ISSN No:-2456-2165

Vitamin-A Deficiency Classification in School Children Using Machine Learning

Leelavathi Arepalli[1]
Sr. Asst Prof.
Department of CSE,
Sri Vasavi Engg. College
Tadepalligudem

Durga Sasindra Vakalapudi[2]
B. Tech III Year,
Department of AIM
Sri Vasavi Engg. College,
Tadepalligudem

Sai Nivesh Bomma[3]
B. Tech III Year,
Department of AIM,
Sri Vasavi Engg. College,
Tadepalligudem

Gowthami Maka[4]
B. Tech III Year,
Department of AIM,
Sri Vasavi Engg. College
Tadepalligudem.

Pavan Kalyan Saila[5]
B. Tech III Year,
Department of AIM
Sri Vasavi Engg. College,
Tadepalligudem.

Nitya Sri Nekkanti[6]
B. Tech III Year,
Department of AIM,
Sri Vasavi Engg. College,
Tadepalligudem.

Abstract:- The main theme of our paper is to early detection of vitamin-A deficiency in school children by using Logistic Regression.[1][2] Vitamin 'A' Deficiency is a significant public health issue affecting millions of children worldwide [4], particularly in developing countries. It can lead to serious health consequences, including impaired vision, and weakened immunity. Early detection and classification of this deficiency in schoolchildren are crucial for implementing interventions to improve their overall health and well-being [3]. This project proposes the application of machine learning techniques, specifically logistic regression, to accurately classify the presence of deficiency in school-aged children based on relevant clinical and demographic factors. The primary objective of this research is to develop a predictive model that can efficiently and accurately identify the children at risk of this deficiency [9]. The proposed project will utilize a dataset collected from schoolchildren in target regions, encompassing key features such as age, sex, location, and symptoms related to Vitamin A [6][9]. These data will be processed and pre-processed to ensure data quality and remove any potential bias. Logistic regression, a widely used classification algorithm in machine learning, will be employed to build the predictive model. The model will be trained on a labeled subset of the dataset, where the presence or absence of Vitamin A deficiency is indicated.

Keywords:- *Machine Learning, Logistic Regression, Scikit-Learn, Jupyter Notebook,Pandas, Matplotlib.*

I. INTRODUCTION

➢ *Introduction to ML*

Machine Learning is a technique of designing a machine or simply we can say that Self - Learning of a machine by training it. We can also say that machine learning is an automatedprocess for machines with less or no human input and the work will be done very fast compared with humans. Nowadays machine learning is used in many sectors such as financial, health, hospitals, and government sectors.

➢ *Applications of Machine Learning:*

- *Online Fraud Detection*
- *Self-driving cars*
- *Speech Recognition*
- *Credit card fraud detection*
- *Image Recognition*
- *Product Recommendations*
- *Email Spam and malware filtering*
- *Automatic Language Translation*
- *Virtual Personal Assistant*
- *Facial Recognition*

II. OBJECTIVE

The objective of the project we have developed is to predictive a model using logistic regression and assess the risk of vitamin A deficiency in children based on relevant dietary and other factors. By achieving this objective, we aim to:

- *Identify High-Risk Individuals*
- *Early Detection*
- *Prevent Health Complications*
- *Promote Nutritional Awareness*
- *Ultimately, Improve Child Health*

III. RELATED WORK

➢ *Scikit-Learn*

Scikit–Learn is also called Sklearn. Sklearn is an open-source machine learning library in Python programming language. It is designed by the both numerical and scientific of NumPy and SciPy. To learn Scikit–Learn the programmer must be aware of Python, NumPy, SciPy & Matplotlib libraries.

