Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 8. No. 2, 2020

Article Citation Format
Folorunsho O.S., Ayinde A.Q., Olagoke M.A. & Ojo T.P (2020): Application of Machine Learning Algorithms to Predict Diabetes in Pregnant Women. Journal of Digital Innovations & Contemp Res. In Science., Engineering & Technology. Vol. 8, No. 2. Pp 167-173
DOI: dx.doi.org/10.22624/AIMS/DIGITAL/V8N2P11

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 11th April, 2020
Review Type: Blind
Final Acceptance: 17th June, 2020
CrossREF Member Listing :: ttps://www.crossref.org/06members/50go-live.html

Application of Machine Learning Algorithms to Predict Diabetes in Pregnant Women

Folorunsho O. S, [1]Ayinde A.Q[2], Olagoke M.A[3] & Ojo T.P[4]
[1]Washington University of Science and Technology Vienna, VA, 22182, USA
[2]Northcentral University, Scottsdale, AZ, 85255, USA
[3] EC-Council University, Albuquerque, NM, 87109, USA
[4]University of Indianapolis, Indianapolis, IN, 46227, USA

ABSTRACT

With the growing numbers of diabetic patients across North America, the United States of America has recorded a higher percentage of people diagnosed with diabetes. The number constitutes 11 percent of the total population in the USA. These numbers spread across various races, ethnicity, gender, and income bracket. In the United States of America, diabetes is common among adults with a family income lower than the federal poverty level. Numerous scholars have developed predicting and classifying models to analyze diabetic data to infer the result generated by their models to support clinical decisions. In this paper, predictive analytics using dynamic machine learning algorithms will detect patterns and relationships from pregnancy data. Instant diagnosis and prediction of diabetes will give patients more time for preventive care and appropriate treatment. The physician used the precision and accuracy of the model to advise patients medically on risk factors and the better way to manage diabetes. The predictive model built in Waikato Environment for Knowledge Analysis (WEKA) was used to analyze pregnancy over 2500 patient data using Naïve Bayes and Decision Tree algorithms. The data served as input data to train the model using these attributes (Pregnancy week, Glucose, Skin Thickness, Blood Pressure, BMI, Age, Insulin, Diabetes Pedigree Score, and Outcome). An incremental data pre-processing method was adopted to remove noisy data. The data was calibrated into 70 percent for training and 30 percent for testing. Decision Tree accuracy and the precision rate is at an average of 80 percent, while Naïve Bayes underperforms because of its inability to learn and identify patterns within the datasets.

Keywords: Pregnancy; Algorithms; CRISP-DM; Patient; WEKA

1. INTRODUCTION

As technology is growing, devices generate large amounts of data daily. There is a global outburst in the availability of data for researchers. The complexity, colossal size, and heterogeneity of data demand one to search discover, and embrace new software tools and means to successfully manage, analyze, and visualize the data [1]. Researchers [2] obtained a review of the literature on big data for more than a decade has shown the importance of machine learning in analyzing big data. Their results show how this field has evolved over the years and the increasing rate of publications in big data. The exponential evolution in big data started in 2010 and thus attracted more researchers.

McKinsey reported that 50 percent of Americans has one or more chronic diseases, and they spend around 80 percent of American medical care fee on treating these chronic diseases [4]. Around three trillion dollars are spent annually on treating those chronic diseases, which is 18 percent of the annual Gross Domestic Product (GDP) of the United States. Big data in the healthcare industry refers to large and complex electronic health datasets for traditional software tools. Healthcare analytics refers to the methodical use of healthcare datasets for business insights, decision-making, planning, learning, early prediction, and detection of diseases by using different statistical, predictive, and quantitative models and strategies.

Current developments in machine learning have radically improved the capability of computers to identify and label images, identify, and translate speech, play games involving skills and higher IQ, predict diseases, and improve decision-making over data. In machine learning applications, the objective is to train a computer to do as humans or better than humans [7]. Traditionally supervised learning algorithms are used for training the model with labeled data, and testing data is used for evaluation using testing data. Earlier studies reported that sex, pregnancy, body mass index (BMI), and metabolic status are associated with diabetes [13,14].

