SUMMER INTERNSHIP PROGRAM 2019

20th May, 2019 - 5th July, 2019

Diabetes Detection

Final Report

MALVIYA NATIONAL INSTITUTE OF TECHNOLOGY

ROBOTICS AND MACHINE ANALYTICS (RAMAN) LAB

Submitted by:

Anshu Raikwar
NIT Manipur
anshuraikwar99@gmail.com

Sanjita phijam
NIT Manipur
phijamsan.jk@gmail.com

Under the supervision of

Prof.(Dr.) Rajesh Kumar

Dept. of Electrical Engineering
Malaviya Nationla Institute of Technology
Jaipur, Rajasthan, India

Contents

1 Introduction / Motivation

1.1 Diabetes

Diabetes is a condition or disease in which blood glucose, or blood sugar, levels in a body are too high,thus impairing the body's ability to process blood glucose. Without careful management, diabetes can lead to a build-up of sugars in the blood, which can increase the risk of dangerous complications, including stroke and heart disease.
Insulin is a hormone made by the pancreas that allows our body to use sugar (glucose) from carbohydrates in the food that we eat for energy or to store glucose for future use. Insulin helps keep our blood sugar level from getting too high (hyperglycaemia) or too low (hypoglycaemia).

Diabetes is divided into three major types:
1.Type 1 diabetes.
2.Type 2 diabetes.
3.Gestational diabetes.

1.2 Dataset Description

Context

This dataset is originally obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of this project is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

Content

This dataset consists of information on 768 of women population near Phoenix, Arizona, USA: 268 tested positive and 500 tested negative instances which indicate whether the patient is diabetic or not, respectively. Each instance is comprised of 8 attributes, which are all numeric. These data contain personal health data as well as results from medical examinations.

Inspiration

Building a machine learning model compare accuracy of seven classification algorithms to predict whether or not the patients in the dataset have diabetes or not.

Attributes:

- Pregnancies: Number of times pregnant

- Glucose: Plasma glucose concentration at 2h in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure
- SkinThickness: Triceps skin fold thickness
- Insulin: 2-h serum insulin
- BMI: Body Mass Index
- DiabetesPedigreeFunction: Diabetes Pedigree Function
- Age: age
- Outcome: Class variable

