MACHINE LEARNING ENGINEER NANODEGREE

JUNE 27 2018

CAPSTONE PROPOSAL

KUNAPARAJU GAYATRI PADMA

JUNE 27 2018

Predicting whether the patient is diabetic or not.

DOMAIN BACKGROUND:

- In this fast-growing world, although the technology had advanced a lot there are many health problems raising at an alarming rate.
- This problem should be addressed because many people are losing their lives because they haven't realized at the earlier time.
- So, I selected one of the most common diseases diabetes. Using the data, we can identify or predict a person whether he is having diabetes are not. This might help the persons to take precautions as they can identify it earlier because it has become common and unlike past days it is not specified to only some particular age, so we can identify it earlier.
- The major motivation for choosing this project is as we have to work on real time problems that may include any field, it is major and common issue, so using this a small change can save many lives.

PROBLEM STATEMENT:

- By using the dataset, we need to find whether the person has diabetes or not.
- In the dataset we have different columns which are used for the prediction of the disease.
- We need to implement machine learning techniques used for prediction of features. Using data visualizations and algorithms we need to predict the person is diabetic or not.

DATASETS AND INPUTS:

- The dataset contains a CSV file which contains different columns used for the prediction.
- Link: https://www.kaggle.com/uciml/pima-indians-diabetes-database

- The dataset consists of following inputs: Pregnancies, Glucose, Blood Pressure, Skin thickness, Insulin, BMI (Body Mass Index), Age, DiabetesPredigreeFunction.
- Every column specified are used to find the prediction based on the values. Using all the input features based on the requirement we predict the outcome feature.

SOLUTION STATEMENT:

- In this fast-growing world, although the technology had advanced a lot there are many health problems raising at an alarming rate.By using the dataset, we need to find whether the person has diabetes or not.
- In the dataset we have different columns which are used for the prediction of the disease.
- So, I am using classification model of supervised learning by passing the current record data to the model and predicting whether the person is diabetic or not. There are no null values.
- I will perform logarithmic transformation on my data to scale the data. I will perform ensemble learning model to know whether a person is diabetic or not finalize only one model based on the fscore.
- If the model is performing better we will perform tuning and the main theme of the model is to find whether the person is diabetic or not.

BENCHMARK MODEL:

- Logistic regression is used as benchmark model. Fbeta score of benchmark model is reference and other model will be judge to perform better if their fbeta score will be greater than Logistic regression model. Accuracy, f-beta score and confusion matrix and will try to get better results in the ensemble learning models.

EVALUATION METRICS:

Accuracy in classification problems is the number of correct predictions made by the model over all kinds of predictions made.

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

In the Numerator, are our correct predictions (True positives and True Negatives) (Marked as red in the fig above) and in the denominator, are the kind of all predictions made by the algorithm (Right as well as wrong ones).

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer. The predicted positives (People predicted as cancerous are TP and FP) and the people having a cancer are TP.

Precision = TP/(TP+FP)

Recall is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer. The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP. (Note: FN is included because the Person actually had a cancer even though the model predicted otherwise).

Recall = TP/(TP+FN)

Fβ score – measures the effectiveness of retrieval with respect to a user who attaches beta times as much importance to recall as precision.

$F\beta=(1+\beta 2)\cdot precision\cdot recall/(\beta 2\cdot precision)+recall$

When beta=0.5 more emphasis is placed on precision. This is called f 0.5 score.

PROJECT DESIGN:

For a successful project, a particular sequence of steps is required.

- Firstly, we should observe the dataset and observe the input features and how they are related, this gives an intuition that how are the features vary and how they are used for output.
- Next, we need to visualize the data if needed and explore the data how they are related.
- The most important part is that we have to pre-process the data, i.e., the data may have NaN values, some features may have different units so all these kinds of data must be cleaned and pre-processed.
- We need to scale the features and transform the features and we should normalize the data because we may have skewed data.
- We need to apply the appropriate models used for prediction and evaluate the results.
- Once we got some results, we may need to optimize the model and check several times which model suits the data and which gives more accuracy.
- Finally, we get a data with good accuracy and we can test the data.