## AWS Build-A-Thon

# Diabetes Prediction Model for Indian Women

-Deepthi V N

Github
(https://github.com/SmartPracticeschool/SPS-1609-Diabetes-Prediction-Model-for-Indian-Women/)

Video
https://youtu.be/PEUAb6fDkbk

## Objective

The diabetes dataset is a binary classification problem where it needs to be analysed whether a patient is suffering from the disease or not on the basis of many available features in the dataset. Different methods and procedures of cleaning the data, feature extraction, feature engineering and algorithms to predict the onset of diabetes are used based for diagnostic measure on Pima Indians Diabetes Dataset.

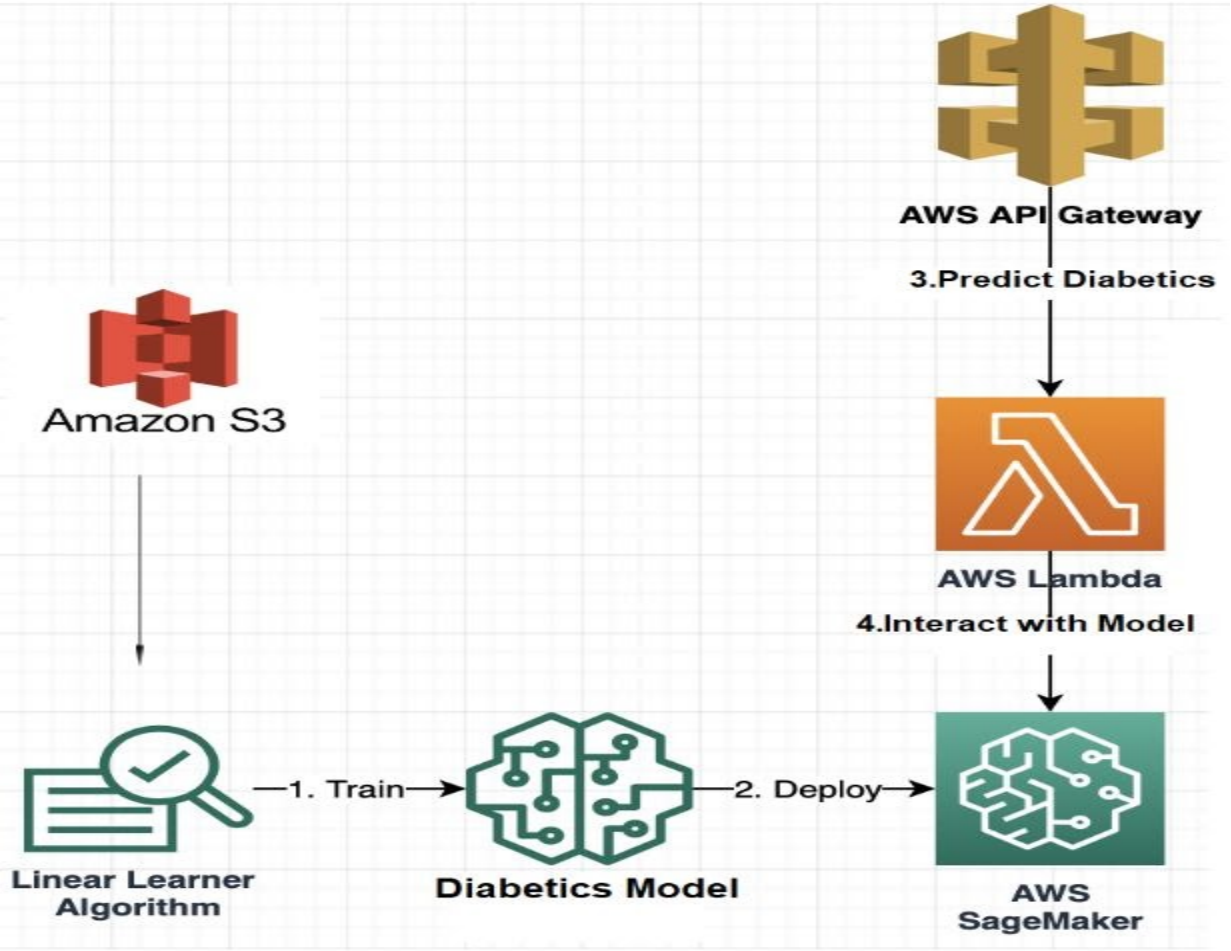
## Method

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.

