RESEARCHERID
THOMSON REUTERS

[ICRTCET-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

PREDICTION OF VITAMIN D AND CALCIUM USING DATA MINING TECHNIQUES

Nawal M. Dandekar[*1] & Rupali Sawant[2]

[*1&2]Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India

ABSTRACT

Health industry is one of the industry which can avail from data mining. Data mining techniques can be used for analyzing and predicting different diseases accurately. Vitamins and minerals are important to the body and ignoring lack of these micronutrients can lead to different diseases and problems in human body. Diagnosing these micronutrients at the earliest is necessary. This paper proposes an approach for prediction of vitamin D and calcium in human body not only based on symptoms but also on body composition, anthropometry, body mineral density whereas the traditional system focuses on symptoms. The approach uses hybrid feature selection technique followed by data mining algorithms for classification. The algorithm used for hybrid feature selection are Boruta and Random forest classifier. For classification, the algorithms used are Artificial Neural Network, Naive Bayes, Random forest, SVM, LDA. Of all the classification algorithms used, random forest gave the best accuracy followed by linear SVM and ANN.

Keywords: *hybrid feature selection, Vitamin d deficiency, Calcium deficiency, data mining.*

I. INTRODUCTION

Data mining is a process used for extraction of useful information from large datasets. It can be used for analysis and prediction purpose. The different fields where data mining can be used include healthcare, Education, Manufacturing Industry, CRM, Detection of frauds, Banking, Bio-informatics, Information Security and many more. Healthcare is one of the industry that takes advantage of data mining. Hospitals generate lots of data. Data mining can be applied in healthcare to early detect and diagnose patient, detection of fraud medi-claim and insurance, to reduce cost of diagnosis , identification of treatment plan.

Vitamins and Minerals

Vitamins and Minerals are nutrients that play a very important role in development and maintenance of human body. Conversion of food into energy , healing wounds, repairing damage cell ,boosting immune system, maintenance of bones are some of the functions of these micronutrients. They are also called as micronutrients since they are required in a very small quantity. [1]

Different vitamins required by human body are Vitamin A, Vitamin B1, Vitamin B2, Vitamin B3, Vitamin B5, Vitamin B6, Vitamin B7, Vitamin B9, Vitamin B12, Vitamin C, Vitamin D, Vitamin E, Vitamin H.

Different minerals required by human body include Calcium, Iron, Iodine, Chloride, Magnesium, Phosphorus, Potassium, Sodium, Zinc, Copper.

Lack of these micronutrients can cause various problems and diseases in human body. Peri-menopausal is a phase prior to menopause till one year after menopause. Different deficiency of vitamins and minerals found in peri menopausal woman. Woman undergo a lot of mood swings. The vitamin that helps for emotional health and for reduction of stress is Vitamin B.

Bone health needs to be taken care of as woman in this phase are at higher risk of losing bone density. Calcium is needed to keep bones strong .Vitamin D helps to absorb calcium in the body. As the age increases woman are at higher risk of developing heart diseases and getting heart attacks. Vitamin C and E are some of the vitamins that can help reduce stress in the body.

ISSN 2348 – 8034
Impact Factor- 5.070

Vitamin D
Vitamin D is needed by the body for maintaining healthy bones and teeth. It controls diabetes by managing insulin level. It helps in supporting lung functioning, cardiovascular health, the health of the immune system, brain, and nervous system.

- Some of the symptoms of vitamin D deficiency are bone and back pain, hair loss, muscle pain, muscle weakness, overweight .
- Sources of Vitamin D are sunlight and dietary source includes fish like salmon, tuna, mackerel and sardines, Egg yolk, cheese, milk and mushrooms.

Calcium
Calcium is a vital micronutrient for overall health. Calcium is used by almost every cell in our body. In our body nervous system, muscles, heart and bone use calcium.

- Some of the symptoms of Calcium deficiency are osteoporosis , muscle cramp, dry skin, fatigue, rickets [18]
- Sources of Calcium are milk, cheese, yogurt, spinach, almond, soya bean, white bean, fishes like salmon, sardine.

