International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 4 June 2020

Supervised Approach of Machine Learning in Medical Diagnosis

[1]Vimal Kumar Awasthi, [2]Ayush Mishra

[1]Assistant Professor, [2]Scholar, [1,2]Department of Computer Science and Engineering,
Kanpur Institute of Technology, Kanpur, India

Abstract: The medical diagnosis for any disease requires a lot of information if it is diagnosed on the basis of the information provided. This prediction task is a very complex task and therefore the expert system that performs the diagnosis needs to use the best and most suitable Machine Learning algorithm to implement it. This research paper provides the approach in which the supervised Machine Learning technique will be used to perform the classification based diagnosis task. In this paper the primarily available supervised learning based classification algorithms like Naïve Bayes, SVM and Decision tree will be fed with the same data of diabetic patients and their accuracy will be tested. This analysis of algorithms of supervised learning will provide a glimpse of the best algorithm to handle and predict the decision with better accuracy.

Keywords: Supervised Machine Learning Technique, naïve Bayes, Decision Tree, SVM, Diabetic Patient Dataset

1. INTRODUCTION

With the increased application of digital concepts and features, researchers and Machine Learning scholars are a bit more involved in the researches related to Medical diagnosis using these digital features. The reason of getting into field of Machine Learning for the medical diagnosis is also primarily dependent on the type of diseases that can be predicted by the details of the patient available as it was predicted by the doctors on the basis of symptoms and learnings of past. These types of diseases provided a concept of using supervised learning to get a basic prediction that should be more accurate and precised on the basis of a huge learning using the knowledge base.

This approach of supervised learning provides a better classification of the disease and the accuracy can be enhanced by the better availability of the training data. This paper focuses on the usability of Machine Learning approach in such a way to classify the disease as per the data provided against various algorithms available. The different algorithms like naïve bayes, SVM and Decision tree will be the main focus to be tested and their accuracy will be compared against the diabetic patient data available.

2. COMPARISION PROBLEM

The comparing of different classifiers sometimes seems to be an easy task as it only requires a basic error counting type logic. This comparison but requires two primary problem in it: The application of that classifier and the user traffic for it. This analysis of primary problems is to be taken into account for the classifier we are considering for the comparison. The comparison also has a major issue when the classifier has various tuning variables that can be tuned as per the requirement. Hence the comparison task becomes a bit complicated to be done without considering some basic standard of those modifiable variables.

In this paper we are considering the supervised algorithms like naïve bayes, SVM, Decision Tree with their standard variable parameter to perform the comparison between these algorithms. These standard parameters are taken into account for these algorithms such that they can provide the basic support to algorithm taking the input data feed as random. These standard parameters didn't have any prior approach to the input data and hence can be easily mapped as the average case for all the classifiers. To consider the best case it might possible that some algorithm may work better whereas some may work worst hence we are considering the average case and taking the standard parameters into consideration for algorithms/classifiers.

3. APPROACH OF THIS PAPER

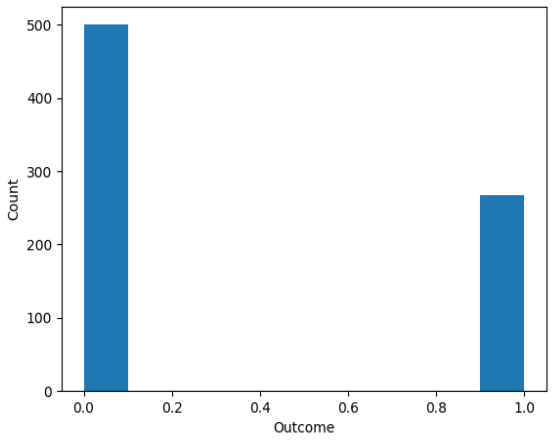
To compare some sort of classification algorithm we require a dataset which can be fed to these algorithms for the task. It will be a nightmare to compare any algorithm without using it to some data and here we will test these algorithms with the medical diagnosis data of diabetic

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 4 June 2020

patient that has well formatted features. This dataset is also in a standard format in quantity and quality both therefore it can easily be used to perform an average case comparison between these classification algorithms. In this paper we are only performing the analysis for the three primary classification algorithms I.e. Naïve Bayes, SVM, Decision Tree. The basic need here in this paper is the dataset and algorithm itself.

