2019 International Conference on Computer Communication and Informatics (*ICCCI* -2019), Jan. 23 – 25, 2019, Coimbatore, INDIA

Food Recommendation using Machine Learning for Chronic Kidney Disease Patients

Anonnya Banerjee, Alaa Noor, Nasrin Siddiqua, Mohammed Nazim Uddin
School of Science, Engineering & Technology, East Delta University, Chittagong, Bangladesh.
anonnyabanerjee95@ymail.com, nazim@eastdelta.edu.bd

Abstract—Chronic Kidney Disease (CKD, also known as the Chronicrenal disease) is a communal problem to the public with an escalating in either technologically advanced or advancing countries. 10% of the populations internationally are diagnosed with the disease of chronic kidney disease (CKD), and many deaths each day because of poor access to proper treatment or absence of awareness. If necessary, precaution is not taken at the right time the treatment gets complicated and possibly reach the final stage. Thus, to provide a better dietary solution our paper proposes a model to basis of their potassium level in blood. Classification of the patients is implemented recommend food for patients suffering from kidney disease on the using WEKA and then further using query-based matching we recommended food for the identified levels based on the seriousness of the disease.

Keywords-Chronic Kidney Disease; Blood Potassium Levels; Diet plans; Machine learning; Potassium zone

I. Introduction

Chronic Kidney Disease (CKD) is characterized as kidney auxiliary harm caused by breaking down capacity of kidneys and is normally estimated with GFR (Glomerular Filtration Rate), where the Glomerular Filtration Rate drops, (GFR) < 60 ml/min/1.73 m2 for three months or so. Such degeneration is problematic to waste and excess fluids formation in the body and impacts the performance of the body, potentially prominent to complications. It often remains overlooked and undiagnosed until the condition of the individual gets worse slowly over time. The disease can reach to end-stage renal disease (complete kidney failure) which takes place when kidney function gets deteriorated to a point where dialysis or kidney transplantation turns out to be the only way for survival. People suffering from blood pressure, diabetes and having a family background of being affected by such diseases or suffering from chronic kidney disease earlier are most likely to suffer from kidney diseases. The rate of death occurrence rose to 956,000 in 2013 from 409,000 in 1990 [1]. It is considered to be a life-threatening disorder affecting a large number of people.

However, this rise in data volume automatically involves the data to be repossessed when needed. Health administrations today are capable of generating and accumulating a bulk amount of data. With the help of data mining and classification techniques in medicinal applications, it is possible to identify relationships and models that support forecasting and decision-making process for analyzing and action planning [2], an action planning may include the do and don't(s) of daily life activities for instances suggesting a meal or diet of an individual. For a CKD patient, following an appropriate diet plan can help to lessen the growth of CKD. So, it is very necessary to identify a suitable diet based on the patient's health condition such as based on the estimated Glomerular Filtration Rate (eGFR), which can be categorized into five stages such as stage0, stage 1, stage 2, stage 3 and stage 4. Till stage 2, patients are considered to be within the safe range or they are able to cope up with the renal functions without gathering excretory products savor potassium or surplus urea in the blood. Henceforth, patients in the stage 0, stage 1, and stage 2 don't require any vital changes in their diet plan. But for patients in the stage 3 and 4 are in difficulty of keeping up the balance of minerals, electrolytes, and liquids inside their body. Be that as it may, for patients in the stage 3 and 4 are in trouble of keeping up the adjustment of minerals, electrolytes, and fluids inside their body.

Diet plans of CKD patients not simply rely upon the stage of the disorder yet additionally with different conditions, such as the level of blood potassium, urea, sodium etcetera [4]. In this paper, diverse data mining techniques have been executed on a dataset containing data about patients' determination for CKD. These methods are – Naïve Bayes, Support Vector Machine, and Random Forest, but the core center is around blood potassium level to distinguish the reasonable eating routine arrangement for a CKD patient.