Osteoporosis
Osteoporosis is a medical condition in which bone density decreases and the body stops producing bones as it did earlier. It happens due to hormonal changes, or deficiency of calcium or vitamin D. It is seen in both males and females. It affects mostly women after menopause, since the estrogen (the hormone which protects against osteoporosis) decreases suddenly. A sedentary lifestyle is one of the reasons for osteoporosis.

Osteopenia
Osteopenia is a medical condition in which density of bone decreases, leading to weak bones and thereby increasing the risk of fracture. It is considered to be a precursor to osteoporosis. Because of the loss of estrogen in post-menopausal women, Osteoporosis and Osteopenia is seen. T score value is used to diagnose Osteopenia and osteoporosis. If the T score value is in between -1.0 and -2.5 it indicates Osteopenia, and if the T score value is lower than -2.5 indicates osteoporosis.

Bone health is extremely important in women. Globally, many studies have been reported which highlight the concern of low bone density and also report the increased prevalence of osteoporosis in post-menopausal women.

However, very few researches are done on peri menopausal women because of which it was considered significant to see the bone changes which may begin to occur much earlier in women who are approaching menopause. Thus awareness for the same can be created through such researches that bone changes which can start at early age can be prevented by taking proper diet, supplements and following good lifestyle inclusive of exercise, etc

If Vitamin D or Calcium is detected early in peri menopausal woman, proper medication can be started. Through our study we can detect whether a person is Vitamin D or calcium deficient using parameters which includes anthropometry(Height, Weight, BMI) ,body composition (fat mass, bone mass, muscle mass, visceral fat) , 24 HR-DIET RECALL and bone mineral density(T-score).

The organization of this paper is as follows. In section 2, A detailed literature review is presented. In section 3, The methodology and different algorithms used for the implementation of the system are mentioned .In section 4,evaluation metrics parameters are discussed. In section 5, results are presented. Finally, section 6 discusses conclusion and summary of the paper.

ISSN 2348 – 8034
Impact Factor- 5.070

II. LITERATURE REVIEW

Different data mining techniques have been used by different researchers to predict or diagnose different diseases. Data mining classifiers such as Naive Bayes, Decision tree, SVM, ANN and many more have been used by the researchers to predict different diseases in human body including heart, liver, diabetes.

Mutammimul Ula et.al[1] have created an expert system which can be used to diagnose vitamins and minerals deficiency in the Body. The expert system was created using 46 symptoms relevant to deficiency of 11 vitamins and 6 minerals. The decision table was created to form rules that helped in detection of deficiency from the symptoms. The search method used in the expert system was forward chaining method. Certainty factor method was used to deal with problem of uncertainty. It basically treats a symptom depending on the weights assigned to it that was acquired during observation.

Dony Novaliendry et.al[2] have a created a desktop-based application for expert system. For building an expert system , decision tree method was used. The decision tree was formed from 13 vitamins and 50 deficiency symptoms of these vitamins. The forward chaining method along with depth first search was used as searching technique in the decision tree.

Subinay Datta [3] the relationship between transmission level of Vitamin D with the important explanatory variables. The variables used in the study were BMI, gender, history of depression, age, smoking habits, food habits, FEV1, comorbidity and some more. The study was analyzed using Linear regression model. Lack of vitamin D is the reason for growth retardation and rickets. Insufficient Vitamin D in adults speed up Osteoporosis and Osteopenia. Some of the studies show that Vitamin D is one of the reasons for diseases like cancer, chronic obstructive pulmonary disease (COPD) and cardiovascular diseases. The result of the study can be summarized as 53% population had inadequate range of Vitamin D in the body.9% were Vitamin D deficient. Transmission of vitamin D in people of higher age is low. Men had more Vitamin D in serum than women. Vitamin D is removed quickly from the blood rather than fat. In short, it can be concluded that anthropometry does has an effect on the Vitamin D levels.

Mafazalyaqeen Hassoon et.al [4] have used a new method comprising of Genetic Algorithm (GA) with Boosted C5.0 classification method to predict liver disease .The dataset was composed of 10 important features along with two classes. Boosted C5.0 algorithm generated 92rules for predicting liver disease. So Genetic algorithm was used to optimize and reduce number of rules which were extracted from Boosted C5.0 algorithm and also to find attributes which contribute more to patients diagnosis. The method gave better performance and throughput compared to earlier work.