ISSN No:-2456-2165

- *Installation:*

We can install this library by using the command prompt or by using conda. The easiest way is to install it from the command prompt by running the following command:

✓ *Pip Install -U Scikit-Learn*
From Conda: **conda install scikit-learn**

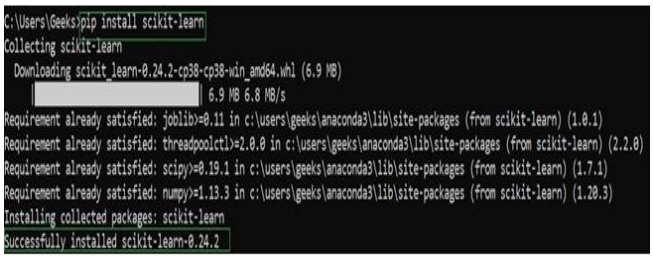

Fig 1 Installing Scikit-Learn

- *Characteristics of Scikit-Learn:*

✓ *Data Clustering*
✓ *Linear Regression*
✓ *Logistic Regression*
✓ *K- means clustering*
✓ *Decision Trees*
✓ *Random forest*
✓ *Support Vector Machines*
✓ *Confusion Matrix*

➢ *Logistic Regression*

This type of statistical model (also known as the logit model) is frequently used for classification and predictive analytics. Logistic regression estimates the probability of an event being, similar as voting or not voting, grounded on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied to the odds—that is, the probability of success divided by the probability of failure. This is also generally known as the log odds or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi)

- *Usage of Logistic Regression:*

Vitamin A insufficiency affects about 190 million preschool-age children, substantially from Africa and South-East Asia [9]. In babies and children, vitamin A is essential to support rapid growth and to help combat infections [5]. Inadequate inputs of vitamin A may lead to vitamin A insufficiency which can cause visual impairment in the form of night blindness and may increase the threat of illness and death from childhood infections, including measles and those causing diarrhea, etc. Beforehand discovery and bracket of VAD in preschool/ academy children [11] are pivotal for enforcing interventions to ameliorate their overall health and well- being. Our Aim for early detection of vitamin A substantially occurs in children[1][2]. For this we make a prophetic model i.e.; logistic regression is a widely used ML algorithm used to make prophetic models and directly classify the presence of vitamin A in the deficient. Logistic regression uses binary classification like 0's and 1's. For illustration: To prognosticate whether the insufficiency is present or not.

- *Key Advantages of Logistic Regression:*

✓ Easier to apply machine literacy styles A machine literacy model can be effectively set up with the help of training and testing. The training identifies patterns in the input data(image) and associates them with some form of affair(marker). Training a logistic model with a retrogression algorithm doesn't demand advanced computational power. As similar, logistic retrogression is easier to apply, interpret, and train than other ML styles.
✓ Suitable for linearly divisible datasets A linearly divisible dataset refers to a graph where a straight line separates the two data classes. In logistic retrogression, the y variable takes only two values. Hence, one can effectively classify data into two separate classes if linearly divisible data is used.
✓ Provides precious perceptivity Logistic retrogression measures how applicable or applicable an independent/ predictor variable is(measure size) and also reveals the direction of their relationship or association(positive or negative).

➢ *Confusion Matrix*

Confusion matrix is a veritably popular measure used while working bracket problems. It can be applied to double bracket as well as to multiclass bracket problems. A confusion matrix is used in machine literacy to assess the performance of a bracket model.

It summarizes the results of bracket by showing the count of TP, TN, FP, and FN prognostications. The values help to estimate a model's delicacy, perfection, recall, and f1 score.