Potassium is a mineral found in huge numbers of the nourishments we eat. It assumes a job in keeping our pulse and muscles working right. It is the activity of regular kidneys to keep the perfect measure of potassium in our body. In any case, when the kidneys are not working properly, we frequently need to confine certain nourishments that can expand the potassium in our blood to an unsafe level. One may feel some shortcoming, numbness and shivering if the potassium is at an abnormal state. In the event that potassium turns out to be too high, it can cause an unpredictable heartbeat or a cardiac arrest.

The paper is divided into five sections. Section II represents the relevant works done so far for prediction of CKD using data mining techniques and how it is implemented in dietary system. In Section III, methodology consists of system model of the recommended food plan and a brief description on how the algorithms are implemented on the

978-1-5386-8260-9/19/$31.00 ©2019 IEEE

model; this is done to provide a better understanding of the whole paper. Experimental results and result analysis are presented in the Section IV. Section V, consists of conclusion and additional future scope.

II. Related Work

In recent days, data mining techniques is contributing a lot in the health care system to solve or detect various diseases. A number of researchers have used different algorithms and methods to classify or to detect CKD, Diabetes and other diseases as well.

The researchers in [5], proposed a work using Naïve Bayes with OneR attribute selector for detecting CKD. It avoids renal disease getting to a more serious or complex level. The suggested system develops rules for present stage in order to proceed the treatment accordingly. Nevertheless, an automated machine learning model was developed by the researcher of the paper [6], to forecast CKD and determine 24 attributes associated to it. Feature selection was made to differentiate the important attributes for detection based on their predictability rank them. Attributes at different level were identified. The results were then assessed using three following classification algorithms such as k-nearest neighbor, random forest and neural networks respectively.

The authors of [7], designed a food recommendation service which formulates a customized service for coronary heart disease patients. It develops a diet taking some important information on accounts such as any significant disease symptom previous family history of the patient and precise food choice if there are any. The service provided is personalized unlike the conventional services that are mostly used. Whereas for the study in [11], the authors used Self-Organizing Map (SOM) and K-mean clustering for food clustering analysis and implemented Food Recommendation System (FRS) for diabetic patients. It recommends the substituted foods based on the connection of eigh*t* significant nutrients of a diabetic patient. Nutritionists evaluated the FRS which performed well and it is useful for diabetic patients. In [12]**,** the researchers developed a machine learning model to increase the quality of CKD diagnosis by using feature selection and ensemble learning. In this paper, to improve the classification of CKD, Correlation-based Feature Selection (CFS) was used for features selection and AdaBoost was used for ensemble learning. Classifiers such as KNearest Neighbor algorithm (kNN), Naive Bayes and Support Vector Machine *(*SVM) were used. The best result was obtained by kNNclassifier with CFS and AdaBoost with 0.981 accuracy rate.

In the following study [8], Chetty et al. utilized order systems that are worked by utilizing Wrapper technique. The strategy decreases the quantity of credits to anticipate CKD. Attribute evaluator and bread-first search were afterwards an improvement for detecting CKD on reduced dataset was observed. Finally, Vijayarani et al. [9], made use of algorithms such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) to predict renal related disease. Central motive of the research was to find out the performance using their execution time and accuracy. The accuracy of ANN outweighs SVM, which is 87%, thus ANN was better. Lastly, Ahmed et al. [10], came up with diagnosis technique with fuzzy logic. MATLAB was utilized; consisting of an in-built toolbox to patterned the state of the patient's kidney.Data has been retrieved from Birdem Hospital in Dhaka. Few attributes have been measured and the fitness or the wellbeing is measured within a boundary between 0 to method reduces the number of attributes to predict CKD.

III. MethodoloGy

This proposed system is used to predict the actual number of CKD patients using machine learning techniques and after evaluating the accuracy, a query is used to provide a list of recommended food based on the seriousness of the disease. In fig 1, the system architecture of the proposed system is shown.