Prediction models can screen pre-diabetes or people with an advanced risk of developing diabetes to help determine the best clinical management for patients [16]. Multiple predictive equations have been proposed to model the risk factors of incident diabetes [15–17]. For instance, researchers [18] presented a tool to predict the risk of diabetes in the US using undiagnosed and pre-diabetes data, and Razavian et al. [19] developed logistic regression-based prediction models for type 2 diabetes occurrence. These have been applied to healthcare models to screen individuals that test positive or are at a high risk of diabetes.

2. METHODOLOGY

The Cross Industry Standard Process for Data Mining- (CRISP-DM) was adopted as the project methodology. The CRISP-DM is divided into six iterative phases. During the business understanding phase, the data points identified we Pregnancy Week, Glucose, Skin Thickness, Blood Pressure, BMI, Age, Insulin, Diabetes Pedigree Score, and outcome.

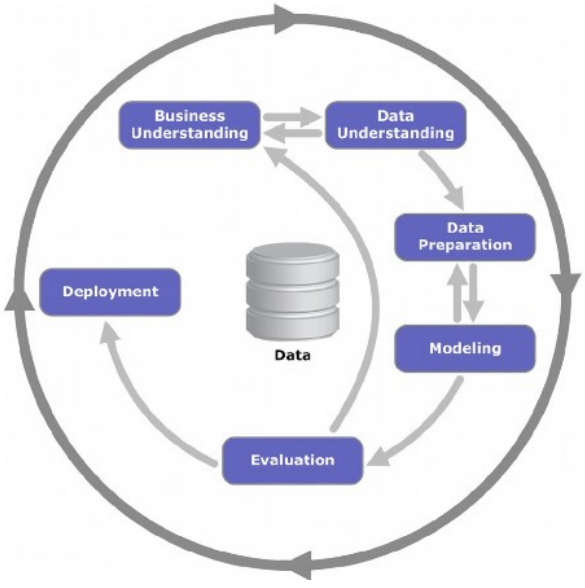

Figure 1: CRISP-DM Model

The outcome is the target variable depending on other data points known as the predicting variables. At the data preparation stage, noisy data were removed, and the data were transformed to an acceptable input into the WEKA machine. Naive Bayes, a supervised learning algorithm, and Decision Tree, an unsupervised learning algorithm, will be selected from WEKA (Waikato Environment for Knowledge Analysis). These algorithms will be applied to predict if a pregnant woman is diabetic. To prevent threat actors from attacking the cloud-based workflow develop in WEKA [8], a third layer security will be added to the workflow during the deployment [9], [10]. The performance of machine learning algorithms in predicting a diabetic patient during pregnancy will be evaluated. The data is partitioned into a training set (comprised of 70 percent) and a test set (comprised of 30 percent). The CRISP-DM is shown in Figure 1.

Business Understanding Phase: During the business understanding phase, the limitations in the review literatures were applied to align the main and objective of this research. Physicians and clinicians were engaged to have a better understanding of the risk factors that were associated with diabetes. These factors were analyzed critically to determine the variables that will be used in this work. The clinicians provided a layman meaning that helped to determine the target variable that other risk factors depend on. An onsite visit to the woman centers to obtain some non-medical data via questionnaire was administered via a Microsoft form. The total of 2500 forms were completed within six months from more than ten hospitals across United States. Data Understanding Phase: The non-medical data gathered from the pregnant women were analyzed to determine the diabetic risk factors that are common among pregnant woman in the United States. The descriptive analysis of the medical data collected from the hospitals is summarized in Figure II. The descriptive analysis is based on the patient age bracket across United States.

