Diabetic Patient's Data Classification and Prediction using Machine Learning Ensemble Algorithm

M. Hemalatha
Department of ECE
Sree Rama Engineering College, Tirupati, A.P, India

ABSTRACT

In this research paper, the diabetic patent dataset is collected from Indian Pima dataset for Indians. The data is understood and visualized by using Pearson correlation statistics method. According to survey 65% of this data set is non-Diabetics and 35% of Indians are Diabetics. The data is understood better by statistics and visualization. A certain pre-processing of data is performed before applying machine learning algorithms. Then machine learning algorithms are carried out on Indian diabetic data set. The ensemble (Random forest) algorithm has got good performance metrics compared to other existing algorithms. The Random forest algorithm gave outperform results compared to MLP (Multi Layer perception classifier) classifier, Support Vector Machine (SVM) classifier and LR (Logistic Regression) algorithms. The performance metrics of machine learning algorithms are calculated using confusion matrix.

Keywords

Machine learning, Random forest, Ensemble, and confusion matrix.

1. INTRODUCTION

Health of public is primary thing for protecting and curing from health hazard diseases [1]. Considerable advances in biotechnology and more specifically high throughput sequencing result incessantly in an easy and economic data production, thereby ushering the science of applied biology into the area of big data [2, 3]. The Governments are expending a huge amount of their gross domestic product (GDP) for the benefit of the public, and measures such as vaccination include prolonged the life hope of the people [4]. However, for the past many years, there has been a significant emergence of chronic and genetic diseases affecting public health. Diabetes mellitus is one of the extremely life-hazardous diseases because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage [5]. Diabetes is a metabolic disorder that impairs an individual's body to process blood glucose, known as blood sugar. -is disease is characterized by hyper glycemia resulting from defects in insulin secretion, insulin action, or both [6]. An absolute deficiency of insulin secretion causes type 1 diabetes (T1D). Diabetes drastically spreads due to the patient's inability to use the produced insulin. It is called type 2 diabetes (T2D) [7]. Both types are increasing rapidly, but the ratio of increase in T2D is higher than T1D. 90 to 95% of cases of diabetes are of T2D. Inadequate supervision of diabetes causes stroke, hypertension, and cardiovascular diseases [8]. To avoid and decrease the difficulties due to diabetes, a monitoring method of BG level plays a outstanding role [9]. A combination of biosensors and advanced information and communication technology (ICT) provides a capable real-time monitoring management system for the health condition of diabetic patients by using SMBG (self-monitoring of blood glucose) portable device. A patient can monitor the variations in glucose level in his blood by himself.

2. MATERIALS AND METHODOLOGY

The Pima Indian Diabetic dataset is downloaded from kaggle site. The tools used for processing and analyzing the data are Jupyter notebook 3.0. Feature distribution statistical analysis is carried out for detecting classification problems, where balance of feature values is needed. If the features are not balanced highly then special handling of data is needed at data preparation state. It is very significant to visualize the data before applying any machine learning models. Here Pearson correlation is applied to various attributes in the data set. The table 1 explains about correlation between the two attributes. If the value of correlation is +1 then it represents positive correlation between two attributes or features. If the value of correlation is -1 then it represents negative correlation between two variables and if correlation is o then it describes that there is no correlation between the two attributes. So, we can easily visualize the data by Pearson correlation statistical method. The Table 2 presents the skewness of the diabetic dataset. The skewness is assumed as similar to Gaussian distribution function with some distortion either in right side or in left side. If skewness is present in the dataset, then correction of dataset is necessary in data preparation stage itself to outperform in terms of accuracy. Most of the datasets are bell shaped curves so it exhibits Gaussian Distribution Function. Here skewness is used to detect the false data and it corrects the data before applying to machine learning algorithms. The various features of diabetic data set are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Diabetic Pedigree Function, BMI, Age, and Outcome. The features such as Blood Pressure and BMI have negative skewness and remaining all features has positive values. If the correlation is strong between any two class values then it is difficult to handle Linear and Logistic Regression algorithms. They perform bad when correlation between two features are high. In Table 2 pregnancies has highest value of skewness.

