Machine Learning Engineer Nanodegree

Capstone Proposal

Akhith Palkamsetti
June 10th, 2018

Proposal

Domain Background

Diabetes mellitus commonly referred to as diabetes, is a type of metabolic disorder in which there are high blood sugar levels over a prolonged period. People diagnosed with diabetes face Symptoms such as frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications which include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes.

Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced. There are three different types of Diabetes: * Type 1 DM results from the pancreas's failure to produce enough insulin.

- Type 2 DM begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses a lack of insulin may also develop.
- Gestational diabetes is the third main form, and occurs when pregnant women without a previous history of diabetes develop high blood sugar levels.

The most common cause for Diabetes is excessive body weight and insufficient exercise. As obesity has increased considerably in today's world Diabetes has become a common sight. It is estimated that 415 million people are living with diabetes in the world, which is estimated to be 1 in 11 of the world’s adult population. 46% of people with diabetes are undiagnosed. It would be beneficial if Diabetes can be diagnosed early or we could predict the risk a person has towards getting Diabetes. In this project I will build a model that can classify whether a person has diabetes or not by looking at some attributes .

Problem Statement

The problem is a machine learning classification problem. In this project we will use the Pima Indians diabetes dataset to predict whether a member has Diabetes or not. The Pima or Akimel O'odham (also called as "river people") are a group of Native Americans living in an area consisting of what is now central and southern Arizona. Pima Indians from the Gila River Indian Community in Arizona have a high incidence rate of type 2 diabetes, and kidney disease attributable to diabetes is a major cause of morbidity and mortality in this population.

The members of the dataset are females of age 21 and above, who belong to the pima indian groups. In this problem we will be considering various potential casues for diabetes and by using Supervised learning algorithms we can predict the likelyhood a person having diabetes.

Datasets and Inputs

National Institute of Diabetes and Digestive and Kidney Diseases provided the Pima Indians Diabetes Database for research purpose to the UCL machine learning dataset web site. The diabetes check has conducted only on female patents in the times of their pregnancy. It contains 9 parameters among that 8 is input parameters and 1 is output parameter. It contains 768 patient's instances.

I was not able to find the dataset in the UCL repository, hence I decided to use the dataset that can be found at [3] provided by Kaggle.

Below are the attributes present in the data set and their descriptions:

Attribute	Type	Description
Pregnancies	numeric	Number of times pregnant
Glucose	numeric	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure	numeric	Diastolic blood pressure (mm Hg)
SkinThickness	numeric	Triceps skin fold thickness (mm)
Insulin	numeric	2-Hour serum insulin (mu U/ml)
BMI	numeric	Body mass index (weight in kg/ (height in m)^2)
DiabetesPedigreeFunction	numeric	Diabetes pedigree function
Age	numeric	Age in years
Outcome	numeric	Zero if Diagnosed negative and One if Diagnosed Positive

Among the 768 entries 34.9% are members who have tested positive and rest are negative. The Pima Indian Diabetes dataset has lot of missing values. To replace the missing values we could simply remove the missing entries from the dataset, however this may not be a correct option because we may loose lot of data and since we have only 768 entries the dataset would become

smaller than it already is, this would lead to difficulty in classification as the model will not generalize well. So instead we will replace the missing value with the separate class mean value or median based on which provides a better result in our end classification.

I believe all the attributes mentioned above would contribute contribute significantly towards determining the presence of Diabetes. I believe attributes like Glucose, Blood pressure, Insulin and BMI would play a major role as discussed earlier about what causes diabetes in the problem statement section. It is also expected that diabetes would be more among the older members of the dataset compared to the younger ones as metabolic rate of a person drops with increase in age.

Solution Statement

First I will have to clean up the dataset, and make sure that the missing values are taken care of. As mentioned earlier the plan is to replace the missing values with either the mean or median of the attributes each of the missing values fall into. Once the refined data is obtained, I will test various models such as SVM, Naive Bayes, Decision Trees, Random Forest etc. and evaluate the best model. The performance of the model will then be evaluated using some of the evaluation metric tools available in scikit learn such as the Confusion matrix.

Benchmark Model

There have been extensive studies of this dataset in the Machine Learning Literature. Various classification algorithms have been applied to the data set, and no algorithm performs exceptionally well. For this project i would like to use one of the models that were mentioned in [1].

A class-wise K-nearest neighbours model was used on the Pima Indians diabetes dataset. Various factors such as Accuracy, Sensitivity and Specificity were compared with simple KNN. The proposed CKNN model gives better classification accuracy as 78.16% compared to simple KNN.

Another model that was used was based on the SVM with Radial basis kernel function which also did really well and gave a classification accuracy of 78%. My goal in this project would also be to try and achieve an accuracy close to 78% as well.

Evaluation Metrics

I will be using The confusion matrix and F-beta_score(beta = 1) as the evaluation metrics for the benchmark model and the proposed project. The mathematical representations for each of the metrics can be seen below.

```
F1-score = 2 (precision * recall)/(precision + recall)

        Precision = TP/( TP + FP )
        Recall = TP/( TP + FN )

        where,

        TP = True Positives

        FP = False Positives

        TN = True Negatives

        FN = False Negatives
```

The F1_score is a measure of the model's accuracy and depends on precision and recall.

Project Design

The project design will be as per the steps below: * Programming Language and Libraries:

```
1) Python 2.7

2) Scikit Learn

3) Matplotlib

4) Numpy
```

- Data Exploration and Preprocessing:

  In this part I will start of by looking at the different atributes of the dataset and what they mean to the problem at hand. I will then work on the missing values of the data set. The dataset needs to be cleaned and the missing values will be replaced with either the mean or the median. Once the refined dataset is obtained I will check the spread of each of the attributes by plotting histograms. From the obtained data, we will look for skewed data if any and later perform normalizing. Normalizing is mainly done so that all attributes are treated equally by the classifier and factors such as higher magnitude in one attribute compared to the other doesn't affect the classification.

- Data Splitting and Training:

  Here we will split the processed data into training and testing sets. The data will be split in the ration of 4:1 or 7:3 with the training set being the larger set. Once the training and testing sets are divided we will use different classifiers available in scikit-learn to build the model that best classifies the data.

- Evaluate Model:

  In this step after training each of the classifiers, It is time to evaluate them using the selected evaluation metrics (F-beta_score and Confusion Matrix), and finalizing on the best classifier.

- Fine tuning the model:

  In this last step, I will look to fine tune the model by playing around with different parameters. I will also check for feature relevance and try to understand how much each of the attributes of the dataset actually affect the classifier. Based on the result, we can look to drop the non-relevant features.

References

[1] https://www.diabetes.co.uk/diabetes-prevalence.html

[2] https://en.wikipedia.org/wiki/Pima_people

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

[4] International Journal of Engineering and Applied Sciences (IJEAS) ISSN: 2394-3661, Volume-2, Issue-5, May 2015.