Predicting Diabetes Diagnoses Using Machine Learning

By Allie Wicklund

"Will using Machine Learning techniques increase accuracy of the model by at least 5% (model accuracy of at least 87.98%), thus significantly improving upon the logistic regression model?"

The current model proposed on Kaggle is a logistic regression model, which has 82.98% accuracy on its testing set of predicting whether a patient is diabetic or not. This dataset is limited in its range to Pima Indian females ages 21 or higher. The study is of interest to the owners of the Pima Indians Diabetes Database, and to those of Pima Indian heritage at risk for diabetes. If machine learning contributes a significant increase of model accuracy, then the database owners would be justified in using machine learning techniques to model this data instead of using logistic regression.

The data can be obtained via Kaggle, but this study will scrape the site, providing the .csv file with BeautifulSoup so that the entire project can be run off of Jupyter Notebook with no downloading of files required. Cleaning and analysis of the data will be conducted in the same Jupyter Notebook. The study will focus on 2-3 different machine learning techniques and their testing accuracy scores, as well as each model's "goodness of fit" factors, which include: computing time, difficulty of implementation, and difficulty of explaining the model to a broader audience. The project will be presented as a slide deck and a project report, which will both be available on GitHub.