To select a model for non-invasive anemia detection for pregnant women

Classification Algorithms

Samruddhi Bora[1], Shreyash Joshi [2], Hiral Bhuptani[3], M.T Kolte

Department of Electronics and Telecommunication Engineering,

Pimpri Chinchwad College of Engineering, Pune, India
1samruddhibora86@gmail.com,
2shreyasjoshi166@gmail.com,
3hiralbhuptani01@gmail.com,
4mahesh.kolte@pccoepune.org

Abstract. The World Health Organization (WHO) identifies anemia, a health hazard condition marked by the deficiency of red blood cells or haemoglobin in the bloodstream, as maligning a quarter of the total world population. There are many different reasons why people get anemia, from common nutritional reasons like iron or folate deficiency, which are relatively easy to treat and cure, to uncommon genetic reasons like sickle cell disease or thalassemia major, which result in severe and chronic anaemia that needs to be closely monitored. Anemia screening or anemia diagnosis are both necessary for the detection of anemia, and both involve varying degrees of measurement precision. First, there is a definite clinical need for simple, widely available tools to screen for anaemia in at-risk populations (such as pregnant women, young children, and elderly patients) or the general public to determine whether a person needs formal confirmatory testing with the gold standard Hgb level test obtained through a complete blood count (CBC). However, there is also a need for non-invasive techniques that can more precisely and formally diagnose anemia and track. Hgb levels in individuals with known or chronic anemia. Because of this, we intend to train a model that will accept input from the patient's finger tips and aid in the detection of Hb levels, which are an indication of anemia.

1 Introduction

In a research by the World Health Organization (WHO), it was determined that between 1995 and 2005, 24.8 percent of the world's population was anemic. The gold standard for identifying anemia is the amount of hemoglobin present in a person's blood. This intravenous procedure calls for specialized medical gear. For lab testing, figure prick blood samples have recently been collected, however, this method is time-consuming and exposes medical staff to the danger of blood-borne illnesses. In many clinics, conjunctival pallor examination is typically done to quickly screen for anemia. Doctors typically draw down the eyelid and ostensibly assess the color of the anterior conjunctival pallor membrane. The clinical sign for anemia identification can be highly helpful in many circumstances, however, the limited sensitivity of anterior conjunctival color and the frequent absence of interobserver agreements can call into question the validity of the visual detection method. Color scale cards, which show the color spectrum and the matching hemoglobin content, are frequently used to reduce the issue of inter-observer disagreement and human error to increase the reliability of the visual detection process. The main component that contributes to the pigmentation seen in human blood is hemoglobin. In contrast to the green component of the light that it mostly absorbs, it has a bias toward reflecting the red component of the light that strikes its surface. The deep reddish color of hemoglobin is primarily due to this. Therefore, it is possible to obliquely estimate the hemoglobin concentration in the human bloodstream by comparing the red and green components of the RGB color spectrum of the conjunctival pallor.

Many people are susceptible to anemia because developing nations lack adequate healthcare and medical facilities. If an indicator of anemia can be assessed without using pricey blood tests, which are unavailable in many of these places, this condition can be improved. Even the availability of medical professionals is sporadic. If the presence of anemia in a patient could be determined using non-invasive techniques that do not require expensive testing or even the presence of a doctor or medical professional, that would be of tremendous assistance. Anemia can be a sign of various conditions, such as jaundice and malnutrition, in a person. The presence of various disorders can also be detected when anemia is present.

2 Methodology

2.1 Software Used: Jupyter Notebook (Anaconda)

The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.The Jupyter Notebook supports more than 40 different programming languages like R, Python, Java, etc. Therefore, most data science professionals tend to use Jupyter Notebooks to create and share documents, including code, equations, visualizations, computational outputs, markdown text, etc. Jupyter provides the implementation of numerous methods for Attributes Selection, which enables the automatic selection of characteristics to build a specific reduced dataset under each category. Choosing a certain algorithm, specifying the desired parameters, and executing the algorithm can be done on the dataset. The data set has been taken from Kaggle as well as manual data entry from real patient data from the Hospital. The data set has 16 attributes which are Age group, Region, ANC visit, Duration of Pregnancy, Body Mass, Breast feeding status, Body Mass Index (BMI), Termination History etc.