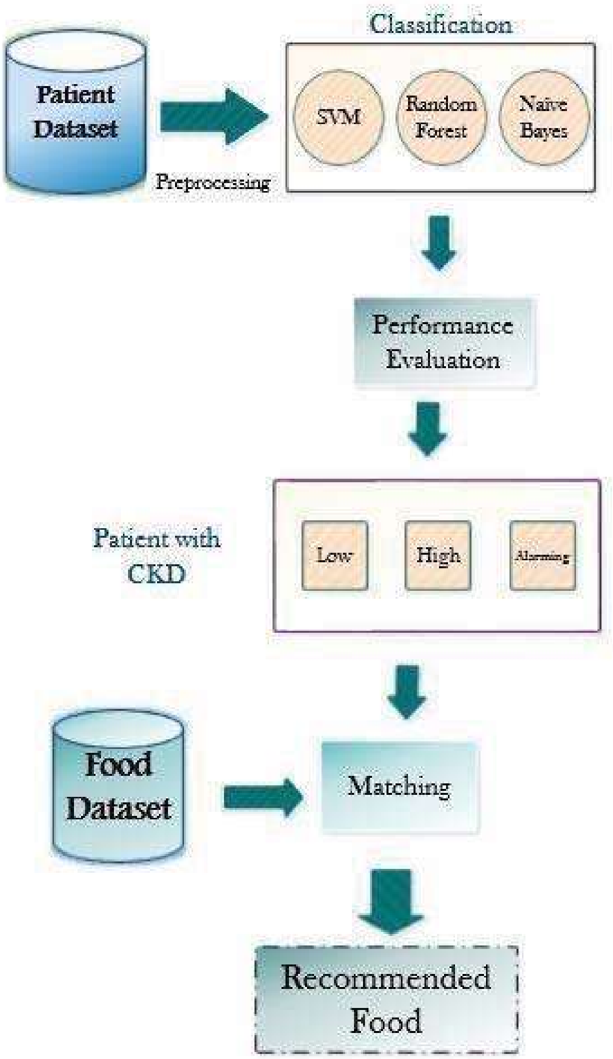

Figure 1. Proposed System Architecture of Recommended Food System

Classification procedure that is also known as supervised learning techniques are most usually used in data mining to classify the data in a raw data. Some classification methods used are Bayesian classification, decision tree, support vector machine (SVM), neural networks, association-based classification.

In the study, various classification methods are applied on the modified data. Three classifiers have been selected that were based on the attributes of dataset.

A. Naive Bayes

The Naive Bayesian classifier [16], is grounded on Bayes' formula with self-determining norms between predictors. The Naive Bayesian classifier is pretty simple to shape, by means of no complex iterative attributes estimate that makes it principally beneficial for vast datasets. In spite of its straightforwardness, the classifier often does astonishingly good and is commonly used because it frequently outclasses more sophisticated classification procedures. This suggestion gives a system for ascertaining the posterior likelihood, P(c|x), from P(c), P(x), and P(x|c). For P(x|c) P(c) category, the system will forecast the diseases which user is probably to get, and will recommend required food with

$$P(c|x) = P(x) \tag{1}$$

Here,

- **P(c|x)** is posterior possibility of target on given attribute
- **P(c)** is previous possibility of class.
- **P(x|c)** is the prospect that is probability ofpredicator given class.
- **P(x)** is previous likelihood of predicator

B. Random Forest Classifier

Random Forest Classifier (random decision forests) is an ensemble classification algorithm. Ensemble algorithm uses several learning algorithms to obtain better predictive validation and also merge more than one similar or different kind of algorithms for classifying data. A series of decision trees are created from a randomly chosen subset of the training set. The final class of the dataset tested is decided by totaling the votes produced by each of the decision trees. Random forest classifier involves a group of classifier that are tree-structured [17].

