PROGNOSTICATE OF CHRONIC KIDNEY DISEASE WITH MACHINE LEARNING ALGORITHMS BASED ON BLOOD POTASSIUM LEVELS

B. NAGA RAJESWARI, S. PRIYANKA, M. USHA MERCY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE
PADUR , TAMILNADU

Abstract

If Kidney damage and diminished function that more than three months is known as Chronic Kidney Disease (CKD). The purpose of this research is to identify the suggest in what level the patients in kidney disease and identify require diet plan for a CKD patient by applying the classification algorithms on the test result obtained from patients' medical records. The aim of this work is to control the disease using the suitable diet plan and to identify that suitable diet plan using classification algorithms. The suggested work pacts with the recommendation of various diet plans by using predicted potassium zone for CKD patients according to their blood potassium level. The experiment is performed on different algorithms like Multiclass Neural Network and Multiclass Logistic Regression. Some of them are performed here.

Keywords

Blood Potassium Level, Chronic Kidney Disease (CKD), Data Mining, Diet Plan, Machine Learning (ML), Potassium Zone

INTRODUCTION

The process of analysis voluminous data in different points of view in order to achieve patterns or trends that lead to business intelligence is called as data mining [1]. Data mining has an important role in Information Technology because it finds facts from past information of different areas. For example, data mining can be utilized to mining medicinal information as health area generates a lot of data about ailments, pathologies and patients [2]. Data mining can use in medicinal applications to predict medical examinations, medications, surgical processes, and for the detection of

connections between equinoctial data and pathological data [3].

When kidneys cannot perform their functions properly, kidney diseases may occur. The human body may end up building several complications if kidneys are no longer be able to remove extra water and waste products from the blood. Kidney damage and diminished function that lasts longer than three months is known as Chronic Kidney Disease (CKD). Medications and lifestyle changes may help to slow the disease progress if CKD diagnosed early [4]. Therefore, following a proper and suitable diet plan can help to slow the progress of CKD. The diet plan may vary from patient to patient depending on many medical reasons. So it is very important to identify the most suitable diet plan by considering their own conditions.

Dietary management of CKD patients relies upon the current CKD stage. Depending on a CKD patient's estimated Glomerular Filtration Rate (eGFR), there are five stages of CKD as stage 1, stage 2, stage 3, stage 4 and stage 5. Up to the stage 3, patients may not have the symptoms of CKD or they might be able to deal with the renal functions without amassing excretory products like potassium or excess urea in the blood. Hence, patients in stage 1, 2 or 3 may not require a very strict dietary control. But in stage 4 and 5, patients are in difficulty of keeping up the balance of minerals, electrolytes, and liquids inside their body. Therefore, they need to be under a proper dietary control. A special renal eating routine is important at this

late stages to control the disease, prevent further deterioration of the disease and also maintain water balance and electrolytes minerals inside the body.

Dietary management of CKD patients not just depend on the stage of the disease but also with other conditions, like the level of blood potassium, urea, calcium, phosphorous and so on [5]. In this study, the main focus is on blood potassium level to identify the suitable diet plan for a CKD patient.

LITERATURE SURVEY

Many kinds of research studies have been conducted to predict results for CKD related problems using various data mining techniques.

S. Shah, A. Kusiak and B. Dixon [6] performed a research in 2005, to predict the survival of CKD dialysis patients using data mining techniques. To find the relationship between CKD patient survival and the selected attributes, a data mining approach is used in this study. For mining information in the type of decision rules, two dissimilar data mining algorithms are used. Data mining is calculated based on data analysis of the patient received hospitals. The study determined that the overall classification accuracy for all data mining algorithms was expressively greater using the single visit dataset over the other dataset which is aggregate dataset. The prediction accuracy of single visit based rule sets improved over the aggregate based rule sets.

S. Bala and K. Krishan [7] presented a literature survey in 2014, on data mining classification techniques used for CKD predictions. This review analysed the results of classification algorithms used for CKD predictions by many researchers. The overall objective was to study the various data mining techniques available to predict the CKD and to compare them to find the best method of prediction. This study has analysed that there is no individual classification technique which produces the best result for every dataset.

