Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 132 (2018) 1578–1585

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

Prediction of Diabetes using Classification Algorithms

Deepti Sisodia[a], Dilip Singh Sisodia[b]

[a]National Institute of Technology, G.E Road, Raipur and 492001, India
[b]National Institute of Technology, G.E Road, Raipur and 492001, India

Abstract

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show Naive Bayes outperforms with the highest accuracy of 76.30% comparatively other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

Keywords: Diabetes; SVM; Naive Byes; Decision Tree; Accuracy; Machine Learning ;

1. Introduction:

Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, Intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if diabetes remains untreated. Some of the severe complications include diabetic ketoacidosis and nonketotic hyperosmolar coma [14]. Diabetes is examined as a vital serious health matter during which the measure of sugar substance cannot

* Deepti Sisodia. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: dsisodia.phd2017.cse@nitrr.ac.in

10.1016/j.procs.2018.05.122

be controlled. Diabetes is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. The early identification is the only remedy to stay away from the complications [29].

Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms [1],[12],[6] works better in diagnosing different diseases. Data Mining [15], [2] and Machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study [8].

This research work focuses on pregnant women suffering from diabetes. In this work, Naive Bayes, SVM, and Decision Tree machine learning classification algorithms are used and evaluated on the PIDD dataset to find the prediction of diabetes in a patient. Experimental performance of all the three algorithms are compared on various measures and achieved good accuracy [11].

The remaining of the research discussion is organized as follows: Section-II briefs Related Work of various classification techniques for prediction of diabetes, Section-III describes the Methodology and brief discussion of Dataset used, Section-IV discusses evaluated Results, and Section-V determines the Conclusion of the research work.

2. Related Work:

Sajida et al. in [20] discusses the role of Adaboost and Bagging ensemble machine learning methods [18] using J48 decision tree as the basis for classifying the Diabetes Mellitus and patients as diabetic or non diabetic, based on diabetes risk factors. Results achieved after the experiment proves that, Adaboost machine learning ensemble technique outperforms well comparatively bagging as well as a J48 decision tree.

Orabi et al. in [19] designed a system for diabetes prediction, whose main aim is the prediction of diabetes a candidate is suffering at a particular age. The proposed system is designed based on the concept of machine learning, by applying decision tree. Obtained results were satisfactory as the designed system works well in predicting the diabetes incidents at a particular age, with higher accuracy using Decision tree[22], [7].

Pradhan et al in [4] used Genetic programming (GP) for the training and testing of the database for prediction of diabetes by employing Diabetes data set which is sourced from UCI repository. Results achieved using Genetic Programming [25], [21] gives optimal accuracy as compared to other implemented techniques. There can be significant improve in accuracy by taking less time for classifier generation. It proves to be useful for diabetes prediction at low cost.

Rashid et al. in [28] designed a prediction model with two sub-modules to predict diabetes-chronic disease. ANN (Artificial Neural Network) is used in the first module and FBS (Fasting Blood Sugar) is used in the second module. Decision Tree (DT)[10] is used to detect the symptoms of diabetes on patient¢¢s health.

Nongyao et al. in [17] applied an algorithm which classifies the risk of diabetes mellitus. To fulfill the objective author has employed four following renowned machine learning classification methods namely Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes. For improving the robustness of designed model Bagging and Boosting techniques are used. Experimentation results shows the Random Forest algorithm gives optimum results among all the algorithms employed.

3. Methodology Used:

3.1. Model Diagram:

Proposed procedure is summarized in figure-1 below in the form of model diagram. The figure shows the flow of the research conducted in constructing the model.