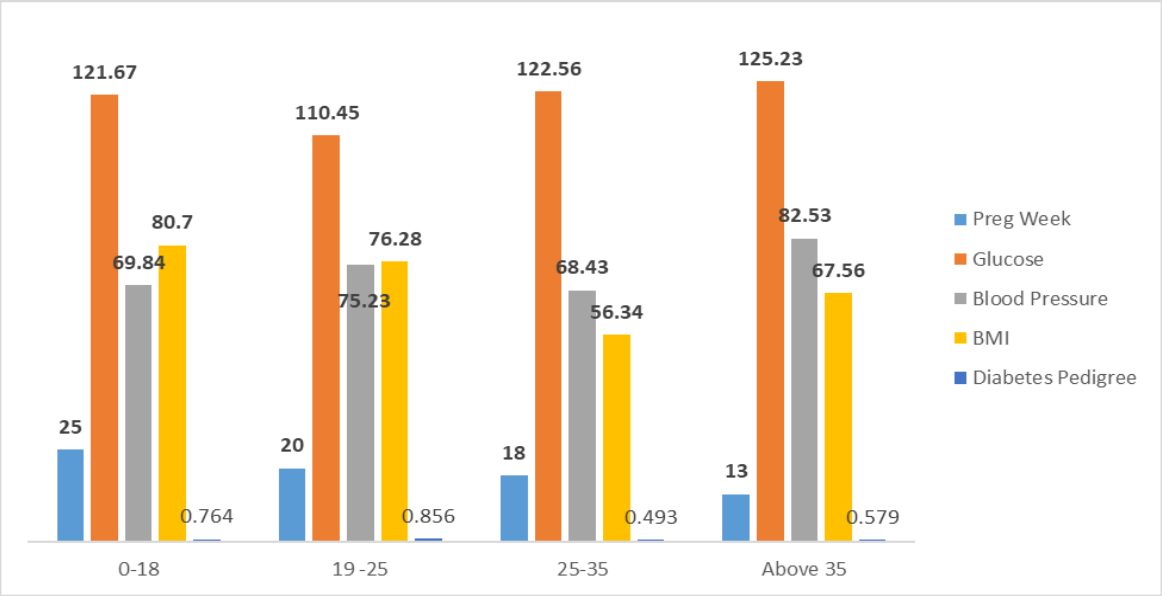

Figure II: Descriptive Analysis

Data Preparation Phase: During the data preparation stage, the insight from the descriptive analysis shown the two prevalent risk factors as the patient glucose level and blood pressure. The chart also revealed the relationship between the two major risk factors and the diabetes pedigree. From this analysis, it was deduced that any patient that has high glucose level and either the blood pressure or BMI is higher will be diabetic. A diabetic patient is coded as 1 and non-diabetic patient is coded as 2. The target variable is the outcome, which is either 0 or 1, the dataset was partitioned into two (70percent training and 30percent testing).

The knowledge base interface of the WEKA was selected and prior to pulling of data into the knowledge environment in WEKA, the data was subjected to a data pre-processing technique to remove noisy data for proper calibration of the model. The pre-processing of data will allow the model to support batch learning, incremental learning, and bagging to improve the precision of the model when processing batch of data once. Modeling Phase: During the modeling phase, the Naïve Bayes and Decision Tree algorithms were selected from the knowledge base environment. The algorithms were trained using the historical data for the machine to learn the patterns and relationships that exist among the datasets. A cross validation of two-folds was selected from the machine to ensure that the algorithms are working perfectly without an error. The batch and incremental learning will enable to machine to process any size of dataset that is fed into it by the modeler. The model is precision and accuracy after loading the 70percent of the dataset into the knowledge environment is 82percent and 92percent respectively. The model was re-calibrated for optimization to increase the precision rate of the model. The Naïve Bayes was re-calibrated to use incremental learning and the Decision Tree was re-trained with batch learning. At the completion of the 30percent of the dataset, the precision and accuracy of the model has improved with excellent True Positive Rate and False Positive Rate for each of the algorithms.

Evaluation Phase: During the evaluation phase, the algorithms were evaluated based on the metrics that estimate the performance of the model. The metrics adopted and applied from the WEKA Knowledge environment are True Positive Rate, False Positive Rate, Precision Rate and Accuracy Rate. Deployment Phase: The result generated from the performance of the algorithms in predicting a diabetic patient as shown in Table 2 reveals that the model is dependable and can analyze the medical variables provided by the hospital to predict if a patient will be diabetic or not.