4. DATA SET

For any algorithm to provide a better accuracy and prediction the most important requirement is a dataset. In this paper we are considering the diabetic patient dataset for performing the comparison task. This dataset is taken from an online platform kaggle and the data is already in a format that can easily be used to perform the feeding task. The data is divided into feature set and the outcome of each entry. This data has a total of 8 features of the patient and one outcome info in form of integer as 0 and 1 where 0 stands for diabetic negative and 1 stands for diabetic positive.



The data is shown here in a histogram format that shows the count of diabetic positive (outcome = 1) and diabetic negative (outcome = 0). This data has nearly 500 diabetes negative patient and approx 300 diabetes positive patients are there. This dataset is a common dataset for male and female hence no any variation according to gender is considered here. The feature set consist of basic 8 features that are used here to perform the algorithm training process and these features are:

Pregnancies: This feature consist of details that how much times the person has been pregnant and for gender other than females, have this count as zero.

Glucose: The glucose count of human body is taken into account and here it is treated as a feature. This is provided here in number values.

Blood Pressure: Blood pressure count is also considered here and it is also provided in a number format for each patient.

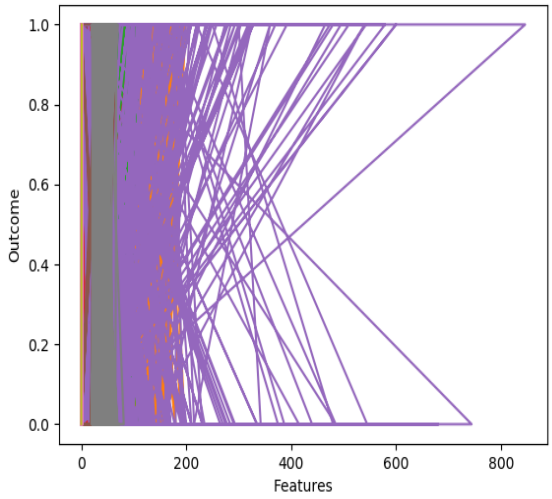
Skin Thickness: This is the measurement of thickness of the skin of patient and it helps to understand the effect of glucose count in body and hence is considered in features.

Insulin: The insulin count is provided in this feature that helps to find the pattern to sustain the understandability of diabetes as per person.

BMI: It is the body Mass Index and is taken into account to understand the pattern of body which is more prone to diabetes on the basis of BMI.

Diabetes Pedigree Function: DPF can be used to provide the relatives with a synthesis of the history of diabetes mellitus and the genetic relationship of certain relatives to the subject. The DPF uses parents, grandparents, full and half siblings, full and half aunts and uncles, and cousins first. It gives a measure of the expected genetic influence of affected and uninfluenced relatives on the eventual risk of diabetes in the subject.

Age: It is a feature that is taken into consideration as it sometimes directly affect the outcome of the prediction as this feature defines the basic health factor of an average person.



5. ALGORITHM

The algorithms applied in this paper are primarily three: Naïve Bayes, SVM and Decision Tree .The dataset will be fed in these algorithms and the comparison will be done the basis of their accuracy for the testing or validation data. This validating data will be randomly selected from the whole data as in here we will use the actual data of 780 people and divided it randomly for training and testing purpose.

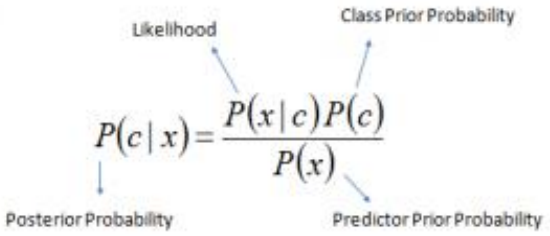
International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 4 June 2020

5.1 Naïve Bayes

It is a classification method based on Bayes ' Theorem in Machine Learning which recognizes an assumption of independence among all predictors / features. In simple terms, the classifier of Naive Bayes considers that the existence of a particular feature in a class is entirely unrelated to the presence of some other feature of that class.