{h(x, Θk),k=1,…}, where Θk is independent indistinguishably conveyed random vectors and here at input x each tree creates a vote for most common class.

Random forests contain three essential parameters where:

Node size = number of each node at each terminal;

$n_{tree}$is the number of decision trees constructed as part of the regression tree ensemble;

m_{try} = number of forecaster variables randomly sampled as contenders at each decision tree node spit, $m_{try}$is measured as follows:

$$m_{try} = \frac{p}{3} \tag{2}$$

C. Support Vector Machine

Support vector machine (SVM), is a machine learning algorithm that is based on hypothetical learning theory. It utilizes a nonlinear mapping to restore the information in to an upper measurement. It depends on the origination of decision planes that distinguishes decision limits. Decision plane is a discrete hyperplane - made in descriptor space training data and compound, are characterized in view of side of the hyper plane it is situated. It accepts data as an input and detects for every one of them, which of the two probable classes includes the input, also for which support vector machine is a non-probabilistic binary linear classifier. SVM has very slow training time but in terms of making predictions it is highly accurate.

D. Fuzzy Lookup

The Microsoft Research created Fuzzy Lookup Add-In for Excel that executes fuzzy matching of textual information in Microsoft Excel. It very well may be utilized to recognize fuzzy duplicate rows inside a particular table or also fuzzy join familiar amid two unique tables. The synchronizing is strong to a wide assortment of faults such as spelling errors, equivalent words, added or missing data, and acronyms. For example, it may distinguish that the lines "Ms. Anonnya Banerjee", "Banerjee, Anonnya." and "Ana Banerjee" all indicate same core entity, returning with a similar score alongside each match. While the default arrangement functions excellently for a wide range of textual information, for example, item names or client addresses, the matching may likewise be modified for particular domains and languages [18].

E. Recommended Food Using Fuzzy Match

Microsoft Power Query is an Excel add-in which can be utilized for data revelation, reshaping the data and consolidating data originating from various sources. In this step, the recommended food is attained by using two queries to identify the three different labels of the CKD patients. Both the preprocessed datasets are merged together in the power query and are matched according to their labels to give or provide a list of food based on their seriousness of the disease.

The food is recommended based on the three classified levels. The recommendation is done on the basis of, high potassium food for the ones with low potassium level, safe for the ones with safe level and low for the ones with alarming blood potassium level.

IV. Experimental Results & Evaluation

TABLE I. Attributes & Description

Attributes	Description
Blood pressure (mm/Hg)	Numerical Values
Sugar	Nominal Values (0,1,2,3,4)
Albumin	Nominal Values (0,1,2,3,4)
red blood cell	Nominal Values (normal, abnormal)
blood glucose random (mgs/dl)	Numerical Values
blood urea (mgs/dl)	Numerical Values
Serum creatinine (mgs/dl)	Numerical Values
Potassium (mEq/L)	Numerical Values
Hemoglobin (gms)	Numerical Values

packet cell value	Numerical Values
white blood cell count (cell/cmm)	Numerical Values
Hypertension	Nominal Value (Yes, No)
diabetes mellitus	Nominal Value (Yes, No)
Appetite	Nominal Value (Good, Poor)
pedal edema	Nominal Value (Yes, No)
Class	Nominal Value (ckd,notckd)

This study uses two publicly accessible datasets [13] [14] which is taken from the UCI repository and Data World website, which consists the records of 400 patients collected from Apollo Hospital India and 61 varieties of raw food.

A. Data Preparation for Patient's Dataset

Originally, the dataset of patients contains 25 attributes with people aged between 2 to 90 years old. Attributes and values are show in the Table 1.

In this dataset, there are 400 instances where 250 are CKD and 150 are not CKD. 25 attributes with 1 zone class, out of them 11 are numeric and 14 nominals. There are missing values and class division is between CKD and not CKD.