2 Problem Development

The dataset consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

2.1 Data Visualization:

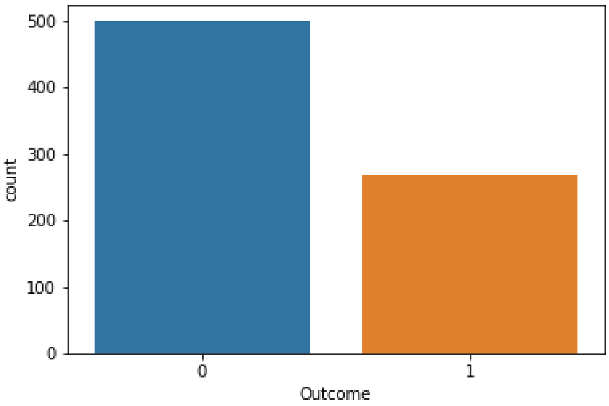

Figure 1: distribution over target variable

2.2 Distribution of predictor variables:

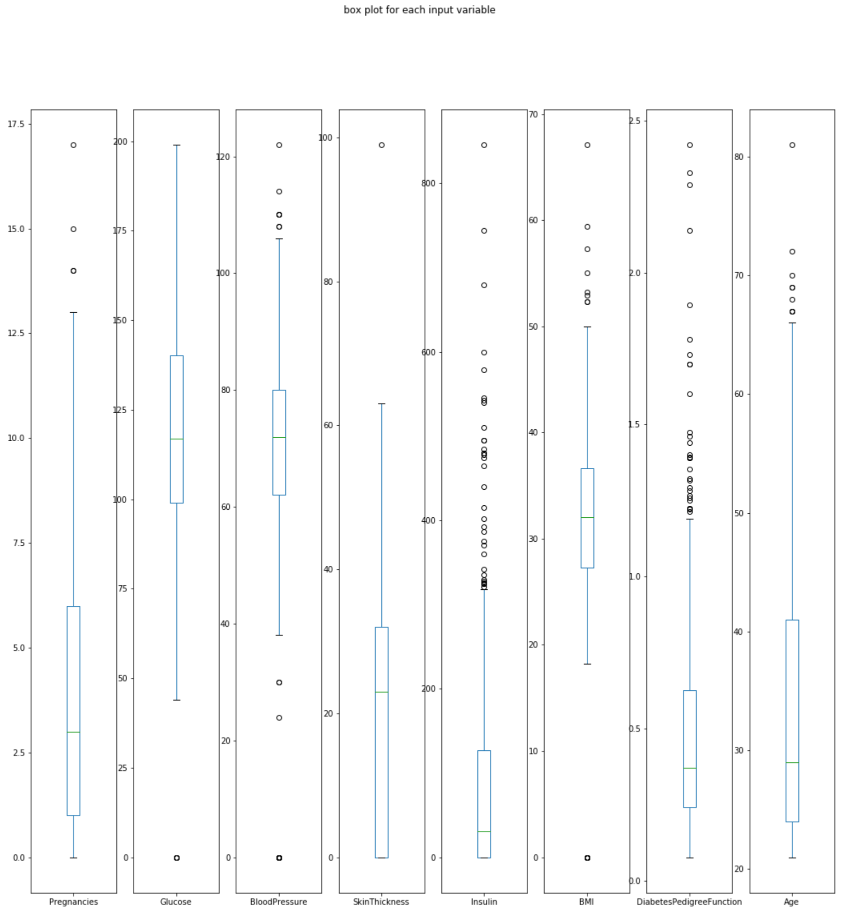

Figure 2: box plot of predictor variables

From the figure above we can conclude that attributes 'glucose' and 'skin thickness' have even distribution of data.

2.3 Correlation between Features:

The statistical relationship between two variables is referred to as their correlation. Variables within a dataset can be related for lots of reasons. It can be useful in data analysis and modeling to better understand the relationships between variables.

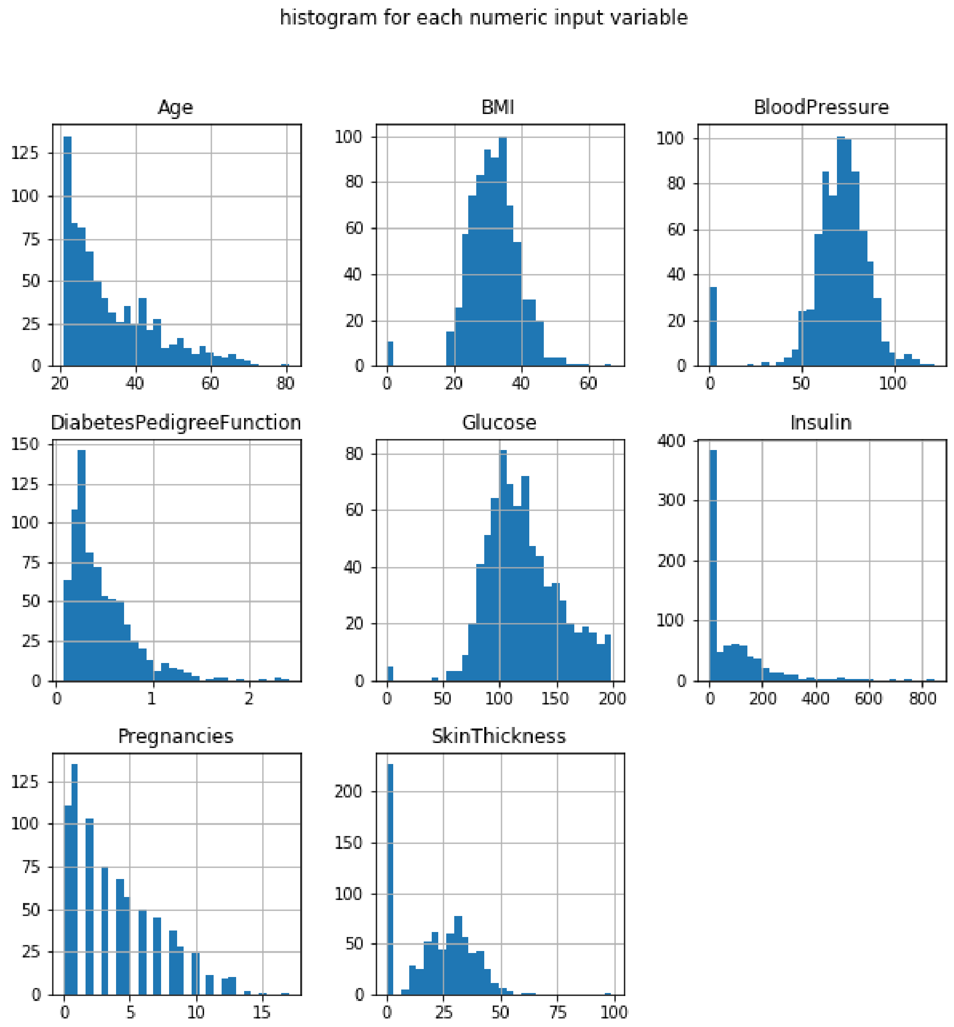

Figure 3: co-relation of predictor variables

3 Methodology

Predictive Modeling

Supervised Algorithms for Classification Problems used in this project:

3.1 KNN Classifier

Arguably the simplest machine learning algorithm. The algorithm finds the closest data points in the training dataset — its "nearest neighbors" to make predictions.