Messan Komi, Jun Li et.al [5] have used different data mining algorithms like SVM , ANN, GMM, Logistic Regression, ELM for early prediction of diabetes. The total attributes that were used in prediction of diabetes is 7.The hidden layers used in ANN were 2 and neurons in both the layer were 5 each. The network function used was sigmoid. Out of all the techniques used ANN gave the highest accuracy of 89%.

Tejaswini U. Mane [6] have used a hybrid method for prediction of heart disease. The system was designed by combining Improved K-Means for clustering and decision tree algorithm i.e. ID3 for classification purpose. Improved K-means was used instead of simple K-means in this system to improve the accuracy of the centroids in the cluster. The system can be used as second opinion as it is helpful in prediction of heart disease on basis of 13 parameters of which some are age, resting Bp , cholesterol ,chest pain, Thalac.

Kamal Nayan Reddy Challa et.al [7] have implemented an automated diagnostic models using Multilayer Perceptron, BayesNet, Random Forest and Boosted Logistic Regression for early prediction of Parkinson's disease. Out of all the machine learning algorithms used Boosted Logistic

Regression performed best with an impressive accuracy of 97.159 % and the area under the ROC curve was 98.9%.

Sumana et.al [11] have implemented a system using a hybrid model where for preprocessing K-means was used for feature selection , the algorithm used are Best First search and Correlation based feature selection (CFS).12 different classification techniques were used and applied on five different datasets.

III. METHODOLOGY

Figure 1.System architecture

A. Data Collection

The dataset has records about 200 perimenopausal women in the age group 40-50 years. The dataset had 145 attributes including anthropometry (Height, Weight, BMI), body composition (fat mass, bone mass, muscle mass, visceral fat), 24 HR-DIET RECALL and bone mineral density (T-score).

B. Data Preprocessing

Data preprocessing is a step in data mining which involves converting raw data into an unambiguous style. The dataset had hardly any missing values. The missing values were filled by taking into consideration the type of data. If the attribute had categorical values then most frequent value was filled in. If the attribute had numeric values then the missing values were substituted with the mean.

Normalization is a scaling method which is done, when there is large difference in the values of the data. Min-Max normalization technique was used in this paper. v –value of the attribute

V1 –value of the attribute after Min-Max normalization min(a)-minimum value of the attribute max(a)-maximum value of the attribute

$$V1 = \frac{v - \min(a)}{\max(a) - \min(a)} \qquad (1)$$

C. Feature Selection

It is the most important step in data mining algorithm, where subsets of the feature available from the dataset are selected for application of a learning model. Feature selection aims to improve data mining performance. All the 145 attributes were given as input to hybrid feature selection technique. The hybrid feature selection technique will give most important features from the data. Hybrid feature selection was done combining Boruta algorithm and random forest classifier.

BORUTA

All the 145 attributes were given to Boruta algorithm. It selected 23 most important features that contribute most to our final predictor (i.e) Deficiency of Vitamin D and Calcium from the entire dataset. Boruta uses Random forest classifier technique for training and feature importance measure is applied to assess the importance of each feature. The default applied is Mean Decrease Accuracy. Higher the value of the result more important is the feature.

RANDOM FOREST

The second algorithm used is Random forest on all 145 attributes. The tree-based strategies which is used by random forests ranks how well they improve the purity of the node. This mean decrease in impurity over all trees (called gini impurity).Nodes with the greatest decrease in impurity happen at the start of the trees, while notes with

the least decrease in impurity occur at the end of trees. Finally, by pruning trees below a particular node, a subset of the most important features is created.18 most important features were taken from random forest classifier.

The union of both the feature selection technique was taken. So finally, 32 attributes came from hybrid feature selection. These 32 attributes were used for prediction of vitamin D and Calcium in human body.

D. Algorithms for Classification

The features from the hybrid method are used as final attributes for the classification algorithm.

NEURAL NETWORK

Artificial neural network is a computing system whose structure and task are similar to biological neural networks. Like humans, ANN's learn by examples. A neural network which is trained can mimic as an expert in that field. The neural networks are extensively used to recognize patterns in data.