For example, if it is orange in colour, circular in shape and around 2.5 inches in diameter, a fruit may be predicted to be an orange. Even if we believe that these characteristics depend on each other or on the presence of the other characteristics, all these properties contribute independently to the likelihood that the fruit is an orange, which is why it is called 'Naive.' The classifier Naive Bayes is simple to implement and is mainly useful for very large data sets. Naive Bayes is well known for outperforming even very advanced techniques of classification.

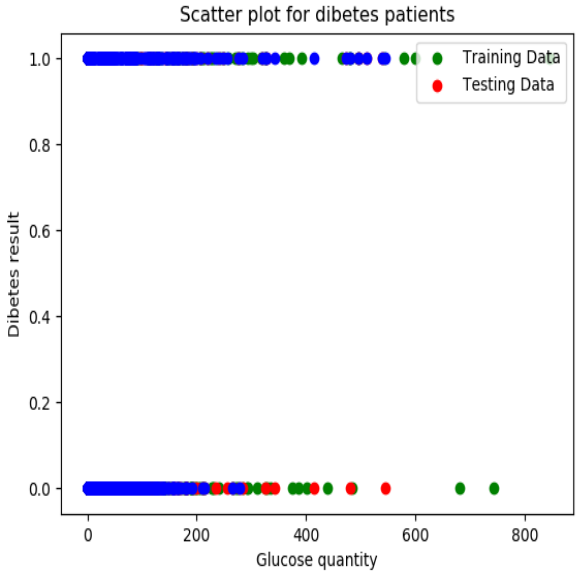
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c) which is the actual logic behind the bayesian prediction.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

The naïve bayes algorithm classifier didn't include much tuning in standard algorithm uses. The classifier can be easily implemented and the data can be fed without the hectic tuning process. The dataset is divided into training and testing part with a proportion of 70-30 where 70% data will be used for training purpose and the rest 30% will be used for the testing purpose. This is done using the method available is scikit i.e. train_test_split() and this uses basic parameter to decide how to split the training and testing data.



The classifier gives an accuracy of 74.89% and the data for training and testing is in 7 : 3 ratio. This graph here gives a glimpse of the data used for training and testing purpose and the predicted output. The graph is plotted between Glucose Quantity and the outcome/result.

Accuracy Score = 0.74891

5.2 SVM

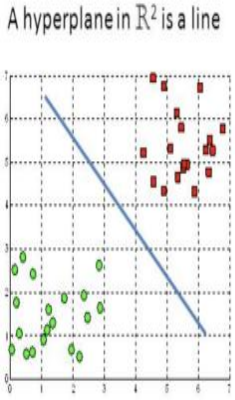
In an N-dimensional space (N — the number of features), the support vector machine algorithm seeks a hyperplane that distinctly classifies the data points. There can be so many different hyperplanes that could be selected to establish a distinction between the two classes of data points. Our goal is to find a plane that can give the maximum margin, that is, the maximum distance between the two classes' data points. Maximizing the margin gap provides some reinforcement to be able to distinguish the next upcoming data points with greater precision and confidence.

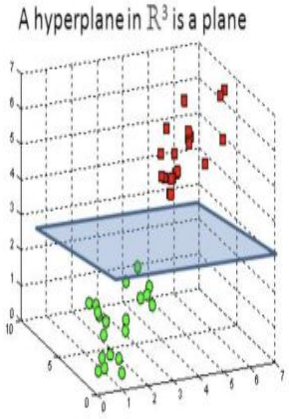
5.3 Hyperplane

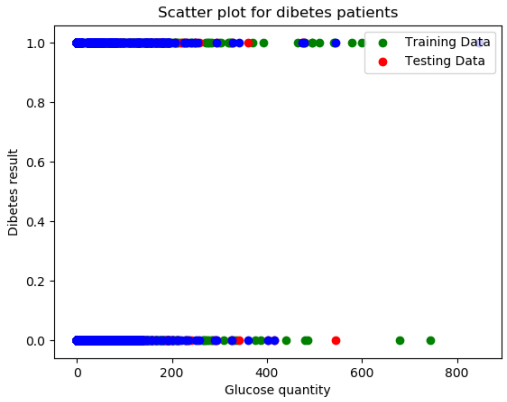
Hyperplanes are limits of decision taking that help to define the data points. You can easily assign data points falling on either side of the hyperplane to various groups. The hyperplane dimension clearly depends on the number of given features.