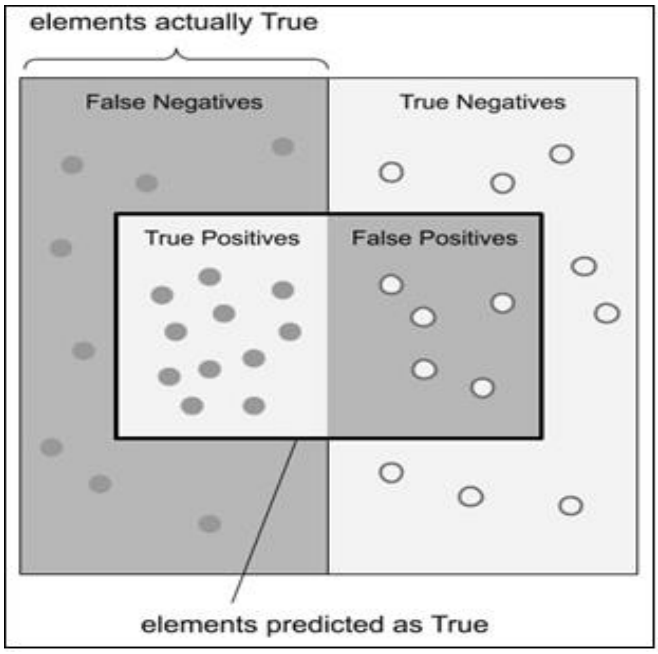

Fig 2 Confusion Matrix

ISSN No:-2456-2165

Confusion matrix represents counts from prognosticated and factual values. The affair “TN (True Negative)” which shows the number of negative exemplifications classified directly. also,

“TP(True Positive)” which indicates the number of positive exemplifications classified directly. The term "FP(False Positive)” value, i.e., the number of factual negative exemplifications classified as positive; and “FN(False Negative)” value which is the number of factual positive exemplifications classified as negative. Performance criteria of an algorithm are delicacy, perfection, recall, and F1 score, which are calculated grounded on the below- stated TP, TN, FP, and FN.

- *Accuracy:*

The accuracy of an algorithm is represented as the ratio of correctly classified patients to thetotal number of patients.

✓ *Accuracy=TN+TP / TN+FP+FN+TP*

- *Precision:*

The precision of an algorithm is represented as the ratio of correctly classified patients with the disease to the total patients predicted to have the disease.

✓ *Precision=TP / TP+FP*

- *Recall:*

Recall metric is defined as the ratio of correctly classified diseased patients (*TP*) dividedby the total number of patients who have the disease. The perception behind recallshow many patients have been classified as having the disease. The recall is also called sensitivity.

✓ *Recall=TP / TP+FN*

- *F1 Score:*

The F1 score is also known as the F-measure. The F1 score states the equilibrium between the precision and the recall.

✓ *F1 score=2*precision*recall / precision + recall*

IV. METHODOLOGY

There are standard way that you ’ve to follow for a Machine Learning design. For any design, first, we've to collect the data according to our business requirements. The coming step is to clean the data and change categorical variables to numerical values. After that training of a model, uses colorful machine learning algorithms. Next, is model evaluation using different criteria like recall, f1 score, delicacy, etc. Eventually, model deployment o and retrain a model.