The raw data contained 768 rows (customers) and 9columns (features). The Outcome column was target variable. 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.

I cleaned the dataset by replacing the cells with value 0 to the mean of the column. I explored the correlation between several features and the target variable.

Then I have built and deployed a Machine Learning model to predict the diabetes using Amazon SageMaker and predictions can be obtained by using its Endpoint.
A python - flask application is also created that interacts with the model deployed on AWS Sagemaker with the help of AWS API Gateway and AWS Lambda Services.




## Model

Amazon SageMaker provides an XGBoost container that we can use to train in a managed, distributed setting, and then host as a real-time prediction endpoint. XGBoost uses gradient boosted trees which naturally account for non-linear relationships between features and the target variable, as well as accommodating complex

interactions between features.

Amazon SageMaker XGBoost can train on data in either a CSV or LibSVM format. Also it should

- Have the predictor variable in the first column
- Not have a header row

Using this technique I have got Overall Classification Rate(accuracy) of 72.7%.

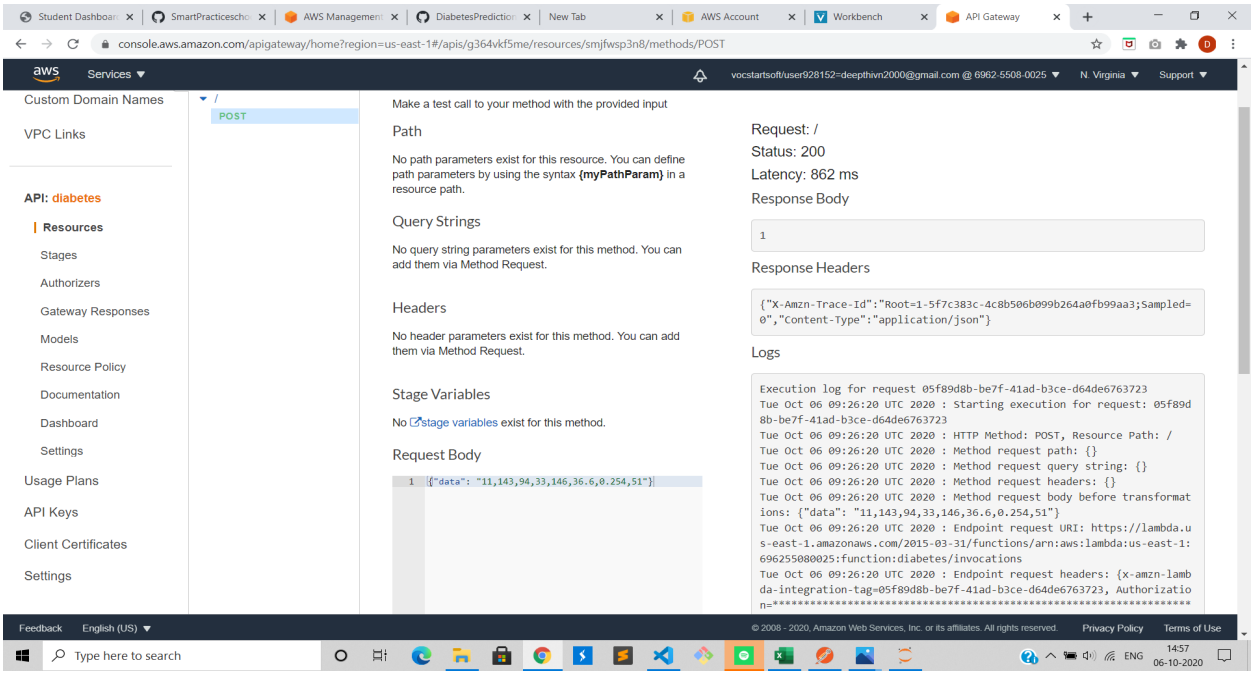
## Conclusion

This machine learning model can be used to predict whether or not the patients in the dataset have diabetes or not .

