Domain background, Proposal statement, Datasets, and inputs

In medical science, it is very useful to be able to forecast the outcome based on certain symptoms or measurements. Machine learning tools can help us in predicting certain outcomes based on the data provided for pertinent measurements. In this project, I will try to make a model for predicting the diabetes status of a patient. I will use a dataset from the study of atleast 21 years old female Pima Indian population near Phoenix, Arizona [1]. National Institute of Diabetes and Digestive and Kidney Diseases performed the study to understand the high diabetes rates in the population. The study took certain diagnostic measurements including the diabetes status of the patient. The goal of this project is to make a prediction for the diabetes status based on the results of diagnostic measurements. The dataset being used in the project is taken from Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database).[2]

There are 768 rows in the dataset. It has eight dependent variables representing various diagnostic measurements and an independent variable showing the diabetic condition of a person. The eight dependent columns are as follows [1]:

'**Pregnancies'**: showing the number of times the female had been pregnant

'**Glucose'**: Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test (GTIT)

BloodPressure: Diastolic Blood Pressure (mm Hg)

SkinThickness: Triceps Skin Fold Thickness (mm)

Insulin: 2-Hour Serum Insulin (µh/ml)

BMI: Body Mass Index (Weight in kg / (Height in in))

DiabetesPedigreeFunction: Diabetes Pedigree Function

Age: Age (year)

All these columns have values in a range of integers or float values whereas the final column showing the diabetes result is has either a value 1 or 0 showing diabetes or no diabetes

Proposed solution, Evaluation metrics and project design:

Since the objective here is to make a classifier that can classify the diabetes status of a female based on multiple variables, we can use one of many regressors available. The classification can be done using SVMs, Decision tree classifiers, AdaBoost, Random forest, naïve Bayes etc. In this project, I will use a few of these classifiers and compare the accuracy of these classifiers against each other.

I will use the classifier that gives the best accuracy, precision and recall (f_beta score) and optimize its parameters to obtain the best using grid search.

In this dataset, it is much important to use sensitivity (rate of true positives) and recall (the ability to detect true positives). Since we want to correctly recognize patient with diabetes and we do not want to miss someone with diabetes. The relative emphasis on precision and recall can be changed by changing the parameter beta in f_beta score. But here we will just set it to 1 to get f1 score, which places equal emphasis on recall and precision.

Benchmark model:

The benchmark model will be the simple logistic regression model.

References:

[1]. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

[2]. https://www.kaggle.com/uciml/pima-indians-diabetes-database