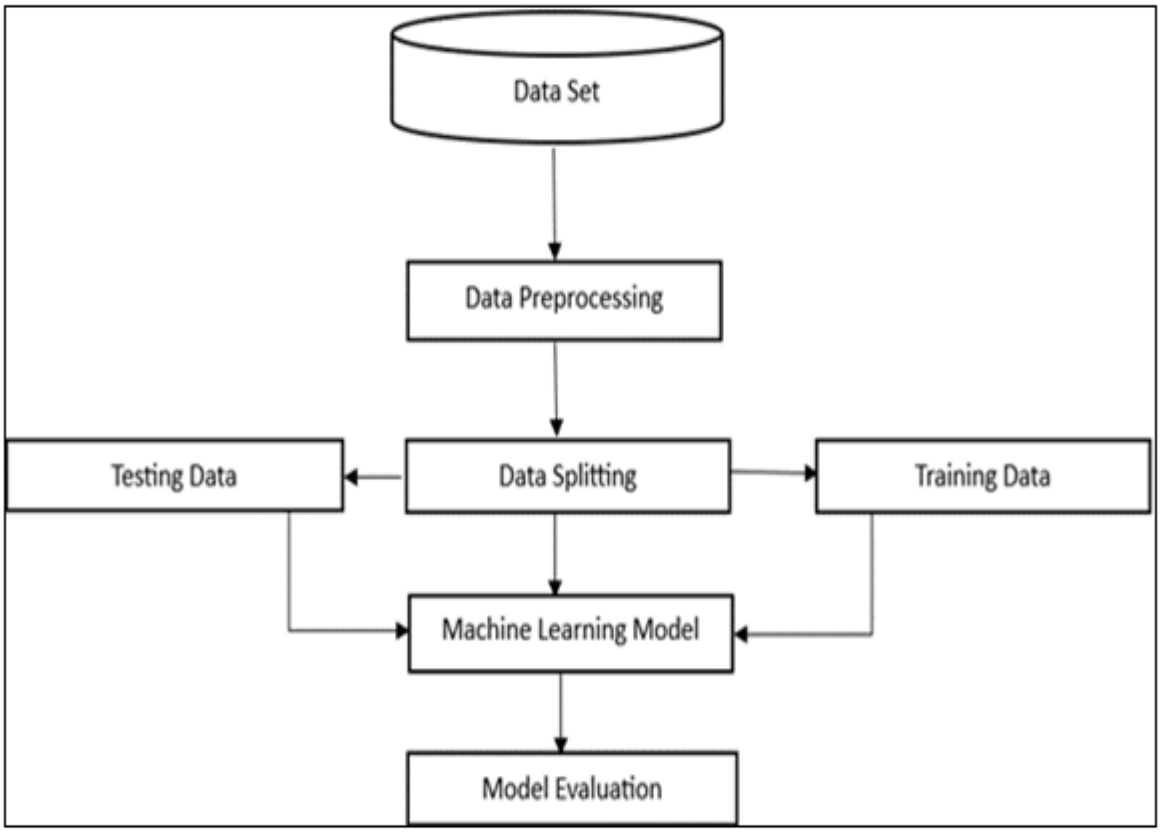

Fig 3 Flow Chart

ISSN No:-2456-2165

➢ *Data Collection*

The dataset is collected from Kaggle and keeps the data in Excel format. The gathered dataset that includes information about individualities and applicable features similar as age, coitus, skin health, symptoms like anemia,etc.

➢ *Data Preprocessing*

Data Preprocessing is for converting raw data into a format accessible by the ML algorithms. For that, we first clean the data by handling missing values, null values, outliers, and inconsistencies and also, convert it into an accessible format.

➢ *Point Selection*

Identify the most applicable features that might contribute to Vitamin A insufficiency similar as, our most important point is age of children and retinol serum situations.

➢ *Data Invoking*

Split the dataset into X-train, Y-train, X-test, Y-test. The training set used for training and the testing set used for evaluate the performance of the model.

➢ *Model Training*

Train the model using the training data So , the model will learn the relationship between the input features and the double outgrowth i.e.; Yes(1) and No(0).

➢ *Model Evaluation*

Validate the model's performance using the testing dataset. Common evaluation criteria for double bracket include delicacy, perfection, recall, and F1- score

V. RESULTS AND DISCUSSION

➢ *Get the Dataset:*

At first, the dataset was collected from a local health care center and it was downloaded as a CSVfile and then it was processed using Python. The collected dataset consists of 150 rows and 13columns.