The features from hybrid method feature selection are given to input layer of neural network .The input layer has neurons equal to our final selected features. Each layer has finite number of neurons .So in this system there will be 31 neurons in input layer. The hidden layer which is in between input and output layer helps neural network in learning more complicated peculiarity of the data .There are two hidden layer used in the network which has 7 and 3 neurons each .The output layer will have one neuron that will predict whether the person has deficiency of Vitamin D/Calcium or not.

The neural network output is given by the formula,

$$y = \sum_{i=1}^{n} (x_i * w_i) + b \qquad (2)$$

where inputs are denoted by x1,x2,..xn ,weights are denotes by w1,w2,..wn weights show how important is a particular input to the final output and b is a bias term used in the network.

RANDOM FOREST

Random forest is a supervised classification algorithm. Random forest is an ensembler meaning it trains multiple decision trees on different parts of the same training dataset and gives that class as the output which is mode of the classes output given by an individual tree . This is done to overcome over -fitting problem. The larger number of trees to train the data more accurate is the result of the prediction. In health care industry .Random forest can help to identify a disease from patients medical reports.

In random forest each tree is constructed by following steps

- ξ First of all 2/3rd training data is used for training each decision tree .The number of trees used for building our model is 50.
- ξ Out of all the predictor attributes or features some are chosen randomly and the best split technique is used
  on them to split the node

Each of these decision tree will predict different class for the same test feature. Then by voting the class which gets highest votes or rather in simple terms all the 50 decision trees that were created predicts a particular class and the class predicted maximum among them will be the final prediction of the random forest algorithm.

NAIVE BAYES

Naive Bayes classifier is built on the technique of Bayesian theorem. Bayes theorem works on the principle of conditional probability. The conditional probability is that probability that tells whether a particular event will

happen depending on an event that has already happened. It is simple to implement but sometimes performs better than complex sophisticated classification.

Algorithm

- Naive Bayes classifier presumes that attributes are independent of each other. So for classifying whether the person is vitamin D/calcium deficient Naive Bayes classifier first transforms the dataset into a table of frequencies.
- The likelihood table is created by finding probabilites.
- The posterior probability for each class is computed using naive bayes equation. The result of the prediction is that class which has the highest posterior probability.

The posterior probability is calculated by the formula

(3)

P(a|y) - posterior probability of category or class(i.e a)

when the predictor is known (i.e y)

P(a) - prior probability of category

P(y|a) - the likelihood which is the probability of predictor given the category.

P(y) -prior probability of predictor.

LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis (LDA) is a supervised classification algorithm. It is developed by R.A. Fisher in the year 1936.It is simple and robust and like complex algorithms it gives good accuracy. The aim of the LDA is projecting the data onto a lower-dimensional space and to get a good separation of two classes.

Algorithm

- Calculate the n-dimensional mean vector for all the categories in the data
- Calculate the scatter matrix for the two categories (one would be for in between the two categories and the other would be within the categories.)
- Calculate the eigen vector(v1,v2,...,vn) and their respective eigen values (Ɛ1, Ɛ2,.. Ɛn) for the said scatter matrix.
- Rank the eigen vectors in the descending order of eigen values. Select some j number of eigen vectors who have the highest values so as to form a matrix (n*j).
- With the help of this new formed matrix (n*j) convert the data samples to a new subspace.

SUPPORT VECTOR MACHINE

In SVM classifier, in n-dimensional space each and every data point is projected where the value of each attribute is a value of a particular co-ordinate in the space. In general it performs classification, by finding a hyper-plane that separates two classes clearly. First it tries to separate linearly by creating hyper plane. If linearly it is not possible then kernel trick is applied.

Different kernels are as follows: polynomial, RBF, sigmoid

Equations for polynomial

ISSN 2348 – 8034
Impact Factor- 5.070

$$K(x_1,x_2)=(x_1 \cdot x_1+1)^d$$

(4)

Equation for RBF

$$K(x_1,x_2)=e^{-\gamma\|x_1-x_2\|^2}$$

(5)

where Û is used to set the widening of the kernel.