2.2 General Process of Building Model

Machine learning algorithms are based on a dataset. The dataset collected should be cleaned before training the algorithm. Machine learning algorithms can’t handle incomplete, incompatible, and noisy data; it has to be cleaned. The lack of attributes in the dataset is referred to as incomplete or incompatible data. The basic step of data preprocessing comprises data cleaning, data integration, data transformation, and data reduction

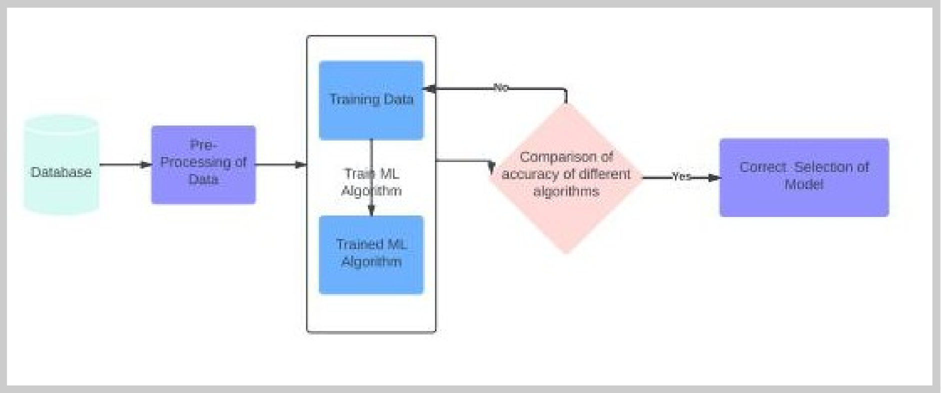


Fig. 1. Block Diagram

The dataset is then imported in Jupyter and is subjected to data preprocessing. The data preprocessing methods used are Data encoding, filling the missing values, and feature scaling.

2.3 Machine Learning Algorithms

The machine learning algorithms can be classified into three broad categories: supervised, unsupervised and semi-supervised. Supervised algorithms refer to the labeled training dataset that is used first to train the algorithm [6]. The trained algorithm is then applied to the unlabeled test data to categorize them into similar categories [8]. This test and train dataset is used to calculate accuracy parameters.

Extreme Gradient Boosting (XG Boost)

XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. This algorithm is described in **fig. 2**. It can work on regression, classification, ranking, and user-defined prediction problems.

Support Vector Machine (SVM)

SVM algorithm can be used to classify both types of data i.e., linear and non-linear data. It is a supervised learning model as well as it can also be used for both classification and regression. Every data item is an n-dimensional attribute space where n refers to the number of attributes.

The hyperplane separates the data items into two classes by maximizing the marginal distance of two classes and minimizing the classification errors [8] The hyperplane changes if the data points are changed. **fig. 3(b)** is a simplified illustration of an SVM algorithm.

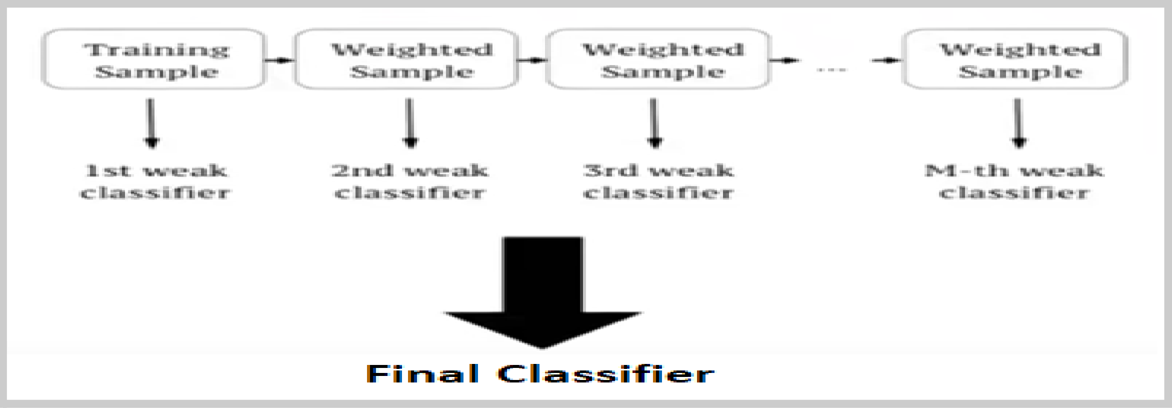
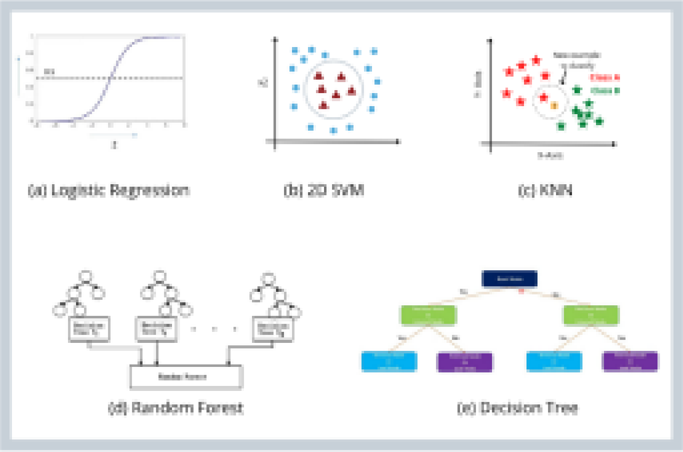