The Features consist of the dataset such as Age, Gender, location, Retinol Serum Levels(µg/dl), Eye Infection state, symptoms like Anemia, etc. and the target class shows whether the children are Deficient (or) not.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Indicator.	Area covered	Population	Gender	Age(in months)	Skin health	Retinol_Serum_Level (per deciliter (µg	Height (in cms)	Weight (in kgs)	Eye Infection state	Having Anemia or not	Socio-economic Status	Deficient / not
2	Retinol binding protein	both urban and rural	Preschool-age children	Female	60	rough and dry	greater than 20	105	17.92	red and swallon eyes	Yes	Low income	No
3	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	rough and dry	less than 20	107.5	17	red and swallon eyes	Yes	Low income	yes
4	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	rough and dry	less than 20	107	17	red and swallon eyes	Yes	Low income	yes
5	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	rough and dry	less than 20	108	18	red and swallon eyes	Yes	Low income	yes
6	Retinol binding protein	both urban and rural	Preschool-age children	Female	60	rough and dry	less than 20	106	17.5	red and swallon eyes	Yes	Low income	yes
7	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	rough and dry	less than 20	107	17.8	red and swallon eyes	Yes	Low income	yes
8	Retinol binding protein	both urban and rural	Preschool-age children	Female	60	normal	greater than 20	106.5	17.4	red and swallon eyes	Yes	Low income	No
9	Retinol binding protein	both urban and rural	Preschool-age children	Female	60	normal	greater than 20	105	17.1	red and swallon eyes	Yes	Low income	No
10	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	normal	greater than 20	106	17.9	red and swallon eyes	Yes	Low income	No
11	Retinol binding protein	both urban and rural	Preschool-age children	Female	60	normal	greater than 20	105	18	red and swallon eyes	Yes	Low income	No
12	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	normal	less than 20	106.5	17.6	red and swallon eyes	Yes	Low income	yes
13	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	scaly and dry	greater than 20	107.5	18	red and swallon eyes	Yes	Low income	No
14	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	scaly and dry	greater than 20	106	17.98	red and swallon eyes	Yes	Low income	No
15	Retinol binding protein	both urban and rural	Preschool-age children	Male	60	normal	greater than 20	105.5	17.4	red and swallon eyes	Yes	Low income	No
16	Retinol binding protein	both urban and rural	Preschool-age children	Female	48	normal	greater than 20	100.5	15.42	red and swallon eyes	Yes	Low income	No
17	Retinol binding protein	both urban and rural	Preschool-age children	Male	48	normal	greater than 20	101	15.5	red and swallon eyes	Yes	Low income	No
18	Retinol binding protein	rural	Preschool-age children	Female	48	normal	less than 20	100	15.3	red and swallon eyes	Yes	Low income	yes
19	Retinol binding protein	urban	Preschool-age children	Female	48	normal	less than 20	99.5	16	red and swallon eyes	Yes	Low income	yes
20	Retinol binding protein	both urban and rural	Preschool-age children	Female	48	normal	less than 20	99	15.9	red and swallon eyes	Yes	Low income	yes
21	Retinol binding protein	both urban and rural	Preschool-age children	Female	48	rough and dry	less than 20	100	15.4	red and swallon eyes	Yes	Low income	yes
22	Retinol binding protein	both urban and rural	Preschool-age children	Male	48	rough and dry	less than 20	103	15.78	red and swallon eyes	Yes	Low income	yes
23	Retinol binding protein	both urban and rural	Preschool-age children	Male	48	rough and dry	greater than 20	99	16	red and swallon eyes	Yes	Low income	No
24	Retinol binding protein	both urban and rural	Preschool-age children	Male	48	rough and dry	greater than 20	101.5	15.9	red and swallon eyes	Yes	Low income	No
25	Retinol binding protein	both urban and rural	Preschool-age children	Female	48	rough and dry	greater than 20	101	15.3	red and swallon eyes	Yes	Low income	No
26	Retinol binding protein	both urban and rural	Preschool-age children	Male	48	rough and dry	greater than 20	102	15	red and swallon eyes	Yes	Low income	No
27	Retinol binding protein	both urban and rural	Preschool-age children	Male	48	normal	greater than 20	100	15.8	red and swallon eyes	Yes	Low income	No
28	Retinol binding protein	both urban and rural	Preschool-age children	Female	48	normal	greater than 20	99.5	16	red and swallon eyes	Yes	Low income	No
29	Retinol binding protein	both urban and rural	Preschool-age children	Female	48	normal	greater than 20	98.5	15.9	red and swallon eyes	Yes	Low income	No
30	Retinol binding protein	both urban and rural	Preschool-age children	Male	48	normal	greater than 20	100	15.7	red and swallon eyes	Yes	Low income	No

Fig 4 Data Set

➢ *Importing Libraries:*

After Collecting the dataset, the next step is to import libraries that are necessary f

- *Import pandas as pd*
- *Import Matplotlib.pyplot as pltfrom sklearn import **

➢ *Importing Data:*

The downloaded data is imported into the Python code file as a Data frame using the pandas module.

```
df=pd.read_csv("Dataset.csv")
```
Python

Fig 5 Importing Data

➢ *Preprocessing step:*

Data Preprocessing is a part of the data analysis and mining process responsible to convert raw data into a format understandable by the ML algorithms. For that, we first clean the data by handling missing values, null values, outliers, and inconsistencies and then, convert it into an understandable format.

```
duplicate_rows = df.duplicated()
duplicate_rows_count = duplicate_rows.sum()

if duplicate_rows_count > 0:
    print("The dataset contains", duplicate_rows_count, "duplicate rows.")
else:
    print("The dataset does not contain any duplicate rows.")
```
Python