S. Vijayarani, S. Dhayanand [8] presented a research in 2015 to predict kidney related diseases by using Artificial Neural Network (ANN) algorithm and Support Vector Machine (SVM) algorithm. To compare the performance of selected two algorithms based on their execution time and accuracy was the goal of this research work. It is concluded that the performance of ANN algorithm is better than SVM algorithm with the accuracy of 87%, from the experimental results.

Lambodar Jena and Narendra Ku. Kamila [9] presented a research in 2015, for prediction of CKD using Naive Bayes, Multilayer Perceptron, Support Vector Machine, J48, Conjunctive Rule and Decision Table. The performance of these techniques is measured by classification accuracy, the time acquired for build a model, the time acquired to test a model, Kappa statistics, mean absolute error, and ROC Area. It is observed from experimental results, the Multilayer Perceptron algorithm gives better result than the other five algorithms with the classification accuracy of 99.7%.

V.Kunwar, K. Chandel, A. Sabitha and A. Bansal [10] compared Artificial Neural Network (ANN) and Naive Bayes data mining classification algorithms to predict and diagnose CKD in 2016. The experimental results showed that Naive Bayes classification technique is the most accurate technique with 100% accuracy when compared to ANN classification technique which produced 72.73% accuracy.

M.A. Ameta and M.K. Jain [11] presented a review of data mining algorithms for the prediction of CKD and treatments of CKD like dialysis in 2017. The research evident that classification is the data mining technique which is highly efficient in prediction for CKD. Also, the study showed that various methods for attribute selection can further improve the classification results.

PROBLEM STATEMENT

Most of the researchers worked on data mining algorithms in different kidney disease survey techniques. In these survey techniques they are using different sources but they are unable to analysis data with visualization techniques to identify the correlation between different attributes. Using limited techniques and they are unable to optimize by increasing efficiency. And they diagnosis disease based on difficult classification only.

Kidney disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency and also reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health. The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in users and patients. The purpose of this project is to prepare predictive modeling for kidney disease data to analyze with different python open source modules and produce prediction outputs with machine learning algorithms and find out accuracy by comparing different algorithms.

Kidney damage and diminished function that lasts longer than three months is known as Chronic Kidney Disease (CKD). The primary goal of this research study is to identify the suitable diet plan for a CKD patient by applying the classification algorithms on the test result obtained from patients' medical records. The aim of this work is to control the disease using the suitable diet plan and to identify that suitable diet plan using classification algorithms. The suggested work pacts with the recommendation of various diet plans by using predicted potassium zone for CKD patients according to their blood potassium level.

PROPOSED SYSTEM

In this proposed system we are collecting data from different data sources and using jupyter notebook to visualize data with help of python and do analysis on the data with different techniques of visualization and implementing machine learning algorithms to identify the kidney disease problem and category they belongs to and find out exact accuracy and make the performance high. And we predicting the disease by blood potassium level and suggest them suitable dietary plan.

The purpose of this project is to prepare predictive modelling for kidney disease data to analyse with different python open source modules and produce prediction outputs with machine learning algorithms and find out accuracy by comparing different algorithms.

The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of diabetes disease

requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The implementation of work is done on diabetes data (UCI) machine learning repository to test on different data mining techniques with Machine learning algorithms.

The main aspects in our proposed system are the potassium levels. The below is level of blood potassium which are used to detect the kidney disease.

THE FORMULA

=if[pot]>0 and [pot]<3.5 then “LOW”
else if [pot]>=3.5 and [pot]<=5.0 then “SAFE”
else if [pot]>=5.1 and [pot]<=6.0 then “CAUTION”
else if [pot]>=6.1 then “DANGER”
else “NO DATA”

METHODOLOGY

CRISP-DM has used as the methodology in this study. The meaning of CRISP-DM is Cross-Industry Process for Data Mining. This methodology gives an organized way to deal with arranging a data mining development. CRISP-DM defines as a well-proven and robust methodology for data mining.

