Detection of Diabetic Patterns using Supervised Learning

Kalpna Guleria, Devendra Prasad, Virender Kadyan

Abstract: ***World Health Organization's (WHO) report 2018, on diabetes has reported that the number of diabetic cases has increased from one hundred eight million to four hundred twenty-two million from the year 1980. The fact sheet shows that there is a major increase in diabetic cases from 4.7% to 8.5% among adults (18 years of age). Major health hazards caused due to diabetes include kidney function failure, heart disease, blindness, stroke, and lower limb dismembering. This article applies supervised machine learning algorithms on the Pima Indian Diabetic dataset to explore various patterns of risks involved using predictive models. Predictive model construction is based upon supervised machine learning algorithms: Naïve Bayes, Decision Tree, Random Forest, Gradient Boosted Tree, and Tree Ensemble. Further, the analytical patterns about these predictive models have been presented based on various performance parameters which include accuracy, precision, recall, and F-measure.***

Keywords: ***Machine Learning, Supervised Learning, Classification, Bio-informatics, Data Mining***

I. INTRODUCTION

Nowadays, diabetes has become one of the most common diseases. Usually, the cases of type 2 diabetes have been reported either in middle age or in old age people. However, in the recent past, various cases of diabetes have also been reported in children. The pancreas is responsible for the production of insulin in our body. Diabetes prevails if the body is unable to use the produced insulin effectively or the pancreas does not produce the required amount of insulin. Therefore, diabetes is considered a major reason for global concern due to severe health hazards which may lead to hyperglycemia [1]. Hyperglycemia is one of the major causes of diabetic retinopathy, cardiac stroke, foot ulcer, nephropathy, and neuropathy. Hence, it has become of utmost important to draw analytics for the early or on-time detection of diabetes to enhance the quality of life and lifetime enhancement of the patients [2-3].

Latest technological developments in the field of engineering and sciences relates to various machine learning applications which include: speech recognition or natural language processing (NLP), computer vision (facial recognition, pattern recognition, character recognition), Google's self-driving cars, recommender system's (Amazon's product recommendations, Netflix, YouTube), stock market/ housing /finance/ real estate predictions, web search engine optimization, photo tagging, spam classification and biomedical/healthcare sector. Major applications of machine learning in bioinformatics include risk assessment and prediction of cardiac attack, cancer classification, and nephropathic analytics, neuropathic risk assessment [4-5].

Machine learning is a science of experiential learning which draws analytics from past experience and improves the performance of a system through predictive modelling [6]. To draw correct and concise analytics from medical information is the main aim of bioinformatics in medical science. Whereas, a lot of unnecessary tests may complicate the diagnosis process/system and results as well. Hence, machine learning can be used to resolve this difficulty by using various classification algorithms [7].

Machine learning is a branch of Artificial Intelligence that builds up predictive models to draw various statistical analytics. Fig. 1. exhibits various steps to develop a predictive model.

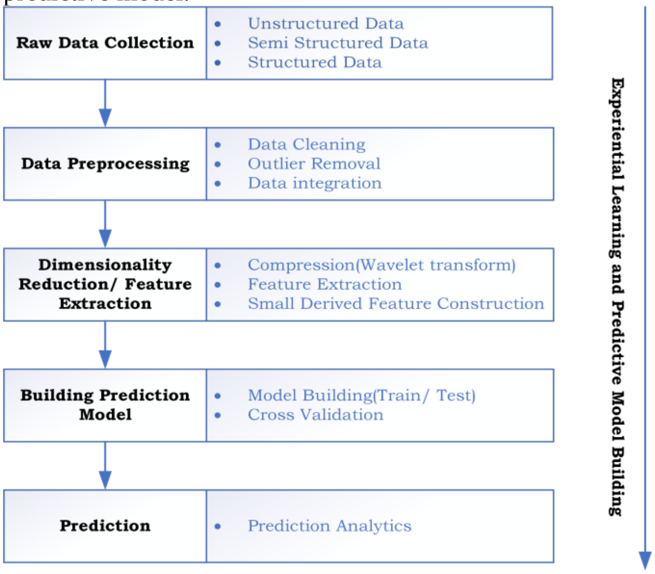
Raw Data Collection
Unstructured Data
Semi Structured Data
Structured Data
Data Preprocessing
Data Cleaning
Outlier Removal
Data integration
Dimensionality Reduction/ Feature Extraction
Compression(Wavelet transform)
Feature Extraction
Small Derived Feature Construction
Building Prediction Model
Model Building(Train/ Test)
Cross Validation
Prediction
Prediction Analytics
Experiential Learning and Predictive Model Building

Fig. 1. Experiential learning and Predictive Model Building Process

The process of learning and predictive model building starts with raw data collection. Data preprocessing focuses on data cleaning (removal of inconsistent and noisy data) and data integration (to combine the different sources of data).