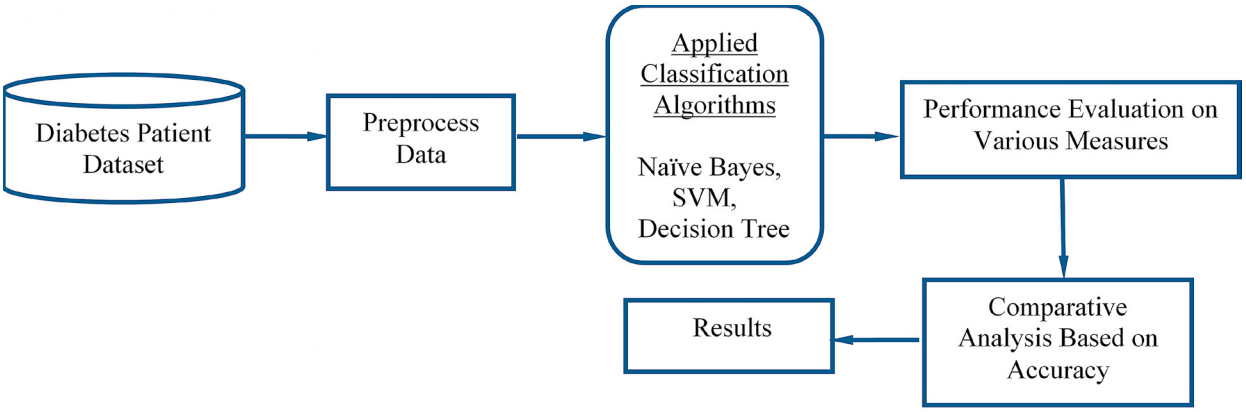

Fig. 1. Proposed Model Diagram

3.2. Brief Description of Algorithms Used:

3.2.1. Support Vector Machine (SVM):

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes[26]. For better generalization hyperplane should not lies closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors [27].

The Accuracy of the experiment is evaluated using WEKA interface. The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyperplane which is defined by $w^T x + b = -1$ and the hyperplane defined by $w^T x + b = 1$ This distance is equal to $\frac{2}{\|w\|.}$ This means we want to solve max $\frac{2}{\|w\|.}$ Equivalently we want min $\frac{\|w\|\|}{2.}$ The SVM should also correctly classify all x(i), which means $y^i(w^T x^i + b) >= 1$, $\forall i \in \{1, ¢¢, N\}$. The evaluated performance of SVM algorithm for prediction of Diabetes [16], [30] using Confusion Matrix is as follows:

Table 1. Confusion Matrix of SVM

	A	B
A-Tested Negative	500	0
B-Tested Positive	268	0

3.2.2. Naive Bayes Classifier:

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with imbalancing problems and missing values. Naive Bayes [24] is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability $P(C|X)$ can be calculated from P(C),P(X) and $P(X|C)$ [23]. Therefore, $P(C|X) = (P(X|C)\ \mathrm{P(C)})/P(X)$

Where,

$P(C|X)$ = target class's posterior probability .

$P(X|C)$ = predictor class's probability.

P(C) = class C's probability being true.

P(X) = predictor's prior probability. The evaluated performance of Naive Bayes algorithm using Confusion Matrix is as follows:

Table 2. Confusion Matrix of Naive Bayes

	A	B
A-Tested Negative	422	78
B-Tested Positive	104	164

3.2.3. Decision Tree Classifier:

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes [11]. The evaluated performance of Decision Tree technique using Confusion Matrix is as follows:

Table 3. Confusion Matrix of Decision Tree

	A	B
A-Tested Negative	407	93
B-Tested Positive	108	160

3.3. Dataset Used:

In this work WEKA tool [3], [9] is used for performing the experiment. WEKA is a software which is designed in the country New Zealand by University of Waikato, which includes a collection of various machine learning methods for data classification, clustering, regression, visualization etc. One of the biggest advantages of using WEKA is that it can be personalized according to the requirements. The main aim of this study is the prediction of the patient affected by diabetes using the WEKA tool by using the medical database PIDD. Table-4 shows a brief description of the dataset.

Table 4. Dataset Description

Database	No. of Attributes	No. of Instances
PIDD	8	768

PIDD-Pima Indians Diabetes Dataset

The proposed methodology is evaluated on Diabetes Dataset namely (PIDD) [13], which is taken from UCI Repository. This dataset comprises of medical detail of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class '0' treated as tested negative for diabetes and value of another class '1' is treated as tested positive for diabetes. Dataset description is defined by Table-4 and the Table-5 represents Attributes descriptions.

