Diabetes Mellitus Prediction

Capstone Proposal

Amal Alabdulkarim
March 8, 2018

Proposal

Domain Background

Diabetes can strike anyone, from any walk of life. Diabetes is a serious life-long health condition that occurs when the amount of glucose (sugar) in the blood is too high because the body can't use it properly. If left untreated, high blood glucose levels can cause serious health complications[1]. Diabetes is responsible for 318,036 regional deaths in adults aged 20-79 years in 2017 (13% of all mortality). About 51.8% of all deaths from diabetes in MENA occurred in people under 60 [2]. Approximately 5 million of the 18 million people with diabetes in the U.S. do not know they have it [3]. Early detection and treatment of diabetes is an important step toward keeping people with diabetes healthy. It can help to reduce the risk of serious complications such as premature heart disease and stroke, blindness, limb amputations, and kidney failure [3].

In [4] the authors reviewed several machine learning and data mining approaches applied in the diabetes mellitus research. For the prediction problem, several algorithms and different approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches and association rule learning in order to achieve the best classification accuracy. Ensemble approaches, which use multiple learning algorithms, have proven to be an effective way of improving classification accuracy. The specific approaches have also been used in DM prediction[5]–[8]. Anderson et al. used a Bayesian scoring algorithm to explore the model space [8]. In [7], authors proposed an ensemble framework with multi-layer classification, using enhanced bagging and optimized weighting, combining seven heterogeneous classifiers. In [6], authors used Rotation Forest (RF), a newly proposed ensemble algorithm, to combine 30 machine learning algorithms. Finally, Han et al. presented an ensemble learning approach, which turns the "black box" of SVM decisions into comprehensible and transparent rules [5].

For me, late diagnosis of diabetes mellitus affected my family, and I chose to do prediction on this dataset because I want to help in improving the process of diagnosis.

Problem Statement

Late diagnosis of diabetes mellitus can have severe consequences. Therefore, in order to prevent these health issues, we need to expedite the process of diagnosing this condition and even predict it before its first obvious symptoms.

The best solution would be to encourage people to go visit the family doctor more often, especially if they were more likely to get it. However, this solution is not feasible to everyone, knowing that many of those people live in developing countries and may not have access to good and affordable healthcare system. Therefore, I propose using the data we have already about the patients and the disease to try and predict its occurrences and make the diagnosis much faster.

Datasets and Inputs

The dataset I will be using for this project is the Pima Indians diabetes database, I got this dataset from Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database). It is originally from National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of 768 labeled record with 8 features. All patients are females at least 21 years of Pima Indian heritage. The features of this dataset are:

1. Pregnancies: Number of times pregnant
2. Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. BloodPressure: Diastolic blood pressure (mm Hg)
4. SkinThickness: Triceps skin fold thickness (mm)
5. Insulin: 2-Hour serum insulin (mu U/ml)
6. BMI: Body mass index (weight in kg/(height in m)^2)
7. DiabetesPedigreeFunction: Diabetes pedigree function
8. Age: Age (years)
9. Outcome: Class variable (0 or 1)