The plot between test accuracy on different number of neighbours id depicted below:

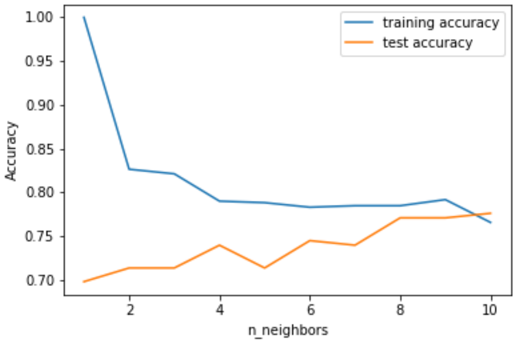

Figure 4: Accuracy of knn algorithm vs no. of neighbours

3.2 Logistic Regression

Logistic regression is one of the most common classification algorithms. It is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis.

3.3 Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Decision trees can handle both categorical and numerical data.

3.4 Random Forest

The model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer. Can be used to build predictive models for both classification and regression problems.

3.5 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

3.6 Naive Bayes Classifier

Naive Bayes classifier is a fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features.

3.7 Support Vector Machine Classifier

Support Vector Machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap as wide as possible.

4 Conclusions and Future Work

The results of the implementation of the seven classification techniques mentioned in previous section for diabetes detection is shown in Table below :

Table 1: Accuracy of algorithms

Algorithm	Accuracy (%)	
	Test	Training
KNN Classifier	77.08	79.16
Logistic Regression	80.20	76.21
Decision Tree	74.48	1.0
Random Forest	75.52	1.0
Gradient Boosting	76.04	80.56
Naive Bayes Classifier	78.12	75.69
Support Vector Machine Classifier	79.16	76.56

We are trying to detect and prevent the complications of diabetes at the early stage through predictive analysis by improving the classification techniques. The **Logistic Regression algorithm** and **Support Vector Machine** giving the highest accuracy of **80.20%** and **79.16%**, respectively holds best for the analysis of diabetic data.

In future, we are targeting to manage bigger and better dataset. If succeeded, then both accuracy and prediction of diabetes can be improved. Which can be a revolutionary step towards handling diabetes. One will easily be able to access any sort of diabetes related treatment and consult with specialists sitting right at their home.

5 Experimental Results

(a) **KNN Classifier** The accuracy of K-nearest neighbour is 77.08% obtained at 9 neighbours as seen in figure 4.

Table 2: confusion matrix - KNN Classifier

		predicted	
		0	1
actual	0	109	16
	1	28	39

(b) **Logistic Regression** Logistic regression gave the best accuracy of 80.20% among all algorithms used.

Table 3: confusion matrix - Logistic Regression

		predicted	
		0	1
actual	0	123	11
	1	27	31

(c) **Decision Tree** The accuracy obtained from Decision Tree algorithm is 74.48%.

Table 4: confusion matrix - Decision Tree

		predicted	
		0	1
actual	0	108	26
	1	23	35

(d) **Random Forest** The accuracy of random forest is 75.52%. It generates tree structure to classify the attributes under various conditions.

Table 5: confusion matrix - Random Forest

		predicted	
		0	1
actual	0	112	22
	1	25	33

(e) **Gradient Boosting** Gradient boosting gives accuracy of 76.04%

Table 6: confusion matrix - Gradient Boosting

		predicted	
		0	1
actual	0	120	14
	1	32	26

(f) **Naive Bayes Classifier** The accuracy of Naive Bayes Classifier is 78.12%.

Table 7: confusion matrix - Naive Bayes Classifier

		predicted	
		0	1
actual	0	114	20
	1	22	36

(g) **Support Vector Machine Classifier** An SVM classification technique is applied on diabetic dataset obtaining the accuracy of 79.16%.

Table 8: confusion matrix - Support Vector Machine Classifier

		predicted	
		0	1
actual	0	123	11
	1	29	29

Bibliography

[1] Basic Links:

- Python: https://www.python.org
- Installing Python: https://docs.python.org/3/installing/index.html
- Anaconda Scientific Python Distribution: https://store.continuum.io/cshop/anaconda/

[2] Dataset information:

- UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
- medilineplus.gov
- medicalnewstoday.com
- Michael Kahn, MD, PhD, Washington University, St. Louis, MO

[3] simple machine learning algorithms for classification:

- Pandas Tutorial: http://pandas.pydata.org/pandas-docs/stable/tutorials.html
- Matplotlib Tutorial: http://matplotlib.org/users/beginner.html
- Artificial Intelligence Tutorial for Beginners: https://www.youtube.com/watch?v=JMUxmLyrhSk

[4] Advanced Machine Learning Classifiers Using Scikit-Learn: http://scikit-learn.org/stable