3.4. Accuracy Measures:

Naive Bayes, SVM and Decision Tree algorithms are used in this research work. Experiments are performed using internal cross-validation 10-folds. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve)measures are used for the classification of this work. Table-6 defines accuracy measures below:

Table 5. Attribute Description [5]

Attribute	Abbreviation of Attributes
1. Number of times pregnant	pr
2. Plasma glucose concentration	pl
3. Diastolic blood pressure (mm Hg)	pr
4. Skin fold thickness (mm)	sk
5. 2-Hour serum insulin (mu U/ml)	in
6. BMI ($weightinkg/(heightinm)^2$)	ma
7. Diabetes pedigree function	pe
8. Age in years	ag
9. Class '0' or '1'	cl

Table 6. Accuracy Measures

Measures	**Definitions**	**Formula**
1. Accuracy (A)	Accuracy determines the accuracy of the algorithm in predicting instances.	A=(TP+TN) / (Total no of samples)
2. Precision (P)	Classifier¢¢s correctness/accuracy is measured by Precision.	P = TP / (TP+ FP)
3. Recall (R)	To measure the classifier¢¢s completeness or sensitivity, Recall is used.	R =TP / (TP+FN)
4. F-Measure	F-Measure is the weighted average of precision and recall.	F=2*(P*R) / (P+R)
5. ROC	ROC(Receiver Operating Curve) curves are used to compare the usefulness of tests.	

Table 7. Comparative Performance of Classification Algorithms on Various Measures.

Classification Algorithms	Precision	Recall	F-Measure	Accuracy %	ROC
Naive Bayes	0.759	0.763	0.760	76.30	0.819
SVM	0.424	0.651	0.513	65.10	0.500
Decision Tree	0.735	0.738	0.736	73.82	0.751

Corresponding classifiers performance over Accuracy, Precision, F-measure, Recall and ROC values are listed in Table-7 and classifiers performance on the basis of classified instances are defined in Table-8.

Where, TP defines True Positive, TN defines True Negative, FP defines False positive, FN defines False Negative. The corresponding classifiers performance on the basis of Accuracy, Precision, F-measure, Recall and ROC values are listed in Table-7 and classifier's performance on the basis of classified instances are shown in Table-8.

4. Results:

Table-7 represents different performance values of all classification algorithms calculated on various measures. From Table-7 it is analyzed that Naive Bayes showing the maximum accuracy. So the Naive Bayes machine learning classifier can predict the chances of diabetes with more accuracy as compared to other classifiers. Performances of all classifier's based on various measures are plotted via a graph in Figure-2. Figure-3 represents ROC area of all classification algorithms.

Table 8. Classifier's Performance on The Basis of Classified Instances

Total no of instances	Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
768	Naive Bayes	586	182
	SVM	500	268
	Decision Tree	567	201