After the evaluation of CKD and NOTCKD has been done we included an extra column named as "ZoneClass". Based on the potassium level in blood, the instances have been categorized into four categories such as: -

- If the blood potassium level is between 0 to 3.4, then it is identified as LOW.
- If the blood potassium level is between 3.5 to 5, then it is identified as SAFE.
- If the blood potassium level is between 5.1 to 6, then it is identified as CAUTION.
- If the blood potassium level is greater than or equals to 6, then it is identified as DANGER [15].

For this study, the range above and equal to 5.1 is categorized into 1 level named as ALARMING. To add the new column "ZoneClass", we used an add-in of Microsoft Excel named Power Query. The following steps are used to evaluate the column: -

- if potassium > 0 &&<3.5 = "Low"
- elseif potassium >= 3.5 &&<=5.0 = "Safe"
- elseif potassium >= 5.1 = "Alarming"
- else "noData".

B. Data Preparation for Raw Food Dataset

Initially, the dataset consists of 16 attributes including the name of the food, total, fats, sodium level, calories, potassium level and so on. In the final dataset, an extra attribute of blood potassium level has been added as the core targeting attribute based on which suitable foods are recommend to CKD patient. Therefore, we are classifying it in 3 levels as LOW, SAFE and ALARMING. Where, if potassium in food,

- >= 5.1, then it is suitable food for LOW.
- >= 3.5 and <= 5.0, then it is suitable food for SAFE
- >=0 and <3.5, then it is suitable food for ALARMING.

To add the new column in the original dataset, we used add-in of Microsoft Excel named Power Query and generated a query.

TABLE II. RESULTS & COMPARISON

Method	CA	F1	Precision	Recall
RANDOM FOREST	0.997	0.997	1.000	0.996
NAÏVE BAYES	0.955	0.955	0.960	0.955
SVM	0.982	0.983	0.983	0.983

In this study, 10-fold cross validation technique is used on the model, where 9 folds are used for training and 1-fold for testing, the overall process is repeated until all the 10 individual folds have been used for testing and the results are evaluated using Naïve Bayes algorithm gives an accuracy of 95.5%, Support Vector Machine (SVM) gives an accuracy of 98.25% and Random Forest algorithm gives an accuracy of 99.75%. It is observed the Random Forest algorithm performs better to predict CKD than Naïve Bayes and SVM. Table II shows the comparison and results. There are distinctive parameters in view of which the performance of a specific classifier method is estimated. Following parameters are utilized for better idea of the obtained results.

Accuracy:

Accuracy denotes the closeness of the calculated value to the real or correct value. The general formula of Accuracy is: -

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

Precision:

Precision explains the number of true positive estimated data. The equation of precision is: -

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

Recall:

Recall explains the number of true predictable data provided all true actual data. The equation for recall is: -

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

F- Measure: -

F-measure is the harmonic average of precision and recall. The equation of f-measure is: -

$$F - Measure = 2\,\frac{Precision * Recall}{Precision + Recall} \quad (6)$$

C. Recommended Food Chart

From the food dataset the following recommended foods were listed within the range. 44 high potassium foods are suggested for CKD patients with Low blood potassium level, 10 moderate range food items were listed for people within the safe range and 7 low potassium food were identified for CKD patients who are at risk or Alarming range.

TABLE III. RECOMMENDED FOOD INTO 3 IDENTIFIED CLASSES

Low (0-3.4)	Safe (3.5 – 5.0)	Alarming (>=5.1)
Apple	Avocado	Green Onion
Asparagus	Cucumber	Iceburg lettuce
Banana	Grape fruit	Lemon
Bell pepper	Green cabbage	lime
Blue Crab	Leaf	Pineapple
Broccoli	Lettuce	Swordfish
Cantaloupe	Onion	Tilapia
Carrot	Pear	
Catfish	Radishes	
Cauliflower	Strawberries	
Celery		
Clams		
Cod		
Flounder/sole		
Grapes		
Halibut		
Green Beans		
Haddock		
Honey dew melon		
Kiwi fruit		
Lobster		
Mushrooms		
nectarine		
Ocean preaches		
Orange roughly		
Oyster		
Peach		
Plums		
Pollock		
Potato		
Rainbow trout		
Rockfish		
Salmon Atlantic		
Salmon Pink		
Scallops		
Shrimp		
Summer squash		
Sweet potato		
Sweet corn		
Sweet cherries		
Tomato		
Tuna		
Watermelon		