Fig. 2. Graphical Illustration of Extreme Gradient Boosting



(a) Logistic Regression

(b) 2D SVM

(c) KNN

(d) Random Forest

(e) Decision Tree

Fig. 3. Graphical Illustration of Machine learning Algorithm

K-Nearest Neighbors (K-NN)

The K- Nearest Neighbors algorithm is based on the Supervised Learning technique. This algorithm saves all present data and classifies a new data point based on its relation to the existing data. it makes no assumptions about the data it's working with which is called a non-parametric algorithm. The category or class of a dataset can be determined with the help of K-NN. Here, ‘K’ in the algorithm refers to the number of nearest neighbors undertaken for a vote. It must be an odd parameter also it must not be a multiple of classes. The sum of the square differences between a new point(x) and an old point(y) with a square root of them is used to evaluate Euclidean distance. The below equation illustrates the mathematical model of this algorithm. Also illustrated (see **Fig. 3(c))**

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]} \qquad (1)$$

Random Forest

Random forest is one of the supervised machine learning techniques. It applied for both Regres sion as well as Classification problems. It uses the method of a combination of multiple classifiers to solve a critical problem and to improve the efficiency of the model. This concept is based on ensemble learning. The random forest contains a number of decision trees. This method averages more than one decision tree which is trained on multiple parts of the same set of training with an aim of obtaining a reduction of the variance. It predicts the final output based on the majority votes of predictions. It can also handle large datasets with high dimensionality which prevents overfitting issues by enhancing the accuracy of the model. The simplified illustration is explained in (see **fig. 3(d))**

Decision Tree (DT)

Decision logic is modeled by a Decision tree. It can construct the model in the form of the tree where the target value takes the place of the leaf node. Nodes of a decision tree have multiple levels. The first node is called the root node. It builds a decision node at each step. In this model, there is no need for normalization or scaling. Also, it is not sensitive to outliners.it is illustrated by **(see fig. 3(e))**

Naïve Bayes

The naïve Bayes algorithm is based on the Bayes theorem. It is a supervised machine learning algorithm. The Bayes theorem describes the probability of occurrence of an event depending on the knowledge of previous conditions related to that event. It is used to predict based on the probability of an object as it assumes the occurrence of a certain attribute is independent of the occurrence of remaining features. If the fruit is classified on the basis of color, taste, and geometry, then red, sweet and spherical fruit is recognized as an apple. Here each feature is individually used to identify that it is an apple independently. Naïve Bayes can also be used for multi-class and binary Classifications.

$$P(A|B) = P(B|A)\, P(A) / P(B) \qquad (2)$$ where,

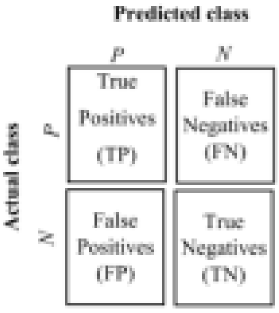
$P(A|B)$ – the probability of event A occurring, given event B has occurred $P(B|A)$ – the probability of event B occurring, given event A has occurred $P(A)$ – the probability of event A

$P(B)$ – the probability of event B

3 Results and Analysis

3.1 Accuracy Parameter.

To estimate the performance of the classifier Accuracy parameter is used . The confusion matrix in machine learning research is used to calculate various accuracy parameters.(see **fig. 4)** shows the basic structure of the confusion matrix. True positives also referred to as TP are positive cases where the classifier is accurately detected. True negatives referred to as TN are negative cases that were successfully detected by the classifier. False positives referred to as FP are negative cases in which the algorithm misidentified them as positive, while false negatives (FN) are positive cases in which the classifier misidentified them as negative. The confusion matrix based measurements listed below are routinely applied to assess the effectiveness of classifiers, particularly those based on supervised algorithms.