There are six steps in CRISP-DM as,

A. Business Understanding
B. Data Understanding
C. Data Preparation
D. Modeling
E. Evaluation
F. Deployment

A. Business Understanding

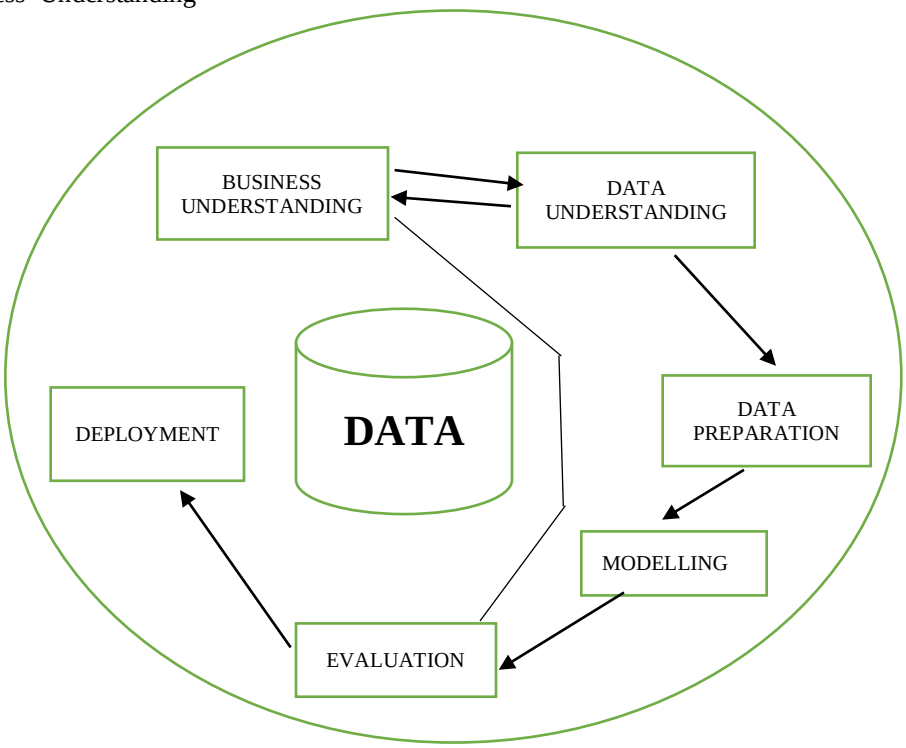

Figure1. CRISP-DM(Cross-Industry Process for Data Mining)

This research has gone through these steps according to the CRISP-DM methodology which is shown in Fig. 1.

This step concentrates on comprehension of the project requirements and objectives from a commercial point of view, and after that changing over this understanding into a data mining problem explanation and a preparatory arrangement [12].

The objective of this study is predicting the most suitable diet plan using blood potassium level for CKD patients to slow their progress of CKD. The main motivation in here is to identify a data mining procedure for a more accurate prediction on variations of blood potassium level of both CKD and none-CKD patients.

B. Data Understanding

This step begins with an underlying data gathering and continues with actions to facilitate acquainted with the information, to distinguish data value problems, to find initial bits of knowledge into the data, or to identify useful subsections to frame theories for unseen data [12].

The dataset for predicting the diet plan for CKD patients is obtained from the UCI data repository [13]. As this dataset is a publicly available dataset, it does not require an ethics approval to use for research works.

Attribute	Description	Allowed Value
age	age in years	Numerical Values
bp	blood pressure (mm/Hg)	Numerical Values
sg	specific gravity	Nominal Values (1.005, 1.010, 1.015, 1.020, 1.025)
al	albumin	Nominal Values (0, 1, 2, 3, 4, 5)
su	sugar	Nominal Values (0, 1, 2, 3, 4, 5)
rbc	red blood cells	Nominal Values (normal, abnormal)
pc	pus cell	Nominal Values (normal, abnormal)
pcc	pus cell clumps	Nominal Values (present, not present)
ba	bacteria	Nominal Values (present, not present)
bgr	blood glucose random(mgs/dl)	Numerical Values
bu	blood urea (mgs/dl)	Numerical Values
sc	serum creatinine (mgs/dl)	Numerical Values
sod	Sodium (mEq/L)	Numerical Values
pot	Potassium (mEq/L)	Numerical Values
hemo	Hemoglobin (gms)	Numerical Values
pcv	packed cell volume	Numerical Values