3. DATA & SUMMARY STATISTICS

Some factors contribute to the diabetes risk factor, including blood pressure, pregnancies for women, age, body mass index, etc. As a component of diabetes control, it would be helpful to know which attributes are related to diabetes. The dataset was collected from a hospital in the Midwestern part of the United States as a case study to predict the risk factors associated with diabetes. The patient's diabetes classification, medical diagnosis regarding laboratory work such as tricep skinfold thickness, the 2-h serum insulin (serum-insulin), number of pregnancies, plasma glucose concentration, body mass index (BMI), diastolic blood pressure, age, and diabetes pedigree function. There are 2500 diabetes patients to be analyzed. The descriptive analysis conducted for a hospital is summarized in Table 1.

Table 1: Patient Data Sample from a Hospital Location

Variable	Description	Mean	Std Dev	Medium
Pregnancy	Pregnancy week	4.65	3.89	4.00
Glucose	Concentration of plasma glucose	145.24	42.68	124.69
Blood Pressure	Diastolic Blood Pressure	67.89	13.79	65.98
Skin Thickness	Skinfold thickness	32.14	7.98	29.26
Insulin	Two-hour serum insulin	146.74	78.41	135.77
Body Mass Index	Body Mass index (kg/m^2)	46.21	7.23	42.29
Pedigree	Degree function of diabetes	0.98	0.23	0.46
Age	Age (Log Years)	37.21	10.56	27.62

4. RESULT DISCUSSION

Classification is used for data mining that allows items in the classes. It comes under the predictive method. The aim of this process is to predict the class for all the predicting attributes related to patient medical data used in this research. The simplest classification issue is the binary classification in which aimed attribute has two possible values, whereas multiclass targets have more than two values. The classification ability of a system is how well it can differentiate one feature from the other. So, there is a need to classify features using Classifiers, so the classification power of the classifiers is well known. In this paper, Bayesian and Naive Bayes classifiers were applied to the dataset to determine which classifier predicts that a pregnant woman is diabetic, considering all the risk factors. A. Bayesian Classifier provides a structural representation of probabilistic relationships between several predicting variables based on the Bayes theorem [20]. The formula below is the Naïve Bayes algorithm, while the target variable A means the patient was diagnosed with diabetes given B (any predicting variables stated in this paper).

$$p(\frac{A}{B}) = p(A)\frac{p(\frac{B}{A})}{p(B)} \qquad (1)$$

p(A) = prior probability of hypothesis A.
p(B) = prior probability of training data B.
p(A/B) = probability of A when B is given.
p(B/A) = probability of B when A is given. A Here A is data and B is hypothesis

Decision Tree
Probability = (x, Y) = (x^1, x^2, x^3,……x^k, xY) (2)
x= occurrence predicting variable in the population
Y= occurrence of target variable in the population

Table 2: Result Table

	True Positive Rate	False Positive Rate	Precision	Accuracy
Naïve Bayes	98.6%	35.9%	98.2%	86.9%
Decision Tree	74.9%	42.8%	89.2%	72%

The model was trained using two-fold cross-validation for testing. It was discovered that Naive Bayes' precision and accuracy are better than Decision Tree simply because it does not support incremental learning when future data are fed into the model. The metrics calibrated in WEKA to measure the performance of the classifiers indicate that Naive Bayes performs better in predicting that a patient is diabetic during pregnancy taking into consideration the risk factors (predicting attributes of the patient). Also, the descriptive prior probability of the target variable and the probability of the target have a probabilistic relationship value of 0.98, supporting the hypothesis that pregnant women can be diabetic. The probability that the developed model will predict that a pregnant woman is diabetic is approximately 1. The accuracy and precision of this model have proven that model can be adopted by hospitals in the United States and worldwide to predict diabetes in pregnant women.

