International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 50-55 March 2026-Sept 2026 www.csjournals.com

Early Risk Prediction: Vitamin Deficiency Diseases Using Clinical and Lifestyle Data

Indu, Dr. Vishal Verma, Ashish
Dept. of Computer Sc. and Appl, Ch. Ranbir Singh University Jind , Haryana , India

Abstract

Machine learning is famous for it learns on its own. Its means instead of telling it every rule, I'm trying to explain it with the help of an example like 1500 bird images after that it learns this is a bird. Then model can find bird in new images which are unseen for this. Model is able to work fast like spam emails scan, medical scan with in seconds. Our smart phones are using it in face unlock, Netflix shows and Google search suggestions. This models are less time consuming. Traditional workflow in which control is totally in your own hand which is rule based on the other hand workflow of machine learning is totally different because this models learns overall the time to learn and find new patterns on required data. This is experimental study in which we are predicting the result for the early risk vitamin deficiency on the basics of clinical and lifestyle factors. We study many papers in those papers clinical dataset is taking as observation but in this case lifestyle factors are avoiding by them. We are doing to study both factors which are able to predict more correctly with the help of machine learning algorithms example random forest, k-nearest neighbors (k-NN), XG Boost(extreme gradient boosting), logistic regression ,decision tree. In which we are comparing this all algorithms and finding which one is best for the predicting the correct output for given input. It is a supervised machine learning in which data type is labeled.

Key words: vitamin deficiency, k-nearest neighbors, XG Boost , logistic regression, decision tree , supervised machine learning.

Introduction

Now days most of the peoples are suffering from vitamin deficiency. Because of that peoples are feeling ill and tired. This problem is arising due to do not take proper diet and exercise. According to biology, there are 13 vitamins are present in which they are divided into two types which is water soluble(B,C) and fat soluble (K,E,D,A).Vitamin deficiency is chronic disease which are long term conditions that often improve slowly. We are making a vitamin deficiency early risk prediction model. There are many models are present which are able to work for it by using machine learning and python language. That is having three types of learning which are supervised learning, unsupervised learning and reinforcement learning. We are working with supervised learning in which those models are on trained labeled trained on labeled dataset. The common algorithms are linear regression, XG boost, decision trees, random forest and support vector machines (SVM). In this we are going to compare different types of model which are able to predict the

early risk vitamin deficiency according to the training of the dataset. At the of comparison we are going to study about different terms which are playing an important role in predicting the correct output related to the given input like F1-score , accuracy , precision and confusion matrix.

In this research, we are targeting to find the role of exercise level and sun exposure, along with clinical measure in predicting vitamin deficiency risk using machine learning models.

Inspiration

From my earlier days, I am too much interesting in computer science with the passage of time this interest is becoming my way of academic study. I am always interesting in those type of work which is related to websites and applications. How system is working in backend side and it is working for the user in frontend side. Machine learning is a powerful tool which is able to help me to follow my own passion. Which is able to handle huge large dataset which is may be structured or not like images, videos. Python language which i used in this project because of it's simplicity and friendly language which is having prebuilt libraries.

Research Methodology

The research methodology involves the following steps:

Collecting Data
↓
Preprocessing Of Data
↓
Exploring The Dataset
↓
Selecting Feature
↓
Model Training

(LR, DT, RF, SVM, XGBoost, KNN)
↓
Evaluation Of Model
↓
Comparison Analysis
↓
Selection of Best Model
↓
Vitamin Deficiency Prediction
↓
Result Generation

Collecting Data : data set which is having text and numerical data in tabular clinical and lifestyle format which is taking from website. It is having 4000 rows and 34 columns which is a supervised data which is having 33 independent features but only a single one target feature in it.

Data preprocessing: which is defined as cleaning and fixing your data with the help of that your model can understand it without making mistakes. In this step we are fixing blanks like their salary which we cannot leave them blank. We are filling those types of missing values with average number. Turning words into numbers in that case as we are very well know that computer is always able to understand only numbers so in this step we says , 'yes is representing by 0 and no is representing by 1'. This

process is able to improve data quality, removes errors and handling missing values.

Exploratory Data Analysis (EDA): that is used for understanding the data with finding its characteristics and patterns. Which are very useful for identifies trends, relationships and potential outliers through visualizations and statistical summaries.