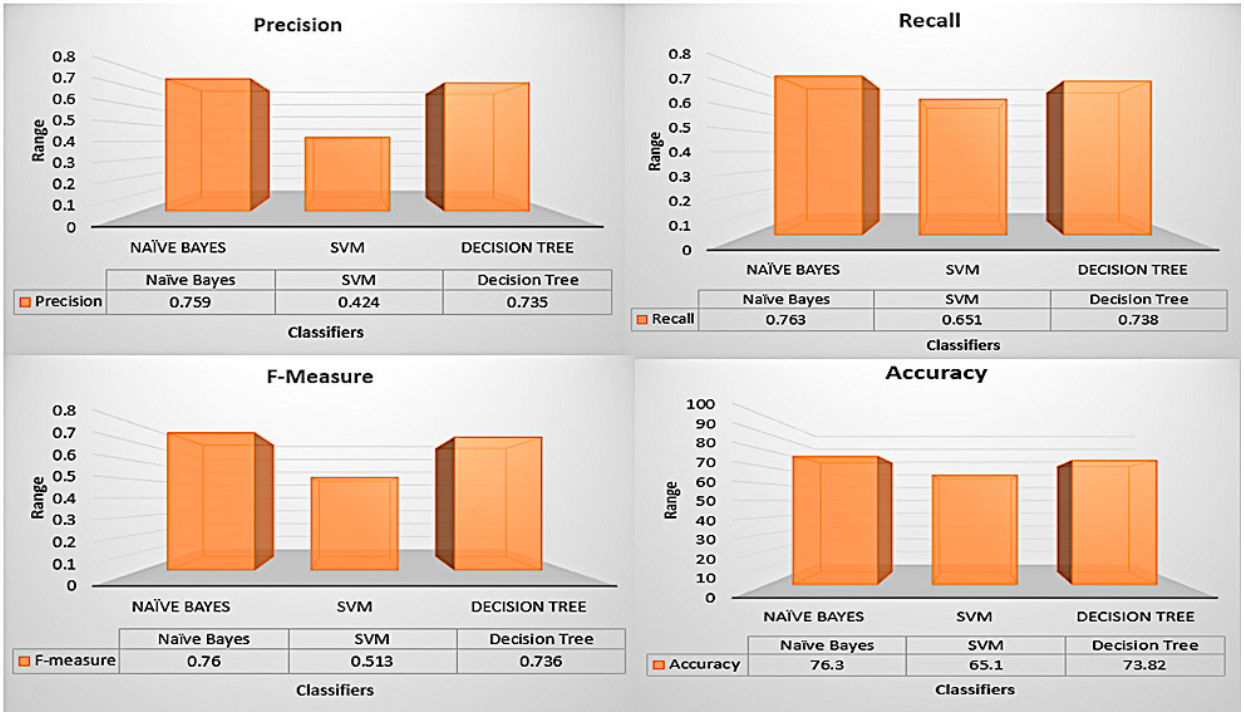

Fig. 2. Classifier¢¢s Performance on Various Measures

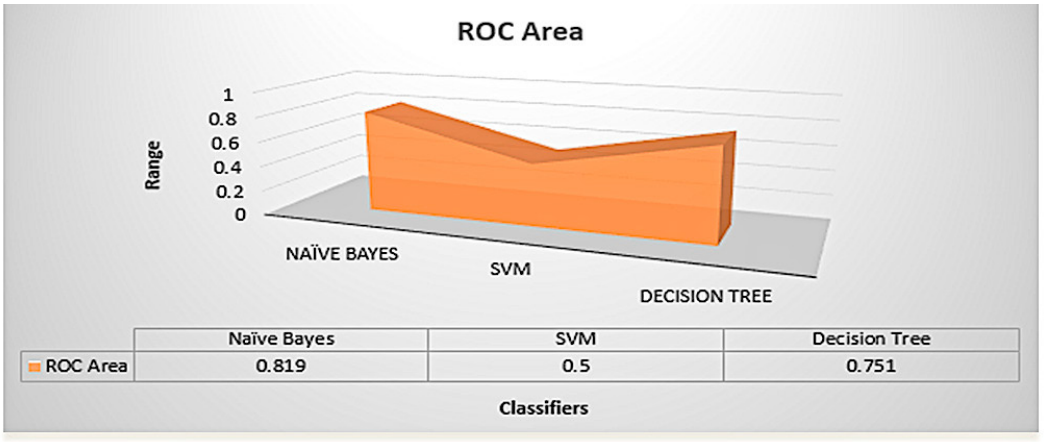

Fig. 3. ROC Area of all Classification Algorithms

Table-8 determines classifier¢¢s performance on the basis of classified instances. According to these classified instances, accuracy is calculated and analyzed. Performance of individual algorithm is evaluated on the basis of Correctly Classified Instances and Incorrectly Classified Instances out of a total number of instances. Figure-4 shows the graphical performance of all classification algorithms on the basis of classified instances. From Table-7 and Table-8 we can conclude that Naive Bayes classification algorithm outperforms comparatively other algorithms. So, Naive Bayes algorithm is considered as the best supervised machine learning method of this experiment because it gives higher accuracy in respective to other classification algorithms with an accuracy of 76.30 %.