5. CONCLUSION

The Naive Bayes performs best when it was re-trained for incremental and batch learning performs optimally when it was re-calibrated to support batch learning. Considering the number of data generated by hospital daily it is important that an enterprise version of the predictive model is developed as solution-as-a-service so that any pregnant woman from the comfort of their home can use the model to predict if there will be diabetic as the risk factors identified. To improve performance and accuracy of the model, future work should focus on hybridizing the Bayesian network and decision tree classifiers to improve the accuracy and precision of the model. While the performance of the developed model is excellent, the healthcare industry can adopt this model to predict early-stage diabetes in pregnant women, and the methodology can be re-modified to predict diabetes in human beings. Pharmaceutical companies producing insulin might study the pattern in the insulin value and other risk factors to develop a more potent insulin based on ethnicity and age.

REFERENCES

[1] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard and K. Najarian, "Big Data Analytics in Healthcare", *Hindawi Publ. Corp.*, vol. 2015, pp. 1-16, 2015.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong and G.-Z. Yang, "Big Data for Health", *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1193-1208, 2015.

[3] E. Ahmed et al., "The role of big data analytics in Internet of Things", *Comput. Networks*, vol. 129, pp. 459-471, December 2017.

[4] The big-data revolution in US health care: Accelerating value and innovation Mckinsey & Company, [online] Available: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care.

[5] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", *IEEE Access*, vol. 5, no. c, pp. 8869-8879, 2017.

[6] L. Zhou, S. Pan, J. Wang and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges", *Neurocomputing*, vol. 237, pp. 350-361, May 2017.

[7] J. B. Heaton, N. G. Polson and J. H. Witte, "Deep learning for finance: deep portfolios", *Appl. Stoch. Model. Bus. Ind.*, vol. 33, no. 1, pp. 3-12.

[8] A.S and A.Q Ayinde, " Cybersecurity Plan for a Healthcare Cloud-Based Solutions", Journal of Cybersecurity, Journal of Cyber Security 2022, 4(3), 185-188. https://doi.org/10.32604/jcs.2022.035446

[9] O.S Folorunsho, A.Q Ayinde and A.S Yusuf, "Evaluating Risk Level for Complex and Distributed System, International Journal of Engineering Research & Technology, Vol 12 Issue 01, ISSN: 2278-0181

[10] O.S Folorunsho, A.Q Ayinde and A.S Yusuf, "Defensive Controls and Processes for Significant Threats, International Journal of Engineering Research & Technology, Vol 12 Issue 01, ISSN: 2278-0181

[11] O.S Folorunsho, A.Q Ayinde and A.S Yusuf, "Predicting Students' Educational Performance, International Journal of Engineering Research & Technology, Vol 12 Issue 01, ISSN: 2278-0181

[12] Wild, S.; Roglic, G.; Green, A.; Sicree, R.; King, H. Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. Diabetes Care **2004**, 27, 1047–1053.

[13] Engelgau, M.M.; Narayan, K.; Herman, W.H. Screening for type 2 diabetes. Diabetes Care **2000**, 23, 1563–1580.

[14] Rolka, D.B.; Narayan, K.V.; Thompson, T.J.; Goldman, D.; Lindenmayer, J.; Alich, K. Bacall, D.; Benjamin, E.M.; Lamb, B.; Stuart, D.O.; et al. Performance of recommended screening tests for undiagnosed diabetes and dysglycemia. Diabetes Care **2001**,24, 1899–1903.

[15] Schwarz, P.E.; Li, J.; Lindstrom, J.; Tuomilehto, J. Tools for predicting the risk of type 2 diabetes in daily practice. Horm. Metab. Res. **2009**, 41, 86–97.

[16] Yu, W.; Liu, T.; Valdez, R.; Gwinn, M.; Khoury, M.J. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Med. Inform. Decis. Mak. **2010**, 10, 1–7.

[17] Naz, H.; Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J. Diabetes Metab. Disord. **2020**, 19, 391–403.

[18] Heikes, K.E.; Eddy, D.M.; Arondekar, B.; Schlessinger, L. Diabetes Risk Calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes Care **2008**, 31, 1040–1045.

[19] Razavian, N.; Blecker, S.; Schmidt, A.M.; Smith-McLallen, A.; Nigam, S.; Sontag, D. Population-level prediction of type 2 diabetes

[20] Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. Front. Genet.**2018**, 9, 515