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

Kalpna Guleria, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. kalpna@chitkara.edu.in

Devendra Prasad*, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India devendra.prasad@chitkara.edu.in

Virender Kadyan, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. varinder.kadyan@chitkara.edu.in

Data set may consist of objects whose values do not relate to the other values in the data set or shows the dissimilarity with the general behavioral characteristics of the data.

Such patterns or trends exhibit irrelevant information which is called an outlier and it is treated as noise or exceptions. To obtain the reduced representation of the dataset various dimension reduction techniques can be applied. Dimension reduction can be applied through compression (wavelet transformation), attribute extraction, construction of a smaller set of derived attributes from the existing large set of features. Supervised learning approaches which utilize classification/regression are used for predictive model building by utilizing available preprocessed knowledge base to train/ test the designed model. Supervised learning algorithms utilize various probabilistic or statistical methods to draw analytics from the existing knowledge base/ past experience [8].

Rest of the paper is organized as follows: Section II gives an insight into supervised learning and elaborates important machine learning techniques, section III presents simulations and results, section IV concludes the paper.

II. SUPERVISED LEARNING TECHNIQUES

In this article supervised machine learning techniques have been applied to Pima Indian Diabetic dataset[9] to explore the various trends/patterns about diabetic patients. The performance analytics about various supervised learning classifiers: Naïve Bayes[10-11], Decision Tree[12-13], Random Forest[14-15], Gradient Boosted[16-17], Tree ensemble[18-19] have been drawn to explore diabetes rends in female patients. The presented models apply supervised learning techniques on PIMA Indian Diabetic Dataset to know the trends of various patterns originated from the analytical analysis.

A. Naïve Bayes
B. Decision Tree
C. Random Forest
D. Gradient Boosted Tree
E. Tree Ensemble

A. Naïve Bayes

Naïve Bayes [10-11] classification algorithm utilizes the famous Bayes theorem which assumes independent behavioral characteristics of predictors in a designated dataset. The assumptions in naïve Bayes state that attributes/features in a given class are independent or unrelated to the features /attributes in other classes. The assumption made by naïve Bayes classifier is named as conditional independence. This classifier is applied to build a model where the data set is having a large number of instances. Bayes theorem states that:

$$P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right)P(c)}{P(x)} \qquad \text{(i)}$$

Where
P(c/x) denotes posterior probability.
P(c) denotes class prior to probability
P(x/c) denotes likelihood.
P(x) denotes predictor prior probability.

Naïve Bayes classifier is a quick learner method and performs prediction of a given dataset in a particular class very quickly. It also outperforms in multiclass prediction. It performs better in comparison to logistic regression as it requires a comparatively less training dataset. Naïve Bayes has varied applications which include recommender system, text classification, spam filtering, sentiment analysis, and multiclass prediction.

B. Decision Tree

The decision tree[12-13] follows supervised learning methodology which can be utilized for categorical and continuous data. It performs population splitting into two or multiple homogeneous subsets depending upon the splitter. A decision tree can be either categorical variable based decision tree if the target variable is categorical in nature or it can be a continuous variable based decision tree if the target variable is continuous in nature. Tree's accuracy is affected by the decision where the spitting process will be performed. Regression and classification trees follow different criteria. The sub node decision criteria are to perform a split in a way so that it achieves a higher level of homogeneity at the next level. The major challenge in the decision tree is overfitting. The overfitting avoidance can be done by defining constraints on the construction of tree size and performing tree pruning as well.

C. Random Forest

Random forest[14-15] is an elixir of the problems related to data science. The versatility of random forest lies in its capability of solving problems related to regression as well as classification. The random forest also deals with dimensionality reduction, handling missing values and outliers. Random forest does the construction of multiple trees rather than a single tree. The classification of a particular object is performed based on features or attributes. The random forest classifies an object to a particular class based upon the maximum number of "vote count" made by trees for a particular class. When the random forest is applied for regression problems it calculates the average of values provided by various trees. The beauty of random forest lies in handling large data sets that possess higher data dimensionality.