IV. EVALUATION METRICS

A confusion matrix is a summary table that helps to understand the performance of the classifier used.

Table.1 Confusion Matrix

	Predicted (yes)	Predicted (no)
Actual(yes)	TP	FN
Actual(no)	FP	TN

True positive (TP) -The actual value is positive and predicted value is positive too

False positive (FP) - The actual value is negative but predicted value is positive.

ISSN 2348 – 8034
Impact Factor- 5.070

False negative (FN)- The actual value is positive but predicted value is negative.

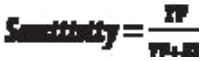
True negative (TN)- The actual value is negative and predicted value is negative too.[4]

Accuracy
Accuracy of a classifier is the ratio of correct predicted values to the total number of values.

(6)

Sensitivity /Recall/True positive rate

Sensitivity of a classifier is the ratio of correctly predicted positive values to the total number of positive values.



(7)

Specificity /True negative rate

Specificity of a classifier is the ratio of correctly predicted negative values to the total number of negative values.

(8)

Precision
Precision of a classifier is the ratio of correctly predicted positive values to the total number of positive values.

(9)

V. RESULTS AND DISCUSSION

Table 2. Without Hybrid feature selection

Classifier	Accuracy	Sensitivity	Specificity	Precision
ANN	54.54%	0%	100%	54.54%
Naive Bayes	63.63%	40%	83.33%	66.67%
Random forest	74.54%	52%	93.33%	86.66%
LDA	67.54%	50%	60%	58%
SVM(linear)	80%	80%	80%	76.92%

SVM (polynomial)	81.81%	84%	80%	77.78%
SVM (radial basis)	54.54%	0%	100%	54.54%
SVM(Sigmoid)	54.54%	0%	100%	54.54%

Table 3. With Hybrid feature selection

Classifier	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
ANN	90.9%	92%	90%	88.46%
Naïve Bayes	83.63%	84%	83.33%	80.76%
Random forest	98.18%	100%	96.67%	96.15%
LDA	85.45%	80%	90%	86.96%
SVM(linear)	96.36%	96%	96.67%	96%
SVM (polynomial)	76.36%	60%	90%	83.33%
SVM (radial basis)	87.27%	76%	96.6%	95%

SVM(Sigmoid)	90.9 %	80 %	93.33 %	91.67 %

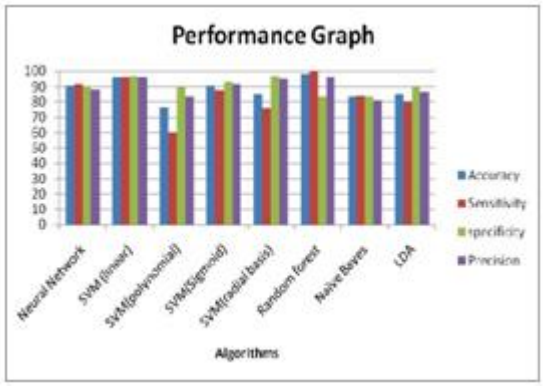


Figure 2: Performance Graph of the system

By using hybrid feature selection, the 145 attributes or features are reduced to 31 which are given as input to the classifiers for prediction of deficiency of Vitamin D and calcium. Because of this not only computational time decreases but also the accuracy of the system increases.

By using hybrid feature selection, it can be seen that performance of the system is increased.

VI. CONCLUSION

Vitamins and minerals are necessary for smooth functioning of the body. The paper has proposed a new approach for prediction of deficiency of vitamin D and Calcium in women. The proposed system can help women to take appropriate diet or multivitamins so as to avoid complications in future .The system for prediction of Vitamin D and Calcium gave better results when hybrid feature selection was used. The accuracy of Random forest was 98.18% followed by linear SVM and ANN whose accuracy was 96.36% and 90.9% respectively. The least accuracy was shown by SVM (polynomial) of 76.36%. The inference that can be drawn from this paper is proper attributes are important for prediction along with proper feature selection and classifiers.

In future, the research can be done using different data mining classifiers and feature selection technique. Also large dataset can be used for the research.