Table 1. Pearson Correlation between Various Attributes

	Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**	**outcome**
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
Blood Pressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
Skin Thickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1.00	0.2	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.2	1.00	0.14	0.04	0.29
Diabetes Pedigree Function	-0.03	0.14	0.04	0.18	0.19	0.19	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	0.04	-0.14	0.03	1.00	0.24
outcome	0.22	0.47	0.07	0.07	0.13	0.13	0.17	0.24	1.00

Table 2: Skewness measurement for various Attributes of sugar patients

Pregnancies	0.90
Glucose	0.17
Blood Pressure	-1.84
Skin Thickness	0.11
Insulin	2.27
BMI	-0.43
Diabetes Pedigree Function	1.92
Age	1.13
outcome	0.64

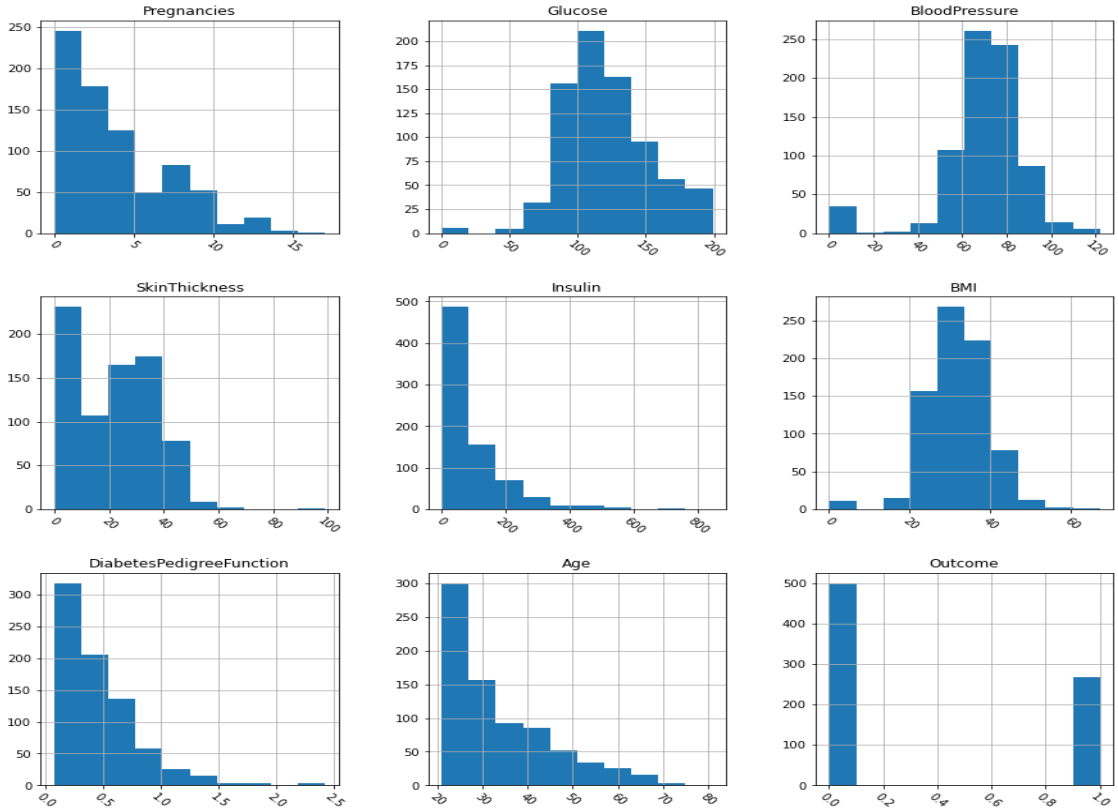

Fig 1: Distribution visualization of Indian diabetic dataset

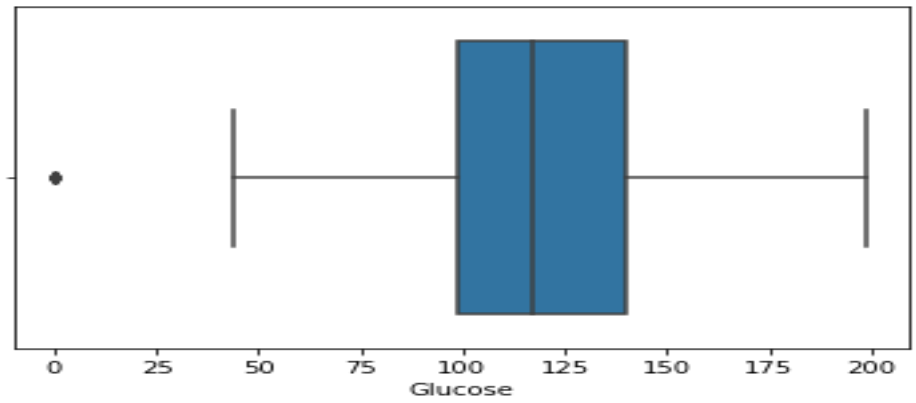

Fig 2: Data cleaning for Glucose attribute

The Figure 1 shows distribution for visualization of Indian diabetic dataset. After data visualization the next step is data cleaning. The Glucose feature after data cleaning is shown in Figure 2. The plot for Blood Pressure for different values of pregnancy is shown in Figure 3.The Zero's in outcome feature describes non-diabetic patients. The one's in outcome feature presents diabetic patients.