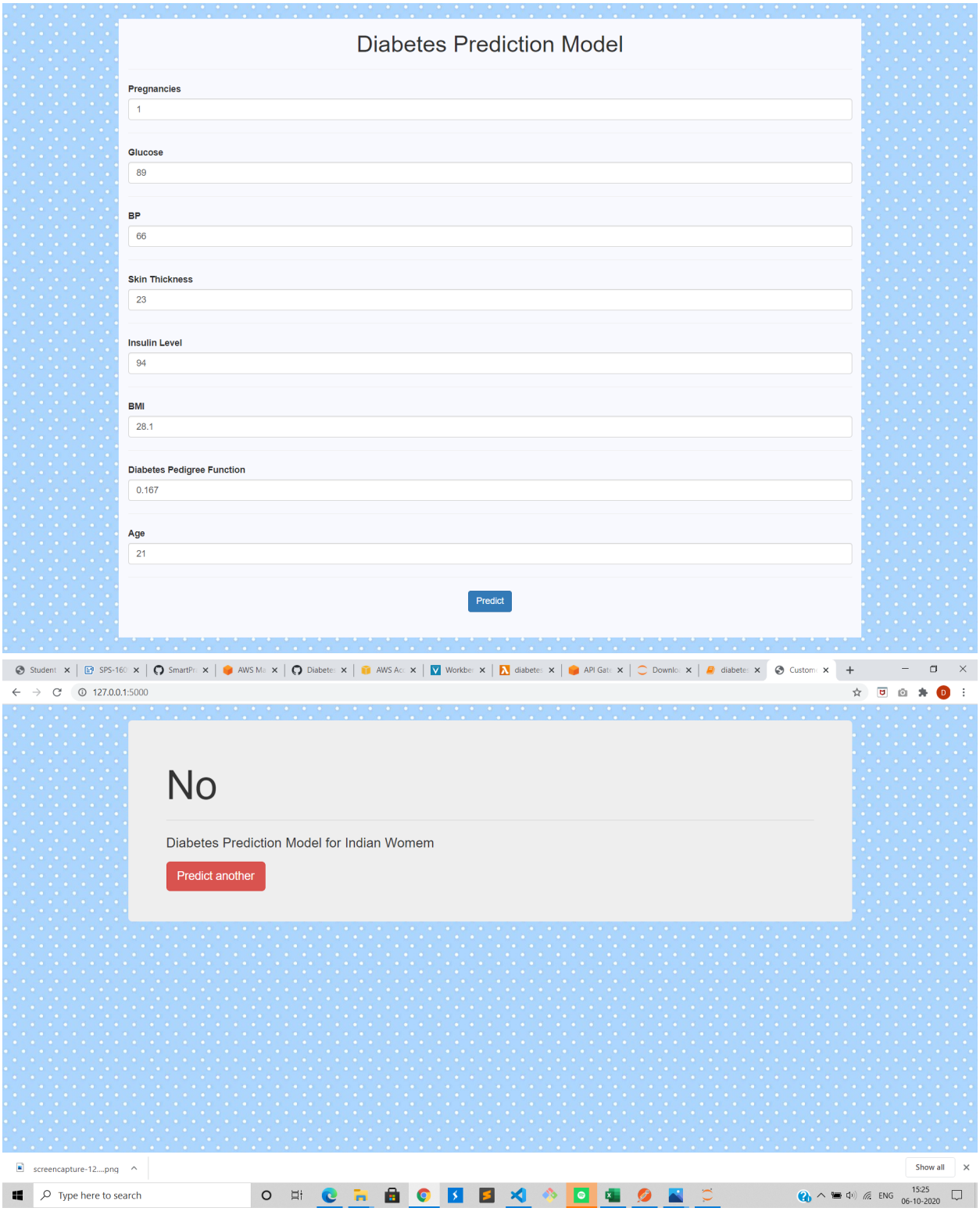
## Future Work

One can optimize the sensitivity of the model (further decrease the number of false negatives). Also one can invest more in feature engineering and try and include additional features from other datasets and make it more generalized.

## Snippets



Testing the API




Web App using Flask