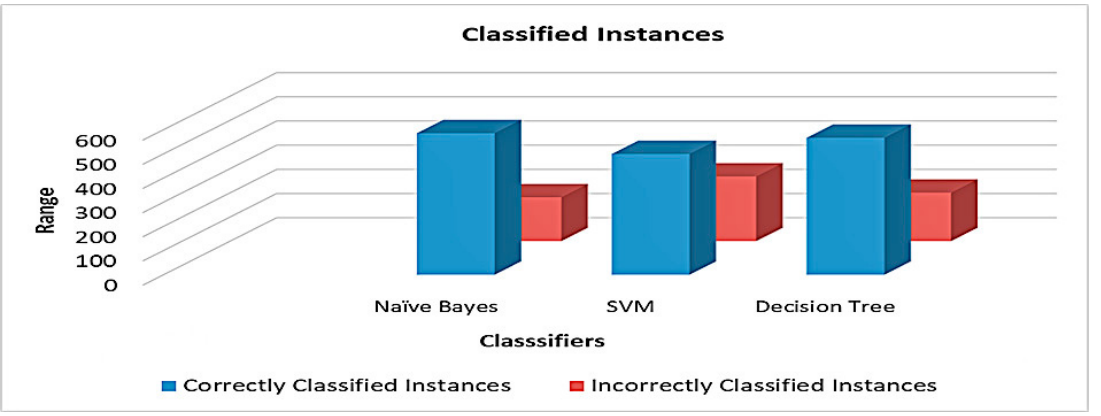

Fig. 4. Classified Instances

5. Conclusion

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

References

[1] Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[2] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[3] Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[4] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[5] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.

[6] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

[7] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

[8] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.

[9] Garner, S.R., 1995. Weka: The Waikato Environment for Knowledge Analysis, in: Proceedings of the New Zealand computer science research students conference, Citeseer. pp. 57–64.

[10] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109.

[11] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.

[12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116. doi:10.1016/j.csbj.2016.12.005.

[13] Kayaer, K., Tulay, 2003. Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP), pp. 181–184.
[14] Kumar, D.A., Govindasamy, R., 2015. Performance and Evaluation of Classification Data Mining Techniques in Diabetes. International Journal of Computer Science and Information Technologies, 6, 1312–1319.
[15] Kumar, P.S., Umatejaswi, V., 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709.
[16] Kumari, V.A., Chitra, R., 2013. Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications (IJERA) www.ijera.com 3, 1797–1801.
[17] Nai-Arun, N., Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014.
[18] Nai-Arun, N., Sittidech, P., 2014. Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931 - 932, 1427–1431. doi:10.4028/www.scientific.net/AMR.931-932.1427.
[19] Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427.
[20] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016.
[21] Pradhan, P.M.A., Bamnote, G.R., Tribhuvan, V., Jadhav, K., Chabukswar, V., Dhobale, V., 2012. A Genetic Programming Approach for Detection of Diabetes. International Journal Of Computational Engineering Research 2, 91–94.
[22] Priyam, A., Gupta, R., Rathee, A., Srivastava, S., 2013. Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and Technology Vol.3, 334–337. doi:JUNE 2013, arXiv:ISSN 2277 - 4106.
[23] Ray, S., 2017. 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python).
[24] Rish, I., 2001. An empirical study of the naive Bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM. pp. 41–46.
[25] Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.
[26] Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 , 554–559doi:10.1109/CICN.2010.109.
[27] Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.
[28] Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract, 2016. An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424, 323–335. doi:10.1007/978-3-319-28031-8.
[29] Vijayan, V.V., Anjali, C., 2015. Prediction and diagnosis of diabetes mellitus A machine learning approach. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) , 122–127doi:10.1109/RAICS.2015.7488400.
[30] Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J., 2010. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 10. doi:10.1186/1472-6947-10-16, arXiv:arXiv:1011.1669v3.