Fig 6 Data Preprocessing

➢ *Splitting X and Y terms:*

After Data Preprocessing , we divide our dataset into X and Y terms as it shown in the below:

```
X,Y= df.iloc[:, 0:-1],df["Deficient/not"]
```

Fig 7 Extracting X, Y variables

➢ *Encoding Categorical Data:*

Categorical data refers to a type of data that represents specific categories or groups. It is a typeof data that is non-numerical and consists of labels or qualitative values rather than numericalvalues. Categorical data is often represented by text or symbols and can be divided into different distinct groups or categories. In machine learning, categorical data is typically represented usingthe “object” or “string” data type. For example, Gender: Categorical variable with categories such as “Male” and “Female.”. In the dataset, we use **Ordinal Encoding** to assign each unique value to a different integer.

```
enc = preprocessing.OrdinalEncoder()
X=enc.fit_transform(X)
```

Fig 8 Encoding the Data

- *In the above code, we have imported the Ordinal Encoder class of the sklearn library.*

➢ *Splitting the Dataset into the X-train, X-test, Y-train and Y-test:*

After the encoding step ,we divide the dataset into a X-train, X-test, Y-train and Y-test 80 of the dataset is resolve into a training dataset and the remaining 20 is resolve into a test dataset.

```
X_train, X_test, Y_train, Y_test=train_test_split( X, Y, test_size=0.2)
```

Fig 9 Splitting the Dataset

➢ *Construction of Model:*

For the model construction, we import our Logistic Regression model from the sklearn library

- *From Sklearn. Linear_Model Import Logisticregression*

Now, construct the logistic regression model and fit the training sets i.e.; x_train, y_train.

```
LR_model = LogisticRegression()
LR_model.fit(X_train, Y_train)
```
Python

```
▾ LogisticRegression
LogisticRegression()
```

Fig 10 Model Construction

- *Prediction of the Test Result*

Our model is well-trained on the training set, so we will now predict the result by using test set data. Below is the code for it:

```
Y_pred_class=loaded_model.predict(X_test)
Y_pred_prob=loaded_model.predict_proba(X_test)
print((Y_pred_class==Y_test).sum()/Y_test.shape[0])
print(loaded_model.score(X_test,Y_test))
```
Python

```
1.0
1.0
```

Fig 11 Predicting the Result

ISSN No:-2456-2165

In the above code, we have created aypred class vector to predict the test set result.

➢ *Test Accuracy of the Result*

Here we will produce the confusion matrix then to check the delicacy of the bracket. To produce it, we need to import the confusion matrix function of the sklearn library. After importing the function, we will call it using a new variablecm_display. The function takes two parameters, substantiallyy_true(the factual values) andy_pred_class(the targeted value returned by the classifier). Below is the law for it

```
confusion_matrix = confusion_matrix(Y_test,Y_pred_class)
                                                    Python

cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix =
             confusion_matrix, display_labels = [False, True])
                                                    Python
```

Fig 12 Test Accuracy of the Result

By executing the above code, we will produce a new confusion matrix. Consider the below image:

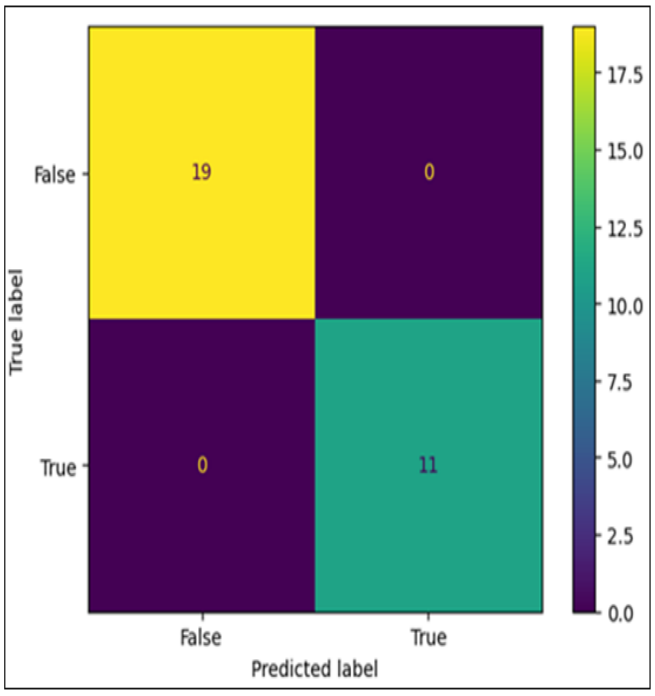

Fig 13 Test Accuracy of the Result

VI. CONCLUSION

The accurate prediction of Vitamin A Deficiency (VAD) is very crucial for avoiding major diseases like Anemia and Blindness in children [2]. However, this can be a challenging task to identify directly but, in this paper, we have concluded that the use of a Logistic regression algorithm for VAD identification is an effective approach [4]. The algorithm has proven its ability to identify complex relationships within the data and provide accurate results. Its agility and predictive accuracy make it the preferred approach over all otheravailable methods.

FUTURE SCOPE

The future scope of Vitamin A Deficiency using machine learning involves the collection of past patients' data who have gone through the tests. Future advancements in Vitamin ADeficiency will likely involve a large amount of highly sophisticated and examined data that can produce more accurate Real-time use cases.

REFERENCES

[1]. Dalmiya N, Darnton-Hill I, Greig A, Palmer A, Wardlaw T. Vitamin A supplementation: progress for child survival. Working paper. New York: UNICEF, December 2006.

[2]. Jayroop Ramesh, Donthi Sankalpa, Amar Khamis, Assim Sabayon, and Fadi Alou Explainable Machine Learning for Vitamin-A Deficiency Classification in Schoolchildren

[3]. Gorstein J, Shrestha RK, Pandey S, Adhikari RK, Pradhan A. Current status of vitamin A deficiency and the national vitamin A control program in Nepal: results of the 1998 national micronutrient status survey. Asia Pac J Clin Nutr 2003; 12:96–103.

[4]. Wiseman EM, Bar-El Dadon S, Reifen R. The vicious cycle of vitamin A deficiency: A review. Crit Rev Food Sci Nutr. 2017 Nov 22;57(17):3703-3714.

[5]. D'Ambrosio DN, Clugston RD, Blaner WS. Vitamin A metabolism: an update. Nutrients. 2011 Jan;3(1):63-103.

[6]. Hombali AS, Solon JA, Venkatesh BT, Nair NS, Peña-Rosas JP. Fortification of staple foods with vitamin A for vitamin A deficiency. Cochrane Database Syst Rev. 2019 May10;5(5): CD010068.

[7]. Harrison EH. Mechanisms involved in the intestinal absorption of dietary vitamin A and provitamin A carotenoids. Biochim Biophys Acta. 2012 Jan;1821(1):70-7.

[8]. Senoo H, Mezaki Y, Fujiwara M. The stellate cell system (vitamin A-storing cell system). Anat Sci Int. 2017 Sep;92(4):387-455.

[9]. Wirth JP, Petry N, Tanumihardjo SA, Rogers LM, McLean E, Greig A, Garrett GS, Klemm RD, Rohner F. Vitamin A Supplementation Programs and Country-Level Evidence of Vitamin A Deficiency. Nutrients. 2017

[10]. Pfeiffer CM, Sternberg MR, Schleicher RL, Haynes BM, Rybak ME, Pirkle JL. The CDC's Second National Report on Biochemical Indicators of Diet and Nutrition in the

[11]. U.S. Population is a valuable tool for researchers and policymakers. J Nutr. 2013Jun;143(6):938S-47S.

[12]. Miller M, Humphrey J, Johnson E, Marinda E, Brookmeyer R, Katz J. Why do children become vitamin A deficient? J Nutr. 2002 Sep;132(9 Suppl):2867S-2880S.