If the number of features on the input is 2, then the hyperplane is only a line. If the number of input features is 3 then the hyperplane becomes a two-dimensional plane. It gets hard to imagine as the number of features increases by more than 3.

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 4 June 2020

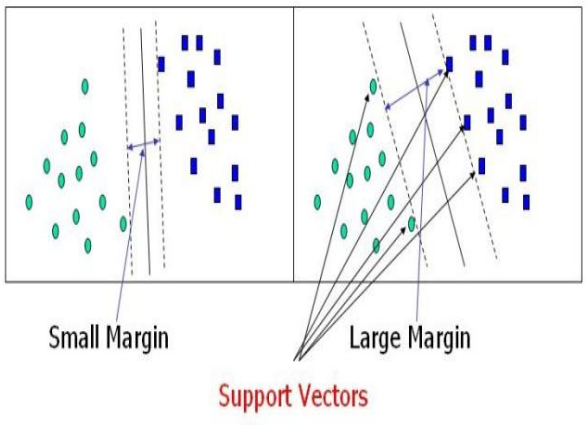






5.4 Support Vectors

Support vectors are data points near to the hyperplane, influencing the hyperplane 's direction and orientation, and moving the hyperplane according to its location. Use these vectors for help, we optimize the margin used in the classifier. Deleting the support vectors will change the hyperplane 's position, and that too is not good. These are the points which help us make our SVM as only one linear vector is unable to maximize the margin and therefore these support vectors provide support for our linear decision boundary to maximize the margin.

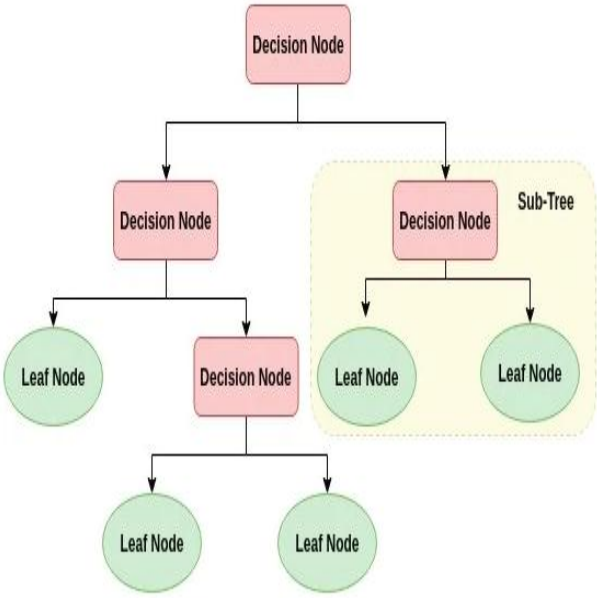


The SVM here gives an accuracy of approx 78 % and this is done by considering the 7:3 ratio in training and testing data. The data is same and is chosen randomly by same method. The classifier uses the kernel parameter as linear which is a standard value/condition.

The accuracy score by this algorithm is: 0.78354 which seems to be better as compared to the Naïve Bayes algorithm when the standard condition is met in both algorithms. The SVM takes a bit more time during training as it calculates error but it is mostly considered better for huge data.

5.5 Decision Tree

Decision Tree algorithm is a supervised technique for learning algorithms. Similar to other supervised learning algorithms, the decision tree algorithm is also used according to its implementation to solve regression and classification issues. The general aim of using the Decision Tree is to construct a training model that could be used to predict output / target variables class by learning the rules of decision created from pre-available data (training data). Comparing with other classification algorithms, the level of understanding of the Decision Tree algorithm is a bit simple. The decision tree algorithm solves the issue with the use of tree representation from the training data for the entire rule logic. An attribute corresponds to each internal node of the Decision tree, and each leaf node relates to a class label.