Selecting Feature : It is the process of choosing the most relevant and important features from a dataset that contribute significantly to the prediction or analysis task. This helps in remove irrelevant, redundant, or less useful features, improving model performance and reducing computational complexity.

Model Training (LR, DT, RF, SVM, XGBoost, KNN): Train different machine learning algorithms using the prepared dataset. Which are enable to learn patterns and relationships from the training data for future predictions.

Evaluation Of Model : it is able to measure the performs of each model. With the help of using metrics such as Accuracy, Precision, Recall, and F1-Score to assess prediction quality.

Comparison Analysis: Compare the performance of all trained models. With the help of that we are able to find which algorithm is performs best for prediction task.

Selection Of Best Model Selection: Select the model with the highest overall performance. The most accurate and reliable model is chosen for prediction.

Vitamin Deficiency Prediction: Predict the selected model to predict the vitamin deficiency status.

Result Generation: Presenting the prediction outcomes in an understandable manner. Which is becoming the final results and performance metrics that support in decision making and further analysis..

Used Techniques :

1. Decision Tree: A Decision Tree works by splitting the data step by step based on simple yes-or-no questions. Each split helps the model make a better decision, and the process continues until the final prediction is reached.

2. Random Forest: this is a collection of many different decision trees. Which are working together as a team to vote on the final prediction.

3. K N N(K-Nearest Neighbors): this model makes a prediction for a new data point by looking at the closest which is most similar data points already in the system.

4. S V M (Support Vector Machine): this model draws a clear boundary line to separate different groups of data as clearly as possible.

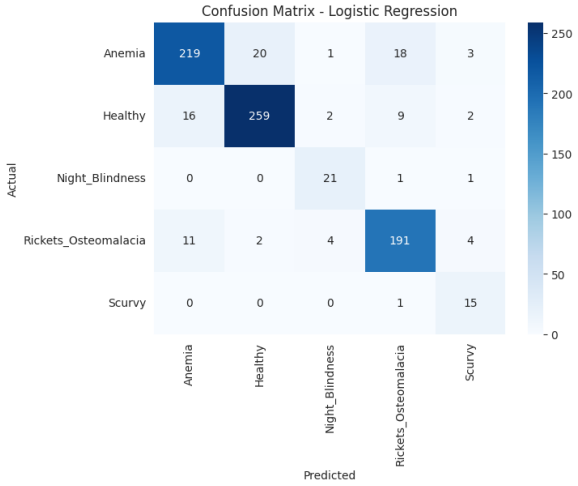
5. Logistic Regression: this model is used to calculate the probability that something belongs to a specific category.

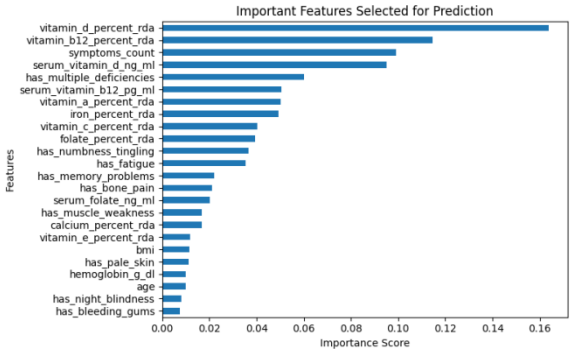
6. XG Boost (Extreme Gradient Boosting): This model uses several decision trees. Each new tree fixes the mistakes of the previous one, which helps improve the final prediction.

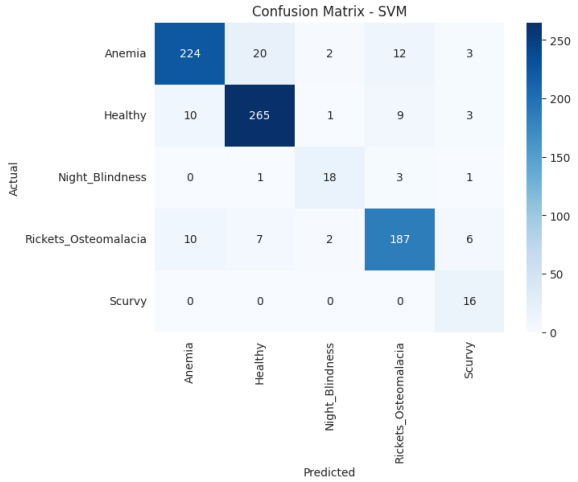
International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 50-55 March 2026-Sept 2026 www.csjournals.com