REFERENCES

1. *Mutammimul Ula, Mursyidah, Yana Hendriana, Richki Hardi "An Expert System for Early Diagnose of Vitamins and Minerals Deficiency On The Body" 2016 International Conference on Information Technology Systems and Innovation (ICITSI) Bandung -Bali, October 24 -27, 2016.*
2. *Dony Novaliendry , Cheng-Hong Yang, Denno Guara Labukti A.Y "The Expert System Application For Diagnosing Human Vitamin Deficiency Through Forward Chaining Method." ICTC 2015.*
3. *Subinay Datta, Mrinal Pal, Anshuman De "The Dependency of Vitamin D Status on Anthropometric Data", Malays J Med Sci. May-Jun 2014; 21(3): 54-61*
4. *Hassoon, Mafazalyaqeen, Mikhak Samadi Kouhi, Mariam Zomorodi Moghadam, and Moloud Abdar "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction." In Computer and Applications (ICCA), 2017 International Conference on, pp. 299-305. IEEE, 2017.*
5. *Komi, Messan, Jun Li, Yongxin Zhai, and Xianguo Zhang "Application of data mining methods in diabetes prediction." In Image, Vision and Computing (ICIVC), 2nd International Conference on, pp. 1006-1010. IEEE, 2017.*
6. *Mane, Tejaswini U. "Smart heart disease prediction system using Improved K-means and ID3 on big data "In Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on, pp. 239-245. IEEE, 2017.*

7. *Challa, Kamal Nayan Reddy, Venkata Sasank Pagolu, Ganapati Panda, and Babita Majhi "An improved approach for prediction of Parkinson's disease using machine learning techniques." In Signal Processing, Communication, Power and Embedded System (SCOPES),2016 International Conference on, pp. 1446-1451. IEEE, 2016.*
8. *Mrigen Kr. Deka, Anil Kumar Malhotra, Rashmi Yadav, Shubhanshu Gupta "Dietary pattern and nutritional deficiencies among urban adolescents". Journal of Family Medicine and Primary Care July 2015 : Volume 4 : Issue 3*
9. *Mohammad Hossein Tekieh, Bijan Raahemi. "Importance of Data Miningin Healthcare: A Survey." 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.*
10. *Jothi, Neesha, and Wahidah Husain. "Data mining in healthcare–a review." Procedia Computer Science 72 ,pp. 306-313,2015.*
11. *Sumana, B. V., and T. Santhanam "Prediction of diseases by cascading clustering and classification." In Advances in Electronics,Computers and Communications (ICAECC), 2014 International Conference on, pp. 1-8. IEEE,2014.*
12. *Bala, Suman, and Krishan Kumar. "A literature review on kidney disease prediction using data mining classification technique." International Journal of Computer Science and Mobile Computing 3, no. 7, pp. 960-967, 2014.*
13. *Amin, Syed Umar, Kavita Agarwal, and Rizwan Beg. "Genetic neural network based data mining in prediction of heart disease using risk factors." In Information & Communication Technologies (ICT), 2013 IEEE Conference on, pp. 1227-1231. IEEE, 2013.*
14. *Heinonen, Petri and Mannelin, Marjo and Iskala, Hannu and Sorsa, Aki and Juuso, Esko "Development of a Fuzzy Expert System for a Nutritional Guidance Application." IFSA/EUSFLAT Conf, pp. 1685-1690, 2009.*
15. *Cilimkovic, Mirza. "Neural networks and back propagation algorithm." Institute of Technology Blanchardstown, Blanchardstown Road North Dublin 15 (2015).*
16. *Weaver, C. M., D. D. Alexander, C. J. Boushey, Bess Dawson-Hughes, J. M. Lappe, M. S. LeBoff, S. Liu, A. C. Looker, T. C. Wallace, and D. D. Wang. "Calcium plus vitamin D supplementation and risk of fractures: an updated meta-analysis from the National Osteoporosis Foundation." Osteoporosis International 27, no. 1 (2016): 367-376.*
17. *Pahwa, Kanika, and Ravinder Guide Kumar. "Prediction of Heart Disease using Hybrid Methodology of Selecting Features." (2017).*
18. *Vitamin and mineral requirements in human nutrition Second edition Authors: World Health Organization, Food and Agricultural Organization of the United Nation.*