D. Gradient Boosted

The main idea behind boosting [16-17] is that a weak learner can be further enhanced to learn better. A weak learner's performance is comparatively little higher than random chance. The main idea behind hypothesis boosting is to perform filtering on observations that can be dealt with by the weak learner and to concentrate on those tedious observations which are difficult to handle. It is an important algorithm to build a predictive model. The gradient boosting incorporates three major factors which include optimizing the loss function, predictions to be made by weak learner and minimization of loss function by additive model. The selection of loss function depends upon the type of problems. There are various standard loss functions however, users can construct error function as per the requirement of the problem. Classification problems usually utilize a logarithmic loss function. Gradient boosting uses the decision tree as a weak learner. Trees construction follows a greedy process that selects the split points based upon purity scores which minimize the loss function. Additive model construction is done by adding one tree at a time in the existing model. A gradient descent strategy is followed to achieve t he minimization of loss function through the additive model building by adding trees.

E. Tree Ensemble

Ensemble learning [18-19] is a collection of various basic techniques that provides an optimal classification solution by building a better predictive model.

The ensemble method's basic working principle is to collectively use several decision trees to make better predictions instead of utilizing only a single decision tree. The idea behind the ensemble learning model is to collectively use various weak learners to enhance the performance of predictive learning by creating a strong learner. BAGGing (Bootstrap AGGregation) [20] is utilized to attain reduced variance in the decision tree. The ensemble of various models is constructed by dividing the dataset into various subsets of training data. Ensemble tree utilizes the average of predictions made by various trees. The results attained through the ensemble tree provide better prediction in comparison to a single decision tree.

III. SIMULATIONS

In this article, the predictive model construction is performed by utilizing supervised learning to extract the pattern of diabetic detection from the PIMA Indian diabetic dataset. The target group in Pima Indian Diabetes dataset is female patients having a minimum of 21 years of age. The dataset has been collected from the UCI repository which includes one dependent and various independent variables. The performance analytics about various algorithms: Naïve Bayes, Decision Tree, Random Forest, Gradient Boosted, Tree Ensemble have been presented to extract the pattern of said disease. The dataset is a collection of 768 instances which has two classes namely diabetic and non-diabetic.

The dataset has eight attributes in total which are pregnancy (Number of times, integer type) ranges from 0-17, plasma glucose(mg/dl, real type) ranges from 0-999, Diastolic Blood Pressure(mm/Hg, real type) ranges from 0-122, tricep skinfold (mm real type) ranges from 0-99, Serum insulin(mu U/ml real type) ranges from 0-846, Body Mass Index(Kg/m^2, real type) ranges from 0 to 67.1, diabetes pedigree(real type) ranges from 0.078-2.42, Age(years, integer type) ranges from 21-81 [9]. During the predictive model building, the data set has been partitioned into training and testing sets. Models utilize 10 fold cross-validation to avoid under fitting and overfitting [21]. The simulations have been carried out in KNime 4.0[22]. Different performance parameters evaluated for the various model are accuracy, precision, recall, and F-measure [23].

IV. RESULT AND DISCUSSION

Accuracy depicts the correctness of applied classifiers for the prediction/detection of a particular problem. Accuracy depicts how much correct evaluation our model shows that the patient is diabetic or non-diabetic. Sensitivity (recall) exhibits the fraction of actual positive diabetic cases which have been rightly classified by the classifier as diabetic [TP/(TP+FN)]. Precision shows that how much the proportion of positive diabetic predictions are diabetic or correct. It is also called a positive prediction [TP/(TP+FP)]. Precision is a parameter to depict the quality or exact results (exactness). However, recall is a parameter to depict the quantity or complete results (completeness). Higher precision exhibits that a classifier provides much more relevant results instead of irrelevant. Higher recall exhibit that a classifier provides most of the relevant results

In the confusion matrix, class 0 represent tested negative class whereas class 1 represents tested positive class.

Table- I depicts the confusion matrix for the Naïve Bayes classifier.

Table- I: Confusion Matrix of Naïve Bayes Predicted Values

		0	1	
Actual Values	0	116	39	77.33
	1	34	42	51.85
		74.84	55.26	68.39

The confusion matrix of Naïve Bayes shows that the class precision for the tested negative class is 77.98% whereas precision for the tested positive class is 66.67%. The recall for the tested negative class is 83.78% whereas it is 57.83% for tested positive class. The overall accuracy of this model is 74.45%.

Table- II depicts the confusion matrix for the Decision Tree classifier. The confusion matrix of Decision Tree shows that the class precision for the tested negative class is 77.33% whereas precision for the tested positive class is 51.85%. The recall for the tested negative class is 74.84% whereas it is 55.26% for the tested positive class. The overall accuracy of this model is 68.39%.

