**Machine Learning Engineer Nanodegree**

**Capstone Proposal**

**Neelam Sinha**

**January 26th 2020**

**Domain Background**

Diabetes is a common physiological health problem among humans across gender, race and age. The term diabetic is applied when an individual is unable to break down glucose, for lack of **insulin**. The human organ called pancreas is responsible for generating the hormone called insulin, which is a very important enzyme that regulates the sugar level in human blood stream. As a result of lack of insulin, the body cells didn’t get the energy they need, and thus elevates the sugar level in the blood, and many problems/ diseases can emerge like heart attack, blindness etc.

Diabetes is not a curable disease; although, fortunately, it is treatable. Diabetes and related complications are responsible for the death of almost 200,000 Americans every year [1, 2]. In past there are many works done using standard statistical techniques such as discriminate analysis, regression analysis, and factor analysis, neural network models to provide prediction but these standard statistical methods may provide disappointing results when:

* The sample size is small.
* The form of the underlying functional relationship is not known.
* The underlying functional relationships involve complex interactions and intercorrelations among several variables.[3]
* Using neural network models[4] will lead black box situation, the understand and interpretation of the model, and its working become very difficult.

That’s why I choose to use simple supervised machine learning models in my project to predict diabetic or non-diabetic patients based on their diagnostic measurements.

**Problem Statement**

In modern healthcare, predicting and properly treating diseases have become of foremost importance in medical prognostics fields. But even in this era, the normal identifying, whether the person is diabetic or non-diabetic process need patients to visit a diagnostic center, consult their doctor, and wait for a day or more to get their report. So, the objective of this project is to use Machine Learning methods to predict whether a person has diabetes or not, in a supervised fashion.

**Dataset Used**

The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases” as part of the [Pima Indians Diabetes Database](https://www.kaggle.com/uciml/pima-indians-diabetes-database/data). Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above.

Dataset Link: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

The data set is saved in diabetes.csv format file and, from the domain knowledge, I have analyzed and found out the ranges of values and its effects on diabetes.

There are 8 independent variables:

1. Pregnancies: No. of times pregnant

2. Glucose: Plasma Glucose Concentration a 2 hour in an oral glucose tolerance test (mg/dl)

A 2-hour value between 140 and 200 mg/dL (7.8 and 11.1 mmol/L) is called impaired glucose tolerance. This is called "pre-diabetes." It means you are at increased risk of developing diabetes over time. A glucose level of 200 mg/dL (11.1 mmol/L) or higher is used to diagnose diabetes.

3. Blood Pressure: Diastolic Blood Pressure(mmHg)

* If Diastolic B.P > 90 means High B.P (High Probability of Diabetes)
* Diastolic B.P < 60 means low B.P (Less Probability of Diabetes)

4. Skin Thickness: Triceps Skin Fold Thickness (mm) –

A value used to estimate body fat. Normal Triceps SkinFold Thickness in women is 23mm. Higher thickness leads to obesity and chances of diabetes increases.

5. Insulin: 2-Hour Serum Insulin (mu U/ml)

6. BMI: Body Mass Index (weight in kg/ height in m2) –

BMI of **18.5 to 25** is within the normal range, BMI between **25 and 30** then it falls within the overweight range. A BMI of **30 or over** falls within the obese range.

7. Diabetes Pedigree Function: It provides information about diabetes history in relatives and genetic relationship of those relatives with patients. Higher Pedigree Function means patient is more likely to have diabetes.

8. Age (years)

9. Outcome: Class Variable (0 if non-diabetic, 1 if diabetic)

The **dependent variable** is whether the patient is having diabetes or not.

**Solution Statement**

**Solution:** As the problem statement have to classify the data into patients having diabetes or not, the best method which can be used is Classification Tree Algorithm or Classification Ensemble Algorithm. Further we can easily classify and predict the outcome using nodes and internodes.

**Software Package Used:** Python, Libraries Used: Pandas, Scikit Learn, Numpy, Scipy, Matplotlib,Seaborn.

**Evaluation Metrics**

As the dataset has classification problem, there are many pre-defined evaluation metrics which we can use. For example, F1\_score, Recall, Precision, Accuracy, ROC\_AUC. Etc. In used two metrics to evaluation.

**Accuracy- How Accurate is your model classifying the testing data.**

The most common metric for classification is accuracy, which is the fraction of samples predicted correctly as shown below:

Accuracy= (TP +TN)/ (TP+TN+FP+FN)

We can obtain the accuracy score from scikit-learn, which takes as inputs the actual labels and the predicted labels.

**ROC (Receiver Operating Characteristic)**- ROC Curve tells us about how good the model can distinguish between two things (e.g If a patient has a disease or no). Better models can accurately distinguish between the two. Whereas, a poor model will have difficulties in distinguishing between the two.

**Project Design**

### Data Cleaning

### In the dataset Data cleaning will take place as data has got lot of Zeroes in some columns like Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, which does not make sense as these columns are important and must have values. So, I consider them as missing value and Impute the columns in accordance with their distribution. I also visualize the data after cleaning which help me to understand in better fashion

### Splitting Dataset

### To have unknown datapoints to test the data and model rather than testing with the same points with which the model was trained, data split is required. This helps capture the model performance much better. So, Here I use train\_test\_split function from sklearn.model\_selection to split data in training set and testing set with testing size of 0.25 i.e randomly ¼ of the dataset will be testing dataset and rest will be used as training dataset to train the model.

### Classification Model

### A screenshot of a cell phone Description automatically generatedModel selection is one of the difficult jobs, but as the problem statement is about classification. So, first to best guess optimal model I use Classifier tree algorithm(DecisionTreeClassifier()), Classifier ensemble algorithm(RandomForestClassifier(), GradientBoostingClassifier(), AdaBoostClassifier()) and Also use KNN and SVM model to check which model performs well on the given dataset.

### *Figure1: Important feature. We can see Glucose, BMI, Age and DiabetesPedigreeFunction*

### After running all model, the accuracy and AUC score I got was not very high, which can be further improved by using only relevant features (figure1) and then Standardize the dataset based on these four important features (first four with highest weight in figure 1).

### After Standardization the Accuracy and AUC were improved. Thus, we select our best classifier model i.e RandomForestClassifier(). Then Finally I tried to optimize the Random forest and optimize my model with an accuracy of 81% and AUC score of 80%. Figure 2 shows the roc\_auc curve plot.

### 

***Figure2: ROC\_AUC curve of the Randomforest Classifier model with an AUC of 0.7983 (~0.80).***

**Conclusion:**

Advantage of this project: The pipeline and algorithm predict and identify patients suffering from diabetes with good accuracy. Further predicting the disease at stage early leads to treating the patient before it becomes critical.

**References**

1. Alberti, K. G. M. M., and Zimmet, P. F.: Definition, Diagnosis and Classification of Diabetes Mellitus and Its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus. In: Provisional report of a WHO consultation. Diabetic medicine, 15(7), 539-553 (1998).

2. National Diabetes Data Group: Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. In: Diabetes, 28(12), 1039-1057 (1979).

3. Simmons, GJ., A Mathematical Model for an Associative Memaory. Sandia Coporation Monograph, SCR-641,1963).

4. Smith, J. W., Everhart, J., Dickson, W., Knowler W., and Johannes, R.: Using the Adap Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In: Proceedings of the Annual Symposium on n Computer Application in Medical Care. American Medical Informatics Association, p. 261 (1988).