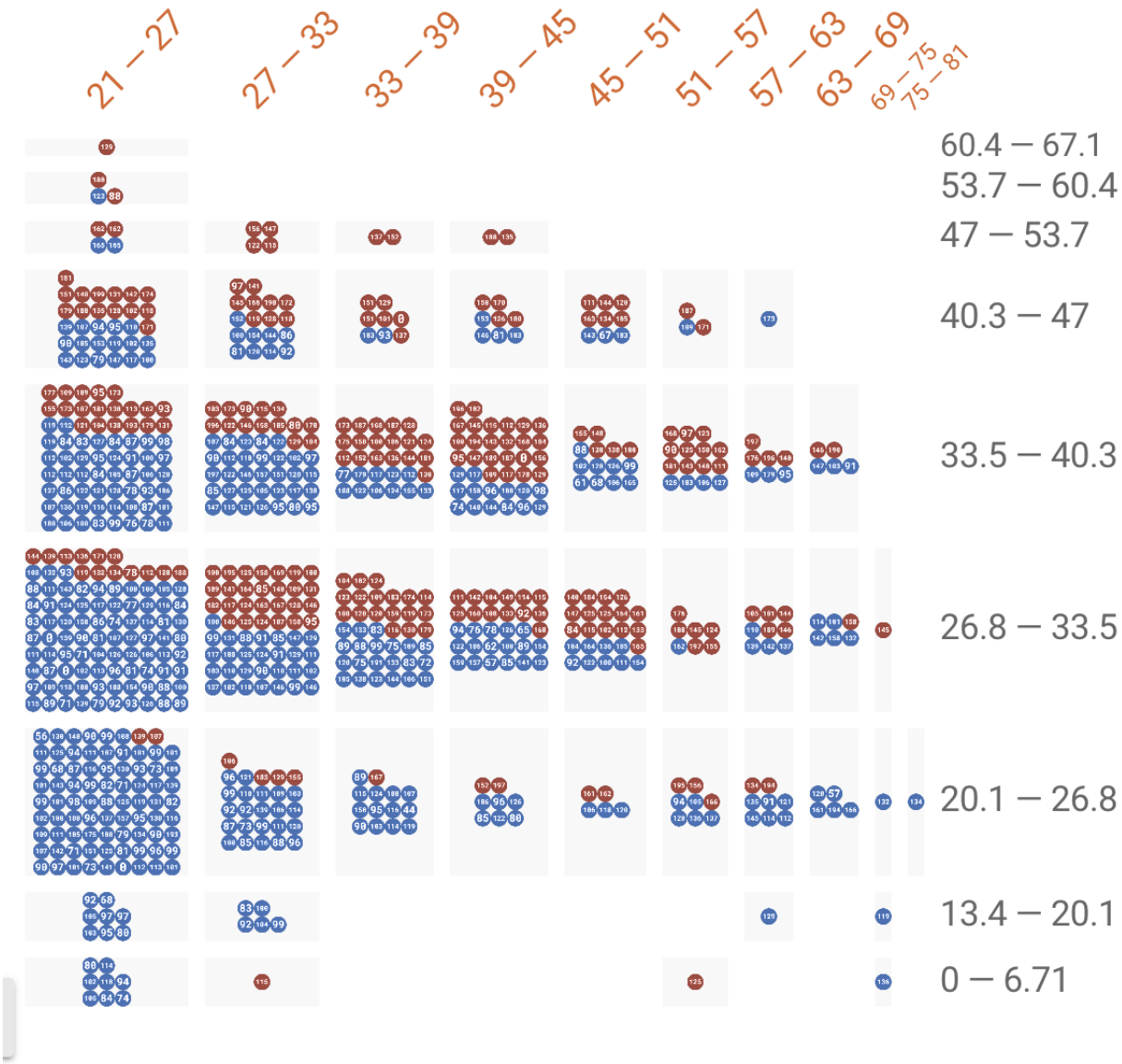

Figure 1: Visualization of the dataset where the rows represent BMI, columns represent Age and color, label on the dots represent glucose, and color represent outcome.

Solution Statement

For my capstone I plane to use XGBoost classification. I chose this model because I am trying to predict if the patient has diabetes or not (0 or 1), and my data are less than 100k. Also, XGBoost is extremely fast and it also performs well in many cases. I will also use k-fold cross validation to use all of my data in the training.

Benchmark Model

From the Kaggle kernels submitted for this dataset, I would say that my benchmark model would be an accuracy > 0.70.

Evaluation Metrics

I would use **F1 score,** because this is a classification problem and The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal [9]. The formula for the F1 score is:

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

Project Design

In this final section, I will summarize the workflow of the capstone project. This project can be divided into the following phases:

1. Analysis:
    a. Data exploration: In this sub-phase I will analyze the diabetes data, look for any abnormalities and get statistical information about this data.
    b. Exploratory visualization: In this sub-phase, I will visualize the data and the relevant features.
2. Data Preprocessing:
   In this phase, I will preprocess the data using the output of the previous phase.
3. Implementation:
   In this phase I will implement my chosen algorithm and document any important part.
4. Refinement:
   In this phase I will improve the algorithm through several iterations until I got the best result from the model evaluation metrics.
5. Justification:
   In this phase I will describe why I assumed the result I got was the best result and is my solution significant enough to solve the diabetes problem.
6. Visualization:
   Here I will visualize my results to show an important quality of the project.

Reference:

[1] "Diabetes: the basics | Diabetes UK." [Online]. Available: https://www.diabetes.org.uk/diabetes-the-basics. [Accessed: 11-Mar-2018].

[2] International Diabetes Federation, *IDF Diabetes Atlas Eighth Edition 2017*. 2017.

[3] M. O'Grady, "Diabetes, A National Plan for Action," *Natl. Diabetes Action Plan*, no. July 17, 2012, 2004.

[4] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017.

[5] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 2, pp. 728–734, Mar. 2015.

[6] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 104, no. 3, pp. 443–451, Dec. 2011.

[7] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *J. Biomed. Inform.*, vol. 59, pp. 185–200, Feb. 2016.

[8] J. P. Anderson, J. R. Parikh, D. K. Shenfeld, V. Ivanov, C. Marks, B. W. Church, J. M. Laramie, J. Mardekian, B. A. Piper, R. J. Willke, and D. A. Rublee, "Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes," *J. Diabetes Sci. Technol.*, vol. 10, no. 1, pp. 6–18, Jan. 2016.

[9] "sklearn.metrics.f1_score — scikit-learn 0.19.1 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score. [Accessed: 11-Mar-2018].