wc	white blood cell	count (cells/cumm) Numerical Values
rc	red blood cell count (millions/cmm)	Numerical Values
htn	hypertension	Nominal Values (yes, no)
dm	diabetes mellitus	Nominal Values (yes, no)
cad	coronary artery disease	Nominal Values (yes, no)
appet	appetite	Nominal Values (good, poor)
pe	pedal edema	Nominal Values (yes, no)
ane	anemia	Nominal Values (yes, no)
class	class	Nominal Values (ckd, notckd)

Table 1. 25 Attributes

This dataset contains data of people from the southern part of India with their ages ranging between 2-90 years. There are 400 instances with 25 different attributes related to CKD. The attribute data view of records is shown in Table I.

C. Data Preparation

This step covers all actions to develop the final dataset from the original raw dataset [12]. The main concentration of this project is to predict Chronic Kidney Disease with help of the value of blood potassium level, they are categorized into different as safe zone, caution zone and danger zone.

- If the blood potassium level is in between 3.5 - 5.0 the patient is in the SAFE zone.
- If the blood potassium level is in between 5.1 - 6.0 the patient is in the CAUTION zone.
- If the blood potassium level is higher than 6.1 the patient is in the DANGER zone [14].

To get a more accurate solution, the current zone of a patient will predict so the most suitable diet plan for that patient can be identified using the predicted result.

When preparing the final dataset, a new attribute named “zone class” has added to the dataset. The purpose of having this attribute in the dataset is to predict the value of this attribute by considering the values of other attributes.

D. Modeling

In the modeling step, the purpose is to select the modeling technique that will be using. Since multiple techniques are applied in here to find the highest accurate technique, this task was performed separately for each technique.

The known problem in this study was to identify the variations of zones by considering the blood potassium level of CKD patients. The prediction has been conducted targeting the zone class to identify the suitable diet plan using the zone class. Since the zone class attribute is predicting five categories, classification algorithms were used. Multiclass Logistic Regression and Multiclass Neural Network algorithms were used in this research to identify the most suitable algorithm with selected attributes in the dataset.

E. Evaluation

Once the model has been built, that appear to have a high quality based on whichever loss functions have been selected, these need to be tested to ensure they generalize against unseen data and that all key business issues have been sufficiently considered. The end result is the selection of the best algorithm for the model [12].

From original dataset, following 21 attributes have been filtered to use in the model.

Filtered attributes: age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, appetite, anemia, class, zone class

F. Deployment

This step will deploy a code representation of the model into an operating system to score or categorize new unseen data as it arises and to create a mechanism for the use of that new information in the solution of the original business problem [12].

In this research, the model was developed in Microsoft Azure Machine Learning Studio. After the deployment of the model, responsible people can use the predicting model to

deals with the suggestion of different diet plans for CKD patients according to their blood potassium level by using the predicted potassium zone.

EVALUATION AND RESULTS

For training and testing of the dataset, 70% of records are selected to train the model and the other 30% of records are selected to test the trained model. The zone class attribute was chosen as the selected column to train the model.

As per the results we obtained, it evaluates Multiclass Logistic Regression algorithm gives the overall accuracy as 0.8917 which is 89.17% and Multiclass Neural Network algorithm gives the overall accuracy as 0.8250 which is 82.50%.