Assumptions

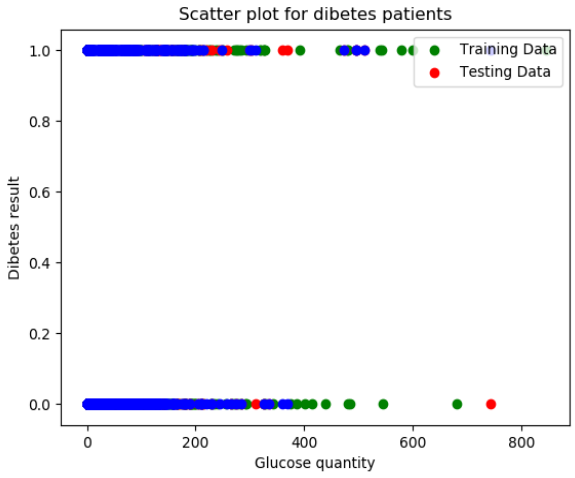
- At the beginning, the entire training dataset is called tree root.
- The feature values are considered primarily categorical. If the values are continuous then they are altered before constructing the model to discrete format.
- Records are distributed in a recursive order on the basis of attribute values.
- Statistical approach is used to decide the order of placing the root and the leaf in the tree.

The key challenge in implementing the decision tree is determining which attribute we will use at each stage as the root node. This method is called selection of the attributes. We have different methodologies of attribute selection to identify the attribute that can be used at each level as the root node.

The popular attribute selection measures:

- Information gain
- Gini index

The algorithm here gives an accuracy of 68.8% which is done using the same format of data and the algorithm is implemented with the standard parameter.



The algorithm gives an accuracy score of: 0.6883. This accuracy score and the accuracy is minimum in the above two algorithms.

6. CONCLUSION

In this paper we have done the comparison of the three supervised learning algorithms for the classification task to diagnose the diabetes disease. We have taken the same data and performed the analysis/prediction on that data with the standard parameter of the algorithms. The results of these prediction are considered in terms of their accuracy scores that are 0.74891, 0.78354 and 0.6883 for Naïve Bayes, SVM and Decision Tree respectively. By the accuracy of these algorithms we can easily decide that the performance of SVM is best in the three for this dataset and the accuracy is quite good with the standard parameters without any tuning. Hence these algorithms will provide a better accuracy and efficiency with the basic medical data for diagnosis and therefore useful in the field of medical diagnosis using the supervised learning techniques.

REFERENCES

[1] Articles related to medical diagnosis using ML from the '**ScholarPedia**'.

[2] Article related to "Diabetes Classification" by **Zhenhuan Cui**, **Crystal Dong**, **Juan Du**, **Lynn Friedmann**, **Yanxing Zhao**

[3] Mario R.Guarracino, Salvatore Cuciniello, DavideFeminiano, Gerardo Toraldo, Panos M. Pardalos,"Current Classification Algorithms for Bio Medical Applications: American Mathematical Society", CRM proceedings and Lecture Notes, Volume 45, 2008

[4] Carolina A.M. Schurink, Stefan Visscher, Peter J. F. Lucas, Henk J. van Leeuwen, Erik Buskens, ReinierG. Hoff, Andy I.M. Hoepelman, Marc J. M. Bonten, "A Bayesian decision Support System for diagnosing ventilator-associated pneumonia, Springer", Intensive Care Med 33: pp. 1379–1386, 2007

[5] Kononenko, I, "Inductive and Bayesian learning in medical diagnosis", Applied Arti□cial Intelligence 7(4): pp. 317-337, 1993

[6] Ruchika Gupta, Alok Sharma, Sompal Singh, Amit K Dinda,"Rule based decision support system in the biopsy diagnosis of glomerular disease", JCP Online. June 2011

[7] Ranjit Abraham, Jay B. Simha,S. SitharamaIyengar,"Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical data mining", International Journal of Computational Intelligence Research, Vol.4, 2008

[8] Peter Lucas, "Bayesian Networks in Medicine", A Model-based Approach to Medical Decision Making

[9] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier",

Elsevier, Expert Systems with Applications, 36, 2009, pp. 10618-10626.

[10] S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[11] Y. C. T. Bo Jin, "Support vector machines with genetic fuzzy feature transformation for biomedical data classification.," InfSci, vol. 177, no. 2, pp. 476–489, 2007.

[12] R. C. Barros, M. P. Basgalupp, A. A. Freitas, and A. C. P. L. F. de Carvalho, "Evolutionary Design of Decision-Tree Algorithms Tailored to Microarray Gene Expression Data Sets," IEEE Trans. Evol. Comput., vol. 18, no. 6, pp. 873–892, Dec. 2014.