Table- II: Confusion Matrix of Decision TreePredicted Values

		0	1	
Actual Values	0	116	39	77.33
	1	34	42	51.85
		74.84	55.26	68.39

Table-III: Confusion Matrix of Random Forest Predicted Values

Actual Values	0	1

0	127	21	77.91
1	36	47	69.12
	85.81	56.63	75.32

Table- III depicts the confusion matrix for the Random Forest classifier. The confusion matrix of Random Forest shows that the class precision for the tested negative class is 77.91% whereas precision for the tested positive class is 69.12%. The recall for the tested negative class is 85.81% whereas it is 56.63% for the tested positive class. The overall accuracy of this model is 75.32%.

Table- IV: Confusion Matrix of Gradient Boosted Predicted Values

		0	1	
Actual Values	0	120	28	76.43
	1	37	46	62.16
		81.11	55.42	71.86

Table-IV depicts the confusion matrix for the Gradient Boosted classifier. The confusion matrix of Gradient Boosted shows that the class precision for the tested negative class is 76.43% whereas precision for the tested positive class is 62.16%. The recall for the tested negative class is 81.11% whereas it is 55.42% for the tested positive class. The overall accuracy of this model is 71.86%.

Table- V depicts the confusion matrix for the Tree Ensemble classifier. The confusion matrix of Tree Ensemble shows that the class precision for the tested negative class is 75.15% whereas precision for the tested positive class is 69.35%. The recall for the tested negative class is 86.99% whereas it is 50.59% for the tested positive class. The overall accuracy of this model is 73.59%.

Table -V: Confusion Matrix of Tree Ensemble Predicted Values

		0	1	
Actual Values	0	127	19	75.15
	1	42	43	69.35
		86.99	50.59	73.59

It is observed from PIMA diabetic dataset that it has 500 instances that represent a non-diabetic class and 268 instances which represent the diabetic class. It shows an imbalanced class. In binary classification where dataset shows imbalanced class accuracy alone cannot represent the performance of the model. F-measure gives a better insight into the performance of a model while dealing with binary classification problems because it provides a good balance between recall and precision. An F-measure value near 1 for a particular classifier represents that model has a better performance. The Naïve Bayes classifier shows F-measure for the tested negative class as 0.8078 and F-measure for the tested positive class is calculated as 0.619. The Random Forest shows F-measure for the tested negative class as 0.81672 and F-measure for the tested positive class is calculated as 0.6225. The Gradient Boosted shows F-measure for the tested negative class as 0.7868 and F-measure for the tested positive class is calculated as 0.5859. The Decision Tree shows F-measure for the tested negative class as 0.7606 and F-measure for the tested positive class is calculated as 0.5350. The Tree Ensemble shows F-measure for the tested negative class as 0.81672 and F-measure for the tested positive class is calculated as 0 .6225.

V. CONCLUSION

This article presents detection of diabetic patterns using supervised learning using Pima Indian Diabetic dataset. Naïve Bayes, Decision Tree, Random Forest, Gradient Boosted, Tree Ensemble methods have been used for building the predictive models for diabetic detection. Accuracy depicts the correctness of applied classifiers for the prediction/detection of a particular problem. It has been observed that the random forest exhibits the highest accuracy among all which is 75.32%. Naïve Bayes classifier also shows very competitive results in comparison to the random forest for accuracy which is 74.45%. The decision tree classifier shows 68.39% accuracy, which is the lowest of all. The diabetic dataset shows an imbalanced class. Therefore, apart from accuracy F-measure is an important parameter to be taken into consideration when the dataset shows an imbalanced class. An F-measure value near 1 for a particular classifier represents that the model exhibits better performance. The Random Forest exhibits the highest value of F-measure as 0.81672 and 0.6225 for tested negative and tested positive class respectively. The Naïve Bayes classifier shows the second highest values of F-measure as 0.8078 and 0.619 for tested negative and tested positive class respectively. However, decision trees show the lowest values for F-measure as well. The versatility of random forest lies in its strength to handle the dataset which has higher dimensionality and it also deals well with imbalanced class data values. Therefore, random forest exhibits higher accuracy and F-measure for building the predictive model for diabetic detection among all applied classifiers.