		Predicted class	
		P	*N*
Actual class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Fig. 4. Confusion matrix

A few more metrics are also used to evaluate the effectiveness of different classifiers. The root mean square error (RMSE) is one such measure. The mean value of all squared errors is also represented by RMSE for distinct pairs of actual and projected values. An error is a discrepancy between the expected value and the actual value. The mean absolute error is another example of such a metric is the absolute difference between an actual and an expected value.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

$$\text{F}_1 \text{ Score} = \frac{2 * TP}{2 * TP + FN + FP} \qquad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (5)$$

$$\text{Sensitivity} = \text{Recall} = \text{True Positive rate} = \frac{TP}{TP + FN} \qquad (6)$$

$$\text{Specificity} = \frac{TN}{TN + FN} \qquad (7)$$

$$\text{False Positive rate} = \frac{FP}{FP + TN} \qquad (8)$$

The recall is the proportion of samples belonging to the positive class which was correctly expected as positive. Accuracy signifies the fraction of correct predictions. Precision signifies the fraction of actual positives among those examples that are predicted as positive.

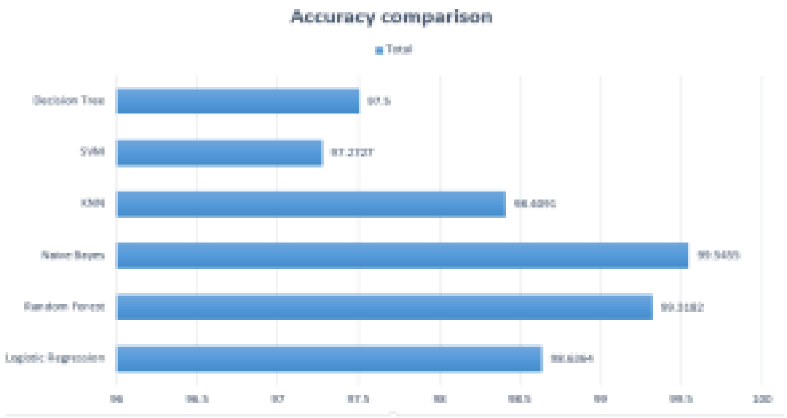


Fig. 4. Graph illustrating the accuracy of algorithms vs. algorithms it is based on table

4 Conclusion

The paper attempted to compare various machine learning classification algorithms for making an informed selection of accurate . The algorithms used for comparisons in this paper are Naive Bayes Decision Tree, Extreme Gradient Boosting , K-NN, Support vector machine, Random Forest. It is found that Random Forest and Naive Bayes are giving the best accuracy among the other algorithms as well as minimum error in comparison to algorithms. This algorithm can be applied for deciding crops for farmers according to their location. It can also be analyzed using unsupervised and reinforcement algorithms.

5 References

1. Balaji Prabhu B.V., Dr. M Dakshayaini "Demand Prediction Model for forecasting AGRI-Needs of Society"-International conference on inventive computing and informatics-2020

2. Aditya Humnabadkar, Omkar Kulkarni, Mrs. Rajani.P.K, " Automation of Vehicle Identification at Night using Light Source Recognition", 5th International Conference on Computing Communication Control and automation (ICCUBEA 2019), Pune, In dia, held on 19th to 21st September 2019

3. Uddin, S., Khan, A., Hossain, M. *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* **19,** 281 (2019).

4. Rajani P. K. and Arti Khaparde "Video error concealment using block matching and frequency selective extrapolation algorithms", Proc. SPIE 10443, Second International Workshop on Pattern Recognition, 104431C (19 June 2017)

5. G. M. Naidu "Applicability of Arima models in the wholesale vegetable market." Int. J. Agric. Stat. Sci., vol. 11, no. 1, pp. 69–72. (2015)

6. Joachims T. Making large-scale SVM learning practical. SFB 475: Komplexitätsreduk tion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Tech. Rep. 1998. p. 28.

7. Rajani, P. K., Neha Motagi, Komal Nair, and Rupali Narayankar. "Autonomous Smart Device for COVID-19 Detection Using Artificial Intelligence." In Handbook of Research on Applied Intelligence for Health and Clinical Informatics, pp. 128-147. IGI Global, 2022