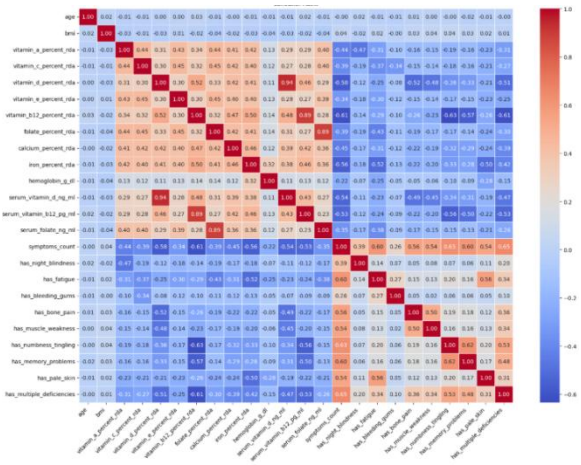
Final Result

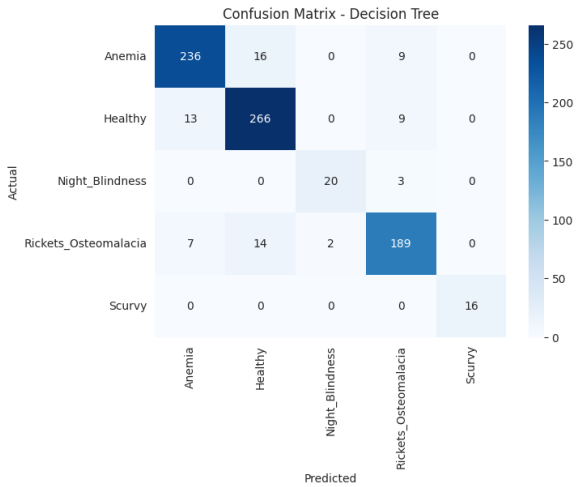
Model	Accuracy	Precision	recall	F1 Score
Logistic Regression	0.88125	0.885888	0.88125	0.882221
Decision Tree	0.90875	0.908912	0.90875	0.908723
Random Forest	0.93000	0.930220	0.93000	0.929896
SVM	0.88750	0.893498	0.88750	0.929896
KNN	0.83750	0.843464	0.83750	0.838498
XGBoost	0.93500	0.935783	0.93500	0.93498