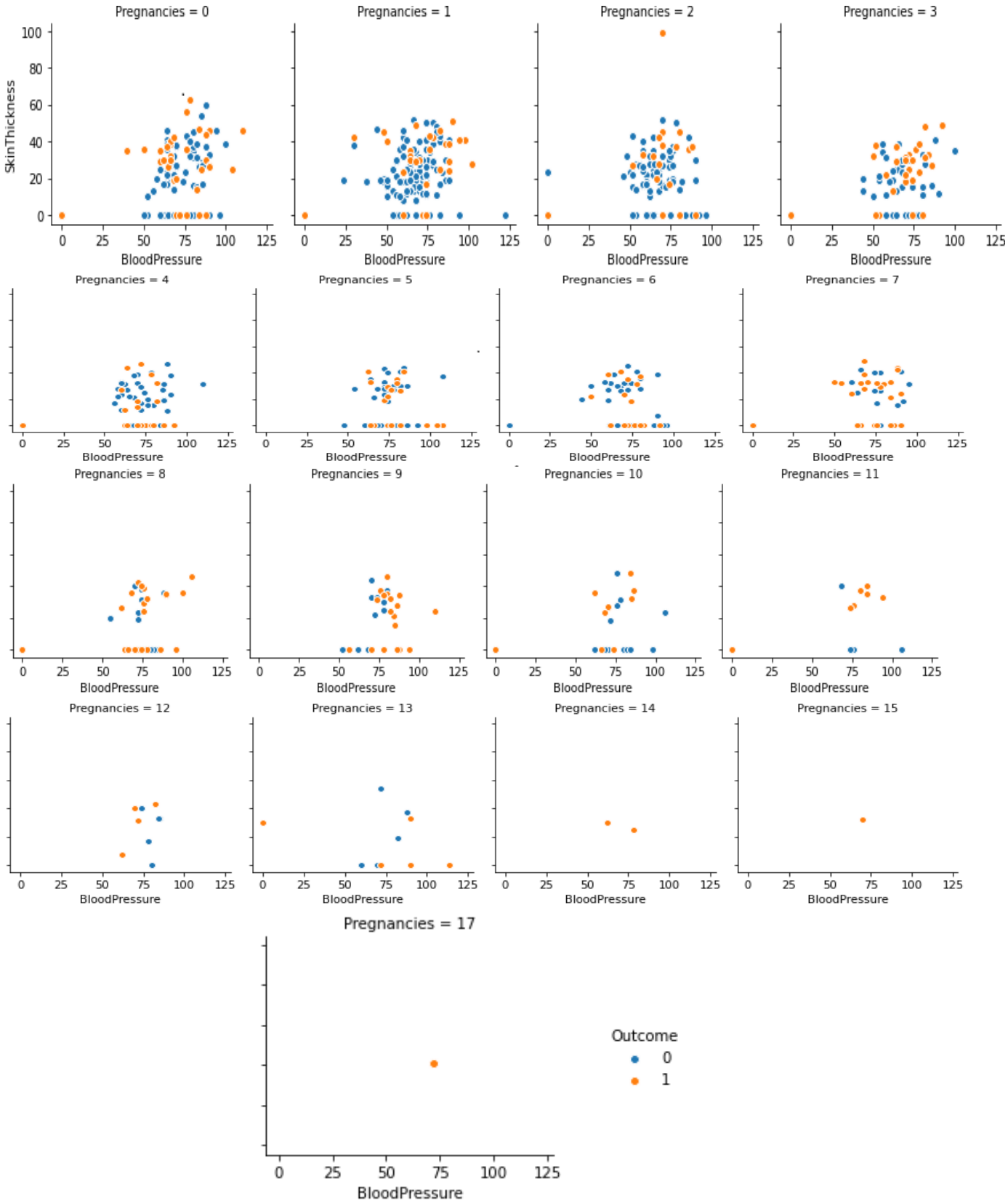

Fig 3: Blood pressure versus pregnancies visualization in two datasets

3. RESULTS AND DISCUSSIONS

The machine learning algorithms such as SVM, MLP, Logistic Regression, and Random forest ensemble algorithms are carried out for Pima Indian Diabetic dataset. The Random Forest algorithm is one of the ensemble classifier in machine learning. It uses decision tress while training the dataset. The confusion or Error matrix is used to calculate various parameters. The Fig 4 shows confusion matrix for Logistic regression. The Table 3 depicts various performance metrics for machine learning algorithms. The proposed method is

suitable for diagnosis of diabetics with 89.8% accuracy, kappa coefficient of 0.72 and F1score of 0.85. The proposed Random forest algorithm outperformed in terms of accuracy, kappa coefficient and F1score.

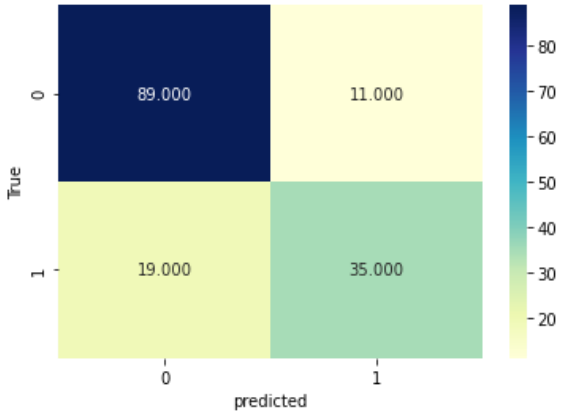

Fig 4: Confusion matrix for Logistic Regression

Table 3: Parameters for machine learning algorithms

Parameter	MLP	SVM	Logistic Regression	Random Forest
Accuracy	80.51%	80.51%	65.8%	89.8%
F1score	0.687	0.70	0.64	0.85
Kappa coefficient	0.549	0.557	0.554	0.72

4. ACKNOWLEDGMENTS

Our sincere thanks to kaggle for providing us Pima Diabetic patient data for the research purpose.

5. REFERENCES