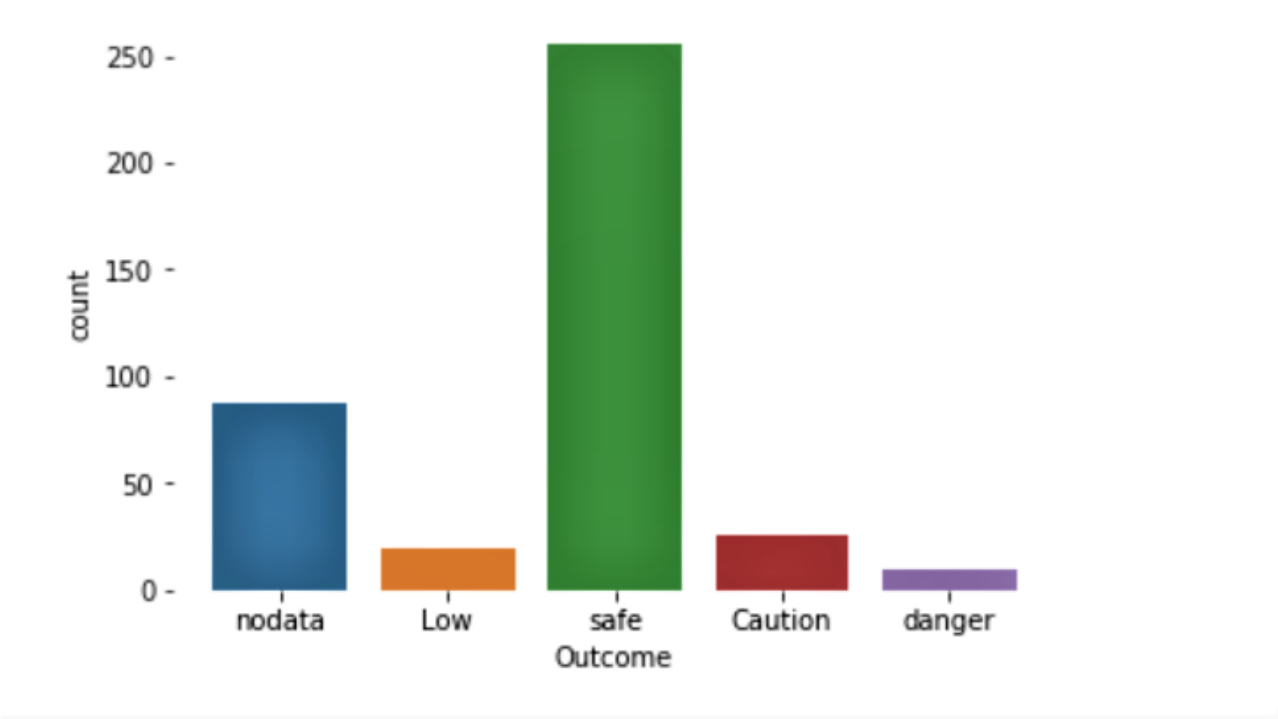

Figure 2. Classification Of patients into their zones

1. The data is plotted according to taken data set. In this we can see that two axes i.e., x and y axes respectively.

   Y-axis - Count (Number of people)

   X-axis - Outcome(This gives which region the Patient belongs to such as no data, low , safe , caution , danger)

The above bar plot shows that most are in safe zone. And some them are in danger zone which means that they should take immediate operation as their condition is serious.

2. In the result part we are comparing some of the attributes and classifying them into different zones as mentioned above. Firstly, we are comparing with the attribute “anemia” which deals with patients classification for respective zones based on the “anemia” levels Count indicates the number of people have taken. And the zones are classified into no data, low, safe , caution , danger.

Here,

X-axis - Outcome(No data , Safe , Caution , Danger)

Y-axis - Count (Number of people)

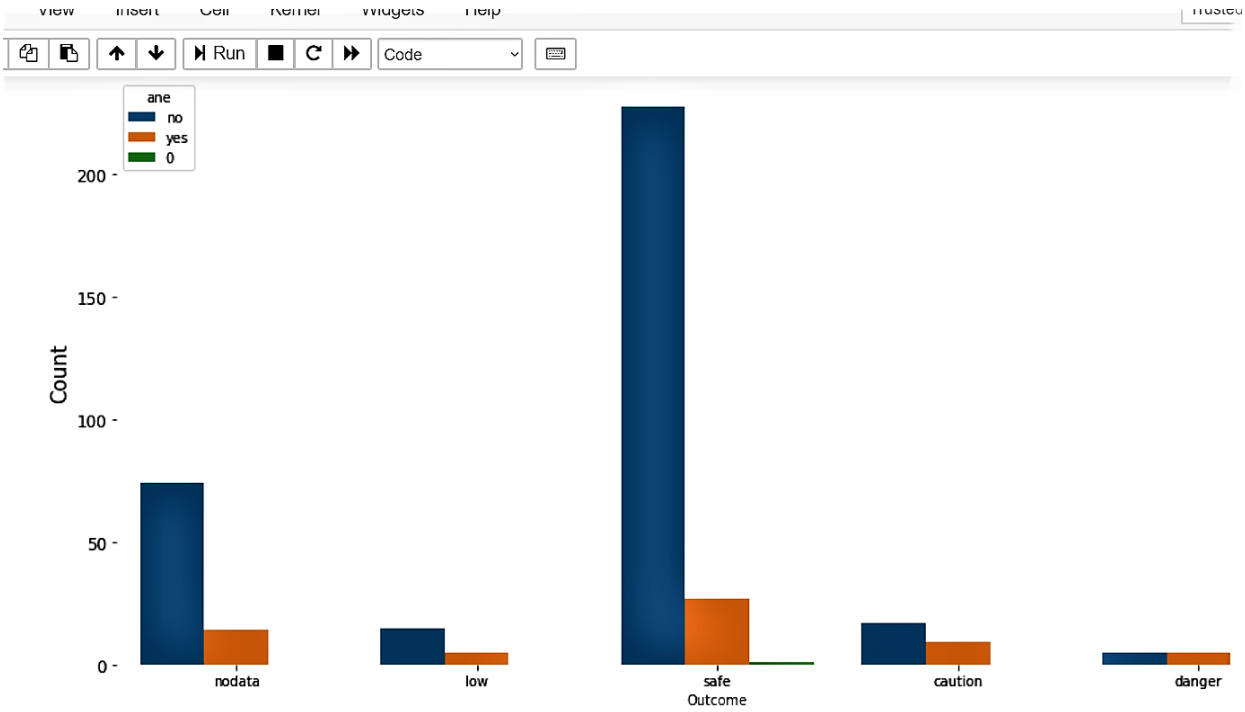

Figure3.Comparision with anemia

3. In this we have compared with bacteria levels and noted as present , not preset and 0 values. Here,

Y-axis – Count(Number of people)

X-axis – Outcome(No data, Safe , Danger and caution)

This graph gives the information about the bacteria levels present or not in the patient body.

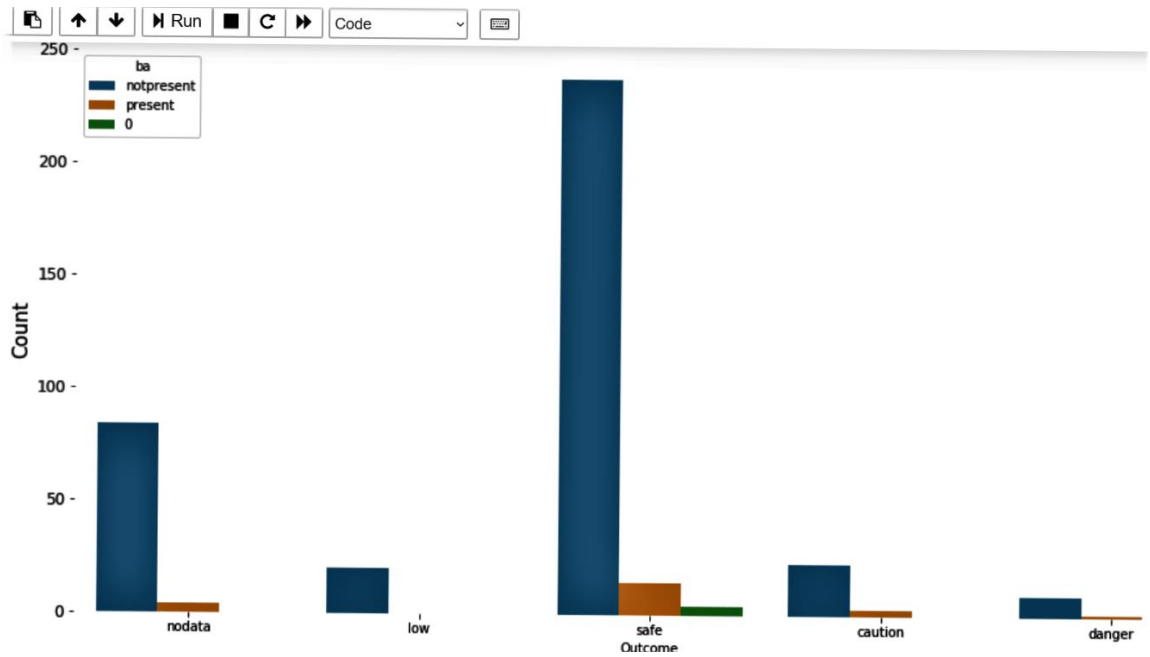

Figure4.Comparision with bacteria

4.

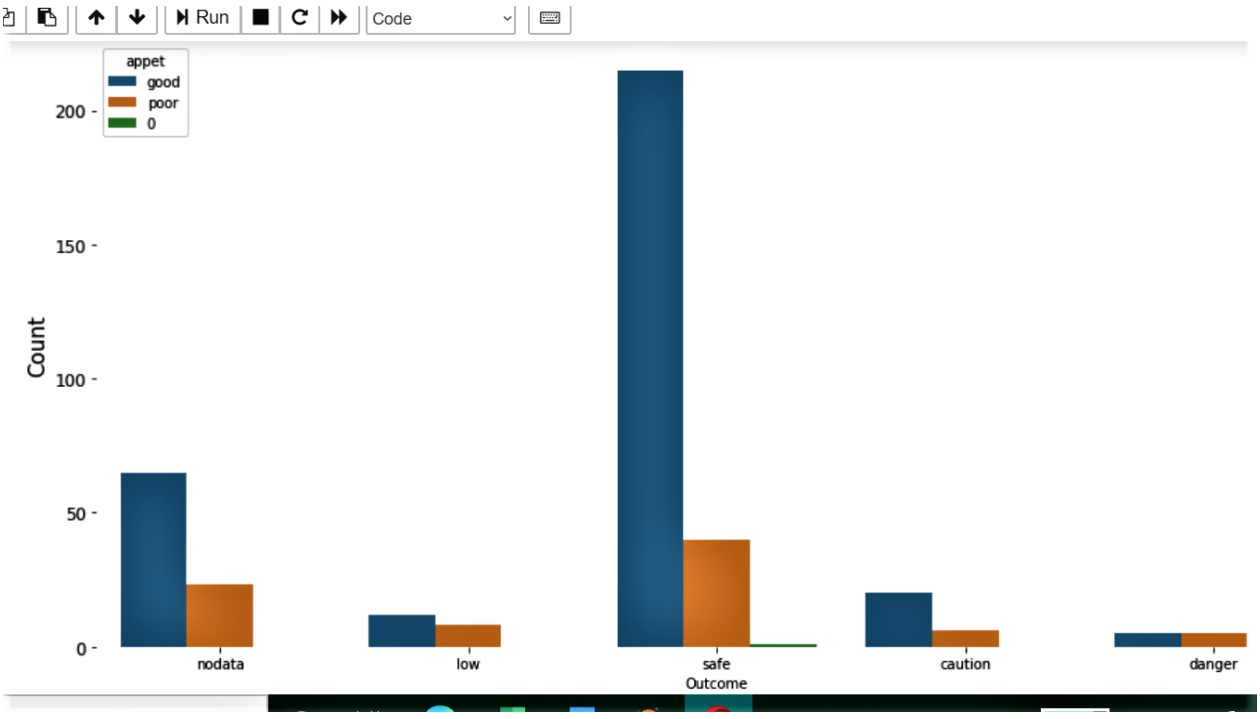

Figure 5.Comparision with appetite

In this we have compared with appetite levels and noted as good, poor and 0 values. Here,

Y-axis – Count(Number of people)

X-axis – Outcome(No data, Safe , Danger and caution)

This graph gives the information about the appetite levels good or poor in the patient body.

5.