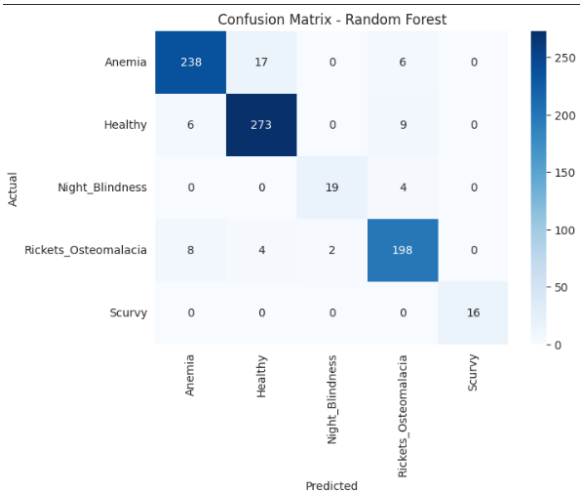


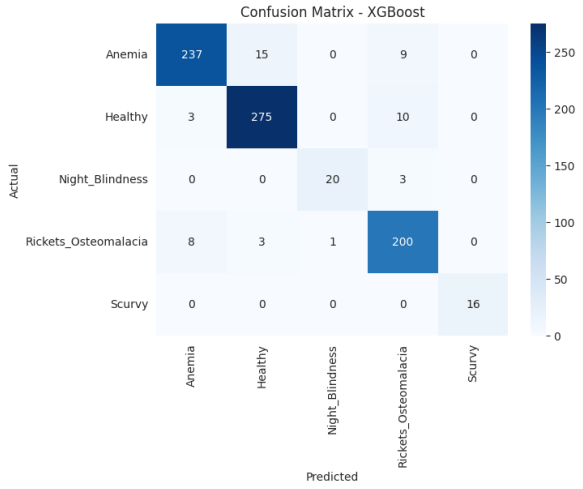


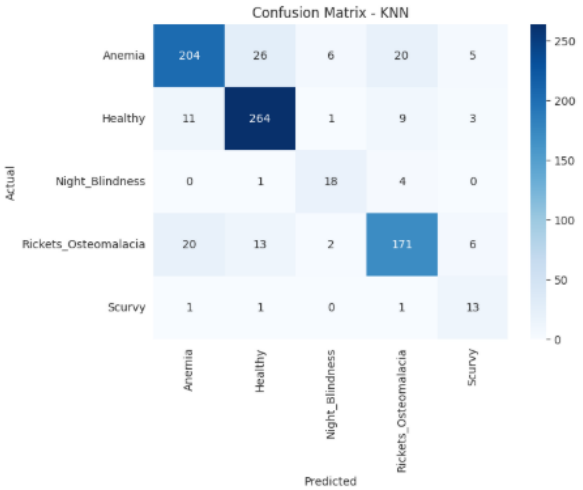




International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 50-55 March 2026-Sept 2026 www.csjournals.com







Conclusion

Vitamins are really compulsory factor deficiency of which is directly impact our health. As we are very well know that day per day many industries are shifting to digital platforms which is able to predict more correctly and less time consume. This model is able to predict early risk of vitamin deficiency on the basics of it's training on a good quality dataset. In this we are going to predict early risk of vitamin deficiency dataset was managed and processed using Python libraries such as Pandas and manipulating data, numpy for numerical calculation, matplotlib and seaborn for graphs and charts. We are using correlation matrics, confusion matrics, accuracy score in percentage, recall, precision and f1 score for getting more truth able model.

Future scope

Our dataset is having small size in volume we are able to replace it with large dataset which makes more accurate result. On the basics of that we are able to make a website or app which is specifically designed for early risk vitamin deficiency which is able to with real-time user inputs and provide the predicted result and this data is stored in database time to time maintain the dataset.

Reference

[1] S. CHEBBAWI , M. TABAA , Hassan BADIR "Investigating the impact of Vitamin D on chronic diseases using Machine Learning" The 16th International Conference on Ambient Systems, Networks and Technologies (ANT) April 22-24, 2025, Patras, Greece

[2] S. Purushotham , S. Karthik Reddy2 , S. Bharath Kumar Reddy , Mr.V.Vidhya4 , Dr. V. Sai Shanmuga Raja5 , Dr. R. Sudhakar6 1, 2, 3Students, Assi Professor, 5, 6Professor Department of Computer Science and Engineering, Dr .M.G.R Educational And Research And Institute, Maduravoyal, Chennai-95, TamilNadu, India

"Vitamin Deficiency Detection Using Machine Learning and Deep Learning Algorithms" International Journal on Science and Technology (IJSAT)

[3] A. Giustina, J. P. Bilezikian, R. A. Adler, G. Banfi, D. D. Bikle, N. C. Binkley, J. Bollerslev, R. Bouillon, M. Luisa Brandi, F. F. Casanueva, L. di Filippo, L. M. Donini, P. R. Ebeling,G. El-Hajj Fuleihan, A. Fassio, S. Frara, G. Jones,C. Marcocci,Adrian R. Martineau, S. Minisola,N. Napoli, M. Procopio, R. Rizzoli, A. L. Schafer,6 C. T. Sempos, F. Massimo Ulivieri, and J.K. Virtanen"Consensus Statement on Vitamin D Status Assessment and Supplementation: Whys, Whens, and Hows" Endocrine Reviews, 2024, 45, 625–654 https://doi.org/10.1210/endrev/bnae009 Advance access publication 27 April 2024

[4] N. Sancar and S.S.Tabrizi" Machine learning approach for the detection of vitamin D level: a comparative study" Sancar and Tabrizi BMC Medical Informatics and Decision Making (2023) 23:219

[5]T.Naga Subhash Polineni, V.Krishna Azith Teja Ganti, K. Kumar Maguluri, Dr. P.R. Sudha Rani AI-Driven Analysis of Lifestyle Patterns for Early Detection of Metabolic Disorders 02.12.2024

[6]https://www.kaggle.com/datasets/nudratabbas/vitamin-deficiency-disease-prediction-dataset