[1] World Health Organization, Global Action Plan on Physical Activity 2018-2030: More Active People for a Healthier World, World Health Organization, Geneva, Switzerland, 2019.

[2] Marx V. Biology: the big challenges of big data. Nature Jun 13 2013;498 (7453): 255–60. http://dx.doi.org/10.1038/498255a. Server/Data Center

[3] Wilson RA, Keil FC. The MIT encyclopaedia of the cognitive sciences. MIT Press; 1999.

[4] R. Williams, S. Karuranga, B. Malanda et al., "Global and regional estimates and projections of diabetes-related health expenditure: results from the international diabetes federation diabetes atlas," Diabetes Research and Clinical Practice, vol. 162, Article ID 108072, 2020.

[5] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," Diabetes Care, vol. 37, no. Supplement 1, pp. S81–S90, 2014.

[6] G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Calibration of minimally invasive continuous glucose monitoring sensors: state-of-the-art and current perspectives," Biosensors, vol. 8, no. 1, 2018.

[7] D. Bruen, C. Delaney, L. Florea, and D. Diamond, "Glucose sensing for diabetes monitoring: recent developments," Sensors, vol. 17, no. 8, 2017.

[8] S. Wadhwa and K. Babber, "Artificial intelligence in health care: predictive analysis on diabetes using machine learning algorithms," in Proceeding of the International Conference on Computational Science and Its Applications, pp. 354–366, Springer, Cagliari, Italy, July 2020.