and they are of Pima Native American heritage.

There are nine variables in this dataset, one being our target variable: *Outcome*.

The other eight variables are:

- *Pregnancies* (Number of times pregnant)
- *Glucose* (Blood sugar levels)
- *BloodPressure* (Diastolic blood pressure, mm Hg)
- *SkinThickness* (Triceps skinfold thickness, mm)
- *Insulin* (Insulin 2 hours later, mu U/ml)
- *BMI* (body mass index, weight in kg / (height in m)^2)
- *DiabetesPedigreeFunction* (scores likelihood of diabetes based on family history)
- *Age* (age of the patient in years)

Based on intuition and research, Glucose, BloodPressure, SkinThickness, Insulin, and BMI have values that are humanly impossible; thus, they are considered as missing values and will be handled as such.

## Methods

- Logistic regression

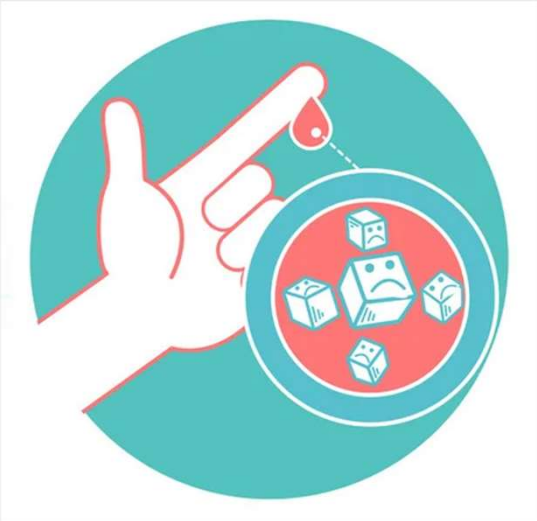
Selected as a baseline to be compared with the other two models.

- Random Forest

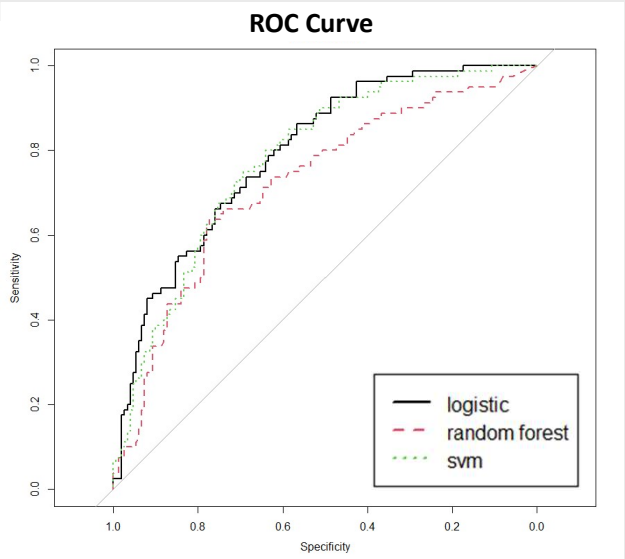
Selected because it can perform well on imbalanced data, such as the one for this project.

- Support Vector Machine (SVM) (using radial kernel)

Selected based on another paper that ran similar tests on predicting diabetes (Abbas et al.).



## Results




## Results - con't

**Metrics after Predictions**

| | Logistic | Random Forest | SVM |
|---|---|---|---|
| **Accuracy** | 0.7217 | 0.7174 | 0.7217 |
| **Kappa** | 0.3902 | 0.3753 | 0.3794 |
| **Sensitivity** | 0.7905 | 0.7815 | 0.7792 |
| **Specificity** | 0.5976 | 0.5949 | 0.6053 |
| **AUC** | 0.7926 | 0.7222 | 0.779 |

## Conclusions

Because the data is imbalanced, we will be mainly focusing on the AUC metric.

- Logistic regression has the best AUC score.
- While the other two models did not achieve the best score, they still performed well.

Using a larger dataset, additional predictors, different preprocessing methods and tuning parameters may improve the accuracy.

## References

Abbas HT, Alic L, Erraguntla M, Ji JX, Abdul-Ghani M, Abbasi QH, et al. (2019) **Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test.** PLoS ONE 14(12): e0219636. https://doi.org/10.1371/ journal.pone.0219636