V. CONCLUSION & FUTURE WORK

In this study, we proposed a predictive approach using machine learning algorithm to identify CKD and NOTCKD patients, where Random Forest was more accurate than SVM and Naïve Bayes, with an accuracy of 99.75%. Based on the obtained results we recommended food for different level of CKD patients using blood potassium level which will help patients to maintain their salt level.

We wish to work with larger dataset in the future which can contribute in the health care sector. Daily meals for patients suffering from CKD can also be recommended.

REFERENCES

[1] GBD 2013 Mortality and Causes of Death Collaborators, "Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013," Lancet, vol. 385 (9963), p. 117–171.

[2] J.C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, W. E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse", Proc AMIA Annual Fall Symposium, pp. 101105, 1997.

[3] Picture available on: - https://nephrologists.nephroconferences.com/

[4] Ministry of Health, Nutrition & Indigenous Medicine, Sri Lanka, "Dietary Guidelines & Nutrition Therapy For Specific Diseases", health.gov.lk[Online].Available: http://www.health.gov.lk/enWeb/publicpubli/Dietaryguidlines.pdf

[5] Uma N. Dulhare and Mohammad Ayesha, "Extraction of Action Rules for Chronic Kidney Disease using Naïve Bayes Classifier," *InternationalConference on Computational Intelligence and Computing Research (ICCIC),*

[6] A. Salekin and J. Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," pp. 262–270, IEEE, Oct. 2016.

[7] J. H. Kim, J. H. Lee, J. S. Park, Y. H. Lee, and K. W. Rim, "Design of Diet Recommendation System for Healthcare Service Based on User Information," Proc. 4th Int'l Conf Computer Sciences and Convergence Information Technology, pp. 516-518, November 2009.

[8] N. Chetty, K. S. Vaisla, and S. D. Sudarsan, "Role of attributes selection in classification of Chronic Kidney Disease patients," in *Computing,Communication and Security (ICCCS)*, 2015 *International Conference* on, pp. 1–6, IEEE, 2015.

[9] Dr. S. Vijayarani and Mr. S. Dhayanand, "KIDNEY DISEASE PREDICTION USING SVM AND ANNALGORITHMS,"*International Journal ofComputing and Business Research (IJCBR),* vol. 6, no. 2, 2015.

[10] S. Ahmed, M. T. Kabir and N. T. Mahmood, " Diagnosis of kidney disease using fuzzy expert system.," in *2014 8thInternationalConference onSoftware, Knowledge, Information Management and Applications (SKIMA)*, Dhaka, 2014, December.

[11] M. Phanich, P. Pholkul, and S. Phimoltares, "Food Recommendation System Using Clustering Analysis for Diabetic Patients," 2010 Int. Conf. Inf. Sci. Appl., pp. 1–8, 2010.

[12] M. S. Wibawa, I. M. D. Maysanjaya, I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose," in *2017 5th International Conference on Cyber and ITService Management (CITSM).*

[13] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[14] https://data.world/adamhelsinger/food-nutrition-information

[15] "Potassium and Your CKD Diet", The National Kidney Foundation. [Online]. Available: https://www.kidney.org/atoz/content/potassium. [Accessed: 24- Aug - 2017].

[16] http://www.sayedsayed.com/naive_bayesian.html [Accessed: 20th March, 2017].

[17] Fuzzy Lookup in Excel:- https://www.microsoft.com/en-us/download/details.aspx?id=1501