CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)

Tejas N. Rao

Project Number

38048

Project Title

Logistic Regression and Decision Tree ML Algorithms to Predict Type-2 Diabetes

Abstract

Objectives/Goals

Compare two Statistical Models to predict Type-2 Diabetes - Logistic Regression and Decision Trees. Determine which patient attributes - Age, Body Mass Index, Glucose Concentration, Genetics, # of time pregnant are most significant for Diabetes

Determine the following for each model to aid comparison: Accuracy, Sensitivity, Specivity, ROC Area under Curve.

Build a simple web application to use the model in mobile phones. Application should accept key patient data and return probability of diabetes
Application should run on phone and browser.

Methods/Materials

UC Irvine Department of Machine Learning Pima Indians Diabetes DataSet. This dataset provides details on 782 Pima Indians for Age, BMI, Pregnancy etc. Scikit-learn: Machine learning in Python Logistic Regression and Decision Tree algorithm packages in Python. Pythonanywhere for Hosting and running Python Applications. Jupyter notebooks running on Azure Cloud.

Methods
Scikit-learn Machine Learning toolkit in Python was used for running Classification Models DataSet has 768 patient records which were divided into 75% (576 records) for Training data and remaining 25% (192 records) for Test data.
Both models Logistic Regression and Decision Trees, are Trained and Scored with training data and test data respectively

Prediction Accuracy is measured as (TP+TN) / (TP+TN+FP+FN)
Sensitivity is measured as TP / (TP+TN)
Specificity is measured as TN / (TN+FP)
HTML5 was used to build a simple webapp that accepts Patient Data in a Form and calls backend Python App.

Results

Logistic Regression Model has

Summary Statement

Prevent Diabetes using Machine Learning Algorithms- Logistic Regression and Decision Trees

Help Received

Mr Wilke (San Mateo High School), Ms Bharathi Udupi (Oracle)