REFERENCES

1. Mamykina, L., Heitkemper, E. M., Smaldone, A. M., Kukafka, R., Cole-Lewis, H. J., Davidson, P. G., Mynatt, E.D., Cassells, A., Tobin, J.N.& Hripcsak, G. (2017). Personal discovery in diabetes self-management: discovering cause and effect using self-monitoring data. *Journal of biomedical informatics*, *76*, 1-8.
2. Papatheodorou, K., Banach, M., Edmonds, M., Papanas, N., & Papazoglou, D. (2015). Complications of diabetes. *Journal of diabetes research*, *2015*.
3. Soumya, D., & Srilatha, B. (2011). Late stage complications of diabetes and insulin resistance. *J Diabetes Metab*, *2*(9), 1000167.
4. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Jama*, *319*(13), 1317-1318.
5. Babič, F., Majnarić, L., Lukáčová, A., Paralič, J., & Holzinger, A. (2014, September). On patient's characteristics extraction for metabolic syndrome diagnosis: predictive modelling based on machine learning. In *International Conference on Information Technology in Bio-and Medical Informatics* (pp. 118-132). Springer, Cham.
6. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, 61-69.
7. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159-190.
8. Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). IEEE.
9. Pima indians diabetes database. "https://www.kaggle.com/uciml/pima-indians-diabetes-database" (July 2019), (Accessed on 12/07/2019)
10. Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007, November). Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)* (pp. 1541-1546). IEEE.
11. Sebe, N., Lew, M. S., Cohen, I., Garg, A., & Huang, T. S. (2002, August). Emotion recognition using a cauchy naive bayes classifier. In *Object recognition supported by user interaction for service robots* (Vol. 1, pp. 17-20). IEEE.
12. Ming, H., Wenying, N., & Xu, L. (2009, June). An improved decision tree classification algorithm based on ID3 and the application in score analysis. In *2009 Chinese Control and Decision Conference* (pp. 1876-1879). IEEE.
13. Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.
14. Van Essen, B., Macaraeg, C., Gokhale, M., & Prenger, R. (2012, April). Accelerating a random forest classifier: Multi-core, GP-GPU, or FPGA?. In *2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines* (pp. 232-239). IEEE.
15. Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012, July). How many trees in a random forest?. In *International workshop on machine learning and data mining in pattern recognition* (pp. 154-168). Springer, Berlin, Heidelberg.
16. Ye, J., Chow, J. H., Chen, J., & Zheng, Z. (2009, November). Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 2061-2064). ACM.
17. Zhao, Q., Shi, Y., & Hong, L. (2017, April). Gb-cent: Gradient boosted categorical embedding and numerical trees. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1311-1319). International World Wide Web Conferences Steering Committee.
18. Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2006). A comparison of decision tree ensemble creation techniques. *IEEE transactions on pattern analysis and machine intelligence*, *29*(1), 173-180.
19. Xu, Y., Cao, X., & Qiao, H. (2010). An efficient tree classifier ensemble-based approach for pedestrian detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *41*(1), 107-117.
20. Fischer, B., & Buhmann, J. M. (2003). Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(11), 1411-1415.
21. van der Aalst, W. M. (2011, June). On the representational bias in process mining. In *2011 IEEE 20th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (pp. 2-7). IEEE
22. Morent, D., Stathatos, K., Lin, W. C., & Berthold, M. (2011). Comprehensive PMML preprocessing in KNIME. In *the 2011 workshop* (pp. 28-31).
23. Gunawardana, A., & Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, *10*(Dec), 2935-2962.

AUTHORS PROFILE

Kalpna Guleria, is an Associate Professor in Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. She is PhD in Computer Science Engineering from Thapar University, Patiala, India. She has received her M. Tech and B. Tech. with honors, majoring in Computer Science Engineering. She has research publications in SCI - Indexed International journals of high repute. Her research interests include Wireless Sensor Networks, Ad hoc Networks, Swarm Optimization techniques and Machine Learning. She is also an active reviewer of various International journals of high repute.

Devendra Prasad, is a Professor in Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. He has received his B.E. (Computer Science & Engineering) degree from Kumaon University, Nainital, India in 1995, M.Tech (Computer Science & Engineering) degree from Kurukshetra University, Kurukshetra, India in 2007 and PhD (Computer Science & Engineering) from M.M. University, Haryana, India in 2011. He has supervised several M. Tech and PhD students. He has research publications in SCI - Indexed International journals of high repute. His research interest includes network security and Fault tolerant mobile Ad-hoc, wireless sensor networks and machine learning.

Virender Kadyan, is an Assistant Professor in Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. He leads the Speech and Multimodal Laboratory on Sign language, Speech Signal Processing and Natural Language Processing. He has been the member of a research project at national level and technical program committee member at International conferences. He is a Co-Principal investigator on IEEE Project for development of Punjabi ASR system. He has research publications in SCI - Indexed International journals of high repute. His research interest includes speech analysis, recognition, synthesis, pattern matching and machine learning.