ANALYTICAL STUDY ON DIABETES PREDICTION USING ML

RAMA BANSAL* & Dr. GAURAV AGGARWAL**

*Ph.D. SCHOLAR, COMPUTER SCIENCE & ENGINEERING DEPARTMENT,

**RESEARCH SUPERVISOR,

JAGANNATH UNIVERSITY, NCR, BAHADURGARH, HARYANA, INDIA

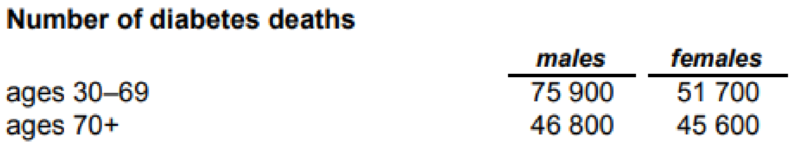
Abstract: Machine learning technique can be utilized to anticipate the diabetes at a beginning phase to benefit society. Now Days Machine learning can foresee any of the issue. The Aim of this study is to analyse and found the impact and relation between various different attributes of diabetes dataset to foresee diabetes disease. Now a Days Diabetes are among one of the leading causes of death in the world. This study will help to build the exhibition and exactness of Prediction of Diabetes Disease utilizing Machine Learning.

Keywords: Machine Learning, Supervised Learning, Unsupervised Learning, Diabetes, Foresee.

I. Introduction

As per the International Diabetes Federation (IDF), in 2020, 463 million individuals were suffered from diabetes in the world. Out of which India had an estimated 77 million individuals with diabetes, that makes India the second most affected country within the world, after China which is affected by diabetes. As per IDF One out of six individuals in the world suffered from diabetes is from India. According to the International Diabetes Federation it will become 134 million by 2045.

As perthe WHO (World Health Organization) report, in 2016, Indian population with age group of [30 to 69], 75, 900 males and 51,700 females died due to Diabetes, and with age group of [70+], 46, 800 males and 45,600 females died due to Diabetes disease.



Number of diabetes deaths

	males	*females*
ages 30–69	75 900	51 700
ages 70+	46 800	45 600

WHO Report (2016)

In Nov 14, 2016, the WHO (The World Health Organization) revealed 422 million adults from all over world were suffered from diabetes and 1.6 million deaths due to it.

As indicated by Above reports and insights we can say that diabetes is an intense and very serious and disease. To save our society and world from the loss of life due to this disease it is very necessary to predict this disease to treat it to reduce death rate due to unawareness of this disease. This study follows machine learning to analysis diabetes disease data set and found the relation between various features. Which will helpful to foresee this disease at a beginning phase to safe human life.

II. RELATED WORK

In recent years, many researchers around the world worked using Machine Learning analytics in medical services and different spaces, to predict or foresee about the future challenges and opportunities about various aspects.

Davidson et. al. (1995) identified that FPG concentrationis usually recommended for those people with one or a lot of risk factorsof diabetes, for the individuals who area unit higher than forty years with any of the subsequent symptomsof parents' history or youngster with diabetes, obesity, African or American, or history of gestational diabetes.

Shanker et. al. (1996) identified that neural networks are proper for foresee the diabetes. the accuracy rateby neural networks is around 78% in training sample and about 81% in the test sample.

Smith et al. (1998) found that After trainingutilizing every one of the 576-trainingset, the ADAP algorithm calculation was placed into a non-learning mode. They found that the sensitivity and specificity were both found around 76 percent.

Chandna et. al. (2014) identified used an approach to diagnose the breast cancer by reduction in the attributes using information gain for Nuralsystem framework.

Saxena et al. (2014) used K- nearest neighbor algorithm for the finding of diabetes mellitus. They have determined accuracy and error rates. The exactness rate is showing that the number of yields of the information of the test dataset are same as the yield of the information of various highlights of the preparation dataset.

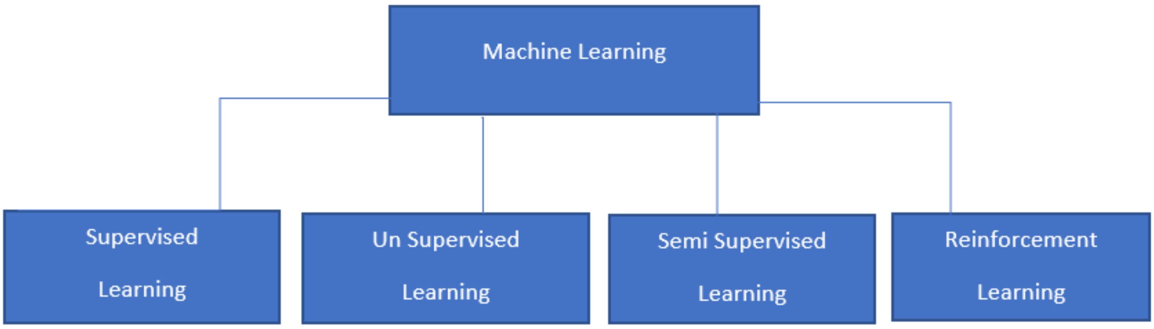
Mounika et al. (2015) utilized three classificationalgorithms utilized in this investigation to be specific, ZeroR, OneR and Naïve Bayes. This investigation shows that Naïve Bayes is the quickest and ZeroR is the slowest.

Simi et al. (2017) explored the importance of early detection of female infertility. Researcher in their research work used 26 attributes and 8 outcomes of female infertility, Results identified that Random Forest algorithmoutflanked other different techniques and gavehigher accuracy.

Sisodia, D.S et. al. (2018) used 3 classification algorithms in their study: Naïve Bayes, Decision Tree, SVM. Their outcomes showed that Naïve Bayes works better compared to the next two algorithms.

III. Research Methodology

Machine Learning is one of thesignificant innovative technology that is currently being utilized in the industry for performing information examination and acquiring understanding into the information.Machine Learning methods, for example, Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Reinforcement Learning gives a pool of tools, libraries and procedures utilizing which crude information can be changed over into some significant information.



Taxonomy of Machine Learning

The Supervised Learningis used to build prescient models. A Supervised model can predict missing databy using other data present in the dataset. Supervised learned model has a bunch of inserted data and furthermore a bunch of output data, and constructs a model to predictaccurate predictions for the result to new data. Supervised learning methods includes Decision Tree, Bayesian Method, Artificial Neural Networks, Ensemble Method,Instance based learning.

In **Unsupervised Learning** we have got betterknown set of inputs however output is unknown. Unsupervised learning is generally used on clustered information. This methodology includes bunch algorithms like k-Means bunch and k-Medians etc.

Semi-supervised Learning uses each labeled and untagged information on training dataset. Classification, Regression techniques comes in Semi supervised Learning. Regression methods go under Semi Supervised Learning. Logistic Regression, Linear Regression are instances of regression strategies.

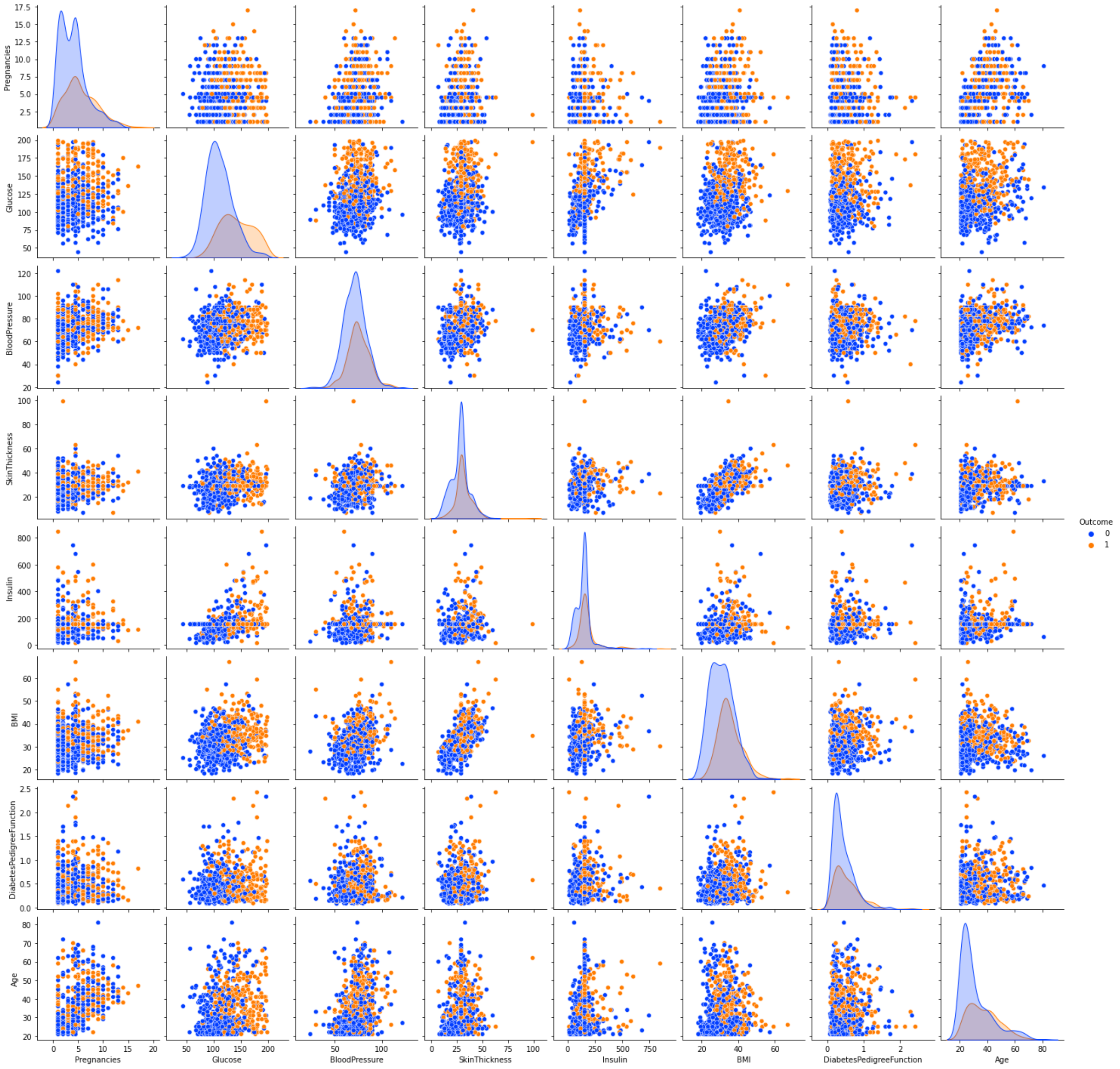
Reinforcement Learning takes feedbacks as input and act and update in code with that feedback. Training for reinforcement focuses on learning from expertise and inconsistencies between unattended and supervised learning. AN agent interacts together with his atmosphere during reinforcement learning setting and gives a reward according to which reinforcement learning program to refine. The agent's goal is to find out the consequences of his actions, like what moves were necessary for winning a game and to use this learning to seek out ways that maximize his rewards. Learning to reinforce could be a style of learning that produces choices supported what actions to require to form the result additional positive. The learner doesn't have any data of what actions to require till a scenario is given. The learner's behaviour will have an effect on future things and their behaviour.

IV. Analysis and Interpretation

This study includes data understanding to study the patterns and trends which helps in prediction and evaluating the results.

i. Pair-wise Relationships

The graphs show pairwise relationships of various attributes. It shows the mapping between various attributes likeGlucose, BMI, Age, Insulin, Pregnancies, Skin Thickness, DiabetesPedigreeFunction and BloodPressure of a dataset onto a column and row in a grid of multiple axes. With the help of platted graphs showing relationship and distribution of data over each and every possibility machine can easily identify the possibility of unknown results.

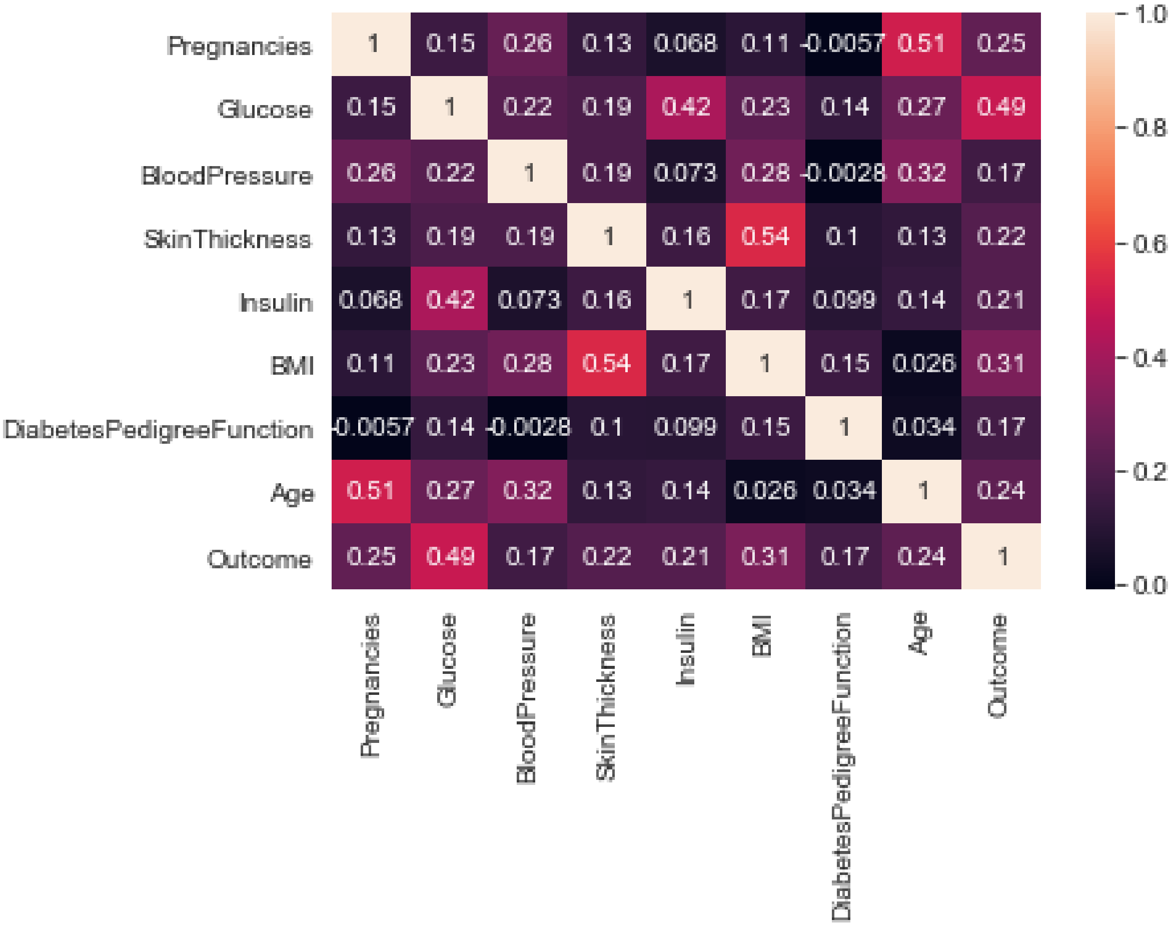


Relation of Attributes

Overlapping in Aboveplotting graphs shows that the mentioned attributes are sufficient enough to get the accurate prediction about diabetes disease. It also shows the relation with various attributes likeGlucose, BMI, Age, Insulin, Pregnancies, Skin Thickness, DiabetesPedigreeFunction and BloodPressure. This study shows the corelation with various attributes. This study shows that Glucose and BMI are two of the main attributes for the prediction of Diabetes based on which we can relate and get the appropriate outcomeusing various Machine Learning Technologies.

ii. Statistical Analysis

This tableis showing the correlation coefficients between variables. Each cell in the table shows the correlation between two variables.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.15	0.26	0.13	0.068	0.11	-0.0057	0.51	0.25
Glucose	0.15	1	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.26	0.22	1	0.19	0.073	0.28	-0.0028	0.32	0.17
SkinThickness	0.13	0.19	0.19	1	0.16	0.54	0.1	0.13	0.22
Insulin	0.068	0.42	0.073	0.16	1	0.17	0.099	0.14	0.21
BMI	0.11	0.23	0.28	0.54	0.17	1	0.15	0.026	0.31
DiabetesPedigreeFunction	-0.0057	0.14	-0.0028	0.1	0.099	0.15	1	0.034	0.17
Age	0.51	0.27	0.32	0.13	0.14	0.026	0.034	1	0.24
Outcome	0.25	0.49	0.17	0.22	0.21	0.31	0.17	0.24	1

Above table shows the corelation between each and every attribute. This study shows that impart factor of Glucose is 49 percentthe prediction of Diabetes, BMI has an impact factor of 31 percent on the prediction, Glucose is highly related to BMI and Insulin. Blood Pressure is highly related to Age, BMI and Glucose level, Insulin is highly related to Glucose. Pregnancies is highly related to Age and SkinThickness is highly related to BMI.DiabetesPedigreeFunction is highly related to BMI factor. This study identified the relationship between various data member elements.

V. Conclusion

This study shows that the corelation with various attributes which shows that Glucose and BMI are two of the main attributes of the Diabetes Dataset. It also shows that Glucose is having an impact of 49 percent on the outcome and the impact of BMI is 31 percent. Plotted graphs show that only one or two attributes are not sufficient enough to get the accurate prediction about diabetes disease. The identified impact and relation of Glucose, BMI, Age, Insulin, Pregnancies, Skin Thickness, DiabetesPedigreeFunction and

BloodPressure on each other will be helpful to foresee the diabetes disease using various Machine Learning to safe human life.

VI. Acknowledgement

I would like to thank the Almighty God and would like to express my great gratitude to my Parents, family members and Supervisor Dr.Gaurav Aggarwal, Associate Professor, Department of Computer Science Engineering, Jagannath University, NCRfor their guidance, support andunending blessings.

VII. References

[1]. International Diabetes Federation.

[2]. Tandon, Nikhil; Anjana, Ranjit M.; Mohan, Viswanathan; Kaur, Tanvir; Afshin, Ashkan; Ong, Kanyin; Mukhopadhyay, Satinath; Thomas, Nihal; Bhatia, Eesh; Krishnan, Anand; Mathur, Prashant, "The increasing burden of diabetes and variations among the states of India: The Global Burden of Disease Study 1990–2016" (2018).

[3]. Kannan, Ramya, "India is home to 77 million diabetics, second highest in the world"(2019).

[4]. Indiaspend, “Indian Diabetics Unaware of Their Condition: Study”, (2019).

[5]. Davidson, Mayer B., Anne L. Peters, and David L. Schriger. "An alternative approach to the diagnosis of diabetes with a review of the literature." Diabetes care 18.7 (1995): 1065-1071.

[6]. Shanker, Murali S. "Using neural networks to predict the onset of diabetes mellitus." Journal of chemical information and computer sciences 36.1 (1996): 35-41.

[7]. Smith, Jack W., et al. "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus." Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association (1988).

[8]. Chandna, Deepali. "Diagnosis of heart disease using data mining algorithm." International Journal of Computer Science and Information Technologies 5.2 (2014): 1678-1680.

[9]. Saxena, Krati, Zubair Khan, and Shefali Singh. "Diagnosis of diabetes mellitus using K nearest neighbor algorithm." International Journal of Computer Science Trends and Technology (IJCST) 2.4 (2014): 36-43.

[10.] Mounika, M., et al. "Predictive analysis of diabetic treatment using classification algorithm." IJCSIT 6 (2015): 2502-2505.

[11]. Lafta et al. (2015) proposed an intelligent recommender system that assists the patients and practitioners about the short-term risk assessment of heart failures.

[12]. R. Lafta, J. Zhang, X. Tao, Y. Li, and V. S. Tseng, “An Intelligent Recommender System Based on Short-Term Risk Prediction for Heart Disease Patients,” in 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, pp. 102–105.

[13]. M. S. Simi, K. S. Nayaki, M. Parameswaran, and S. Sivadasan, “Exploring female infertility using predictive analytic,” in 2017 IEEE Global Humanitarian Technology Conference (GHTC), 2017, pp. 1–6.

[14]. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Computer Sci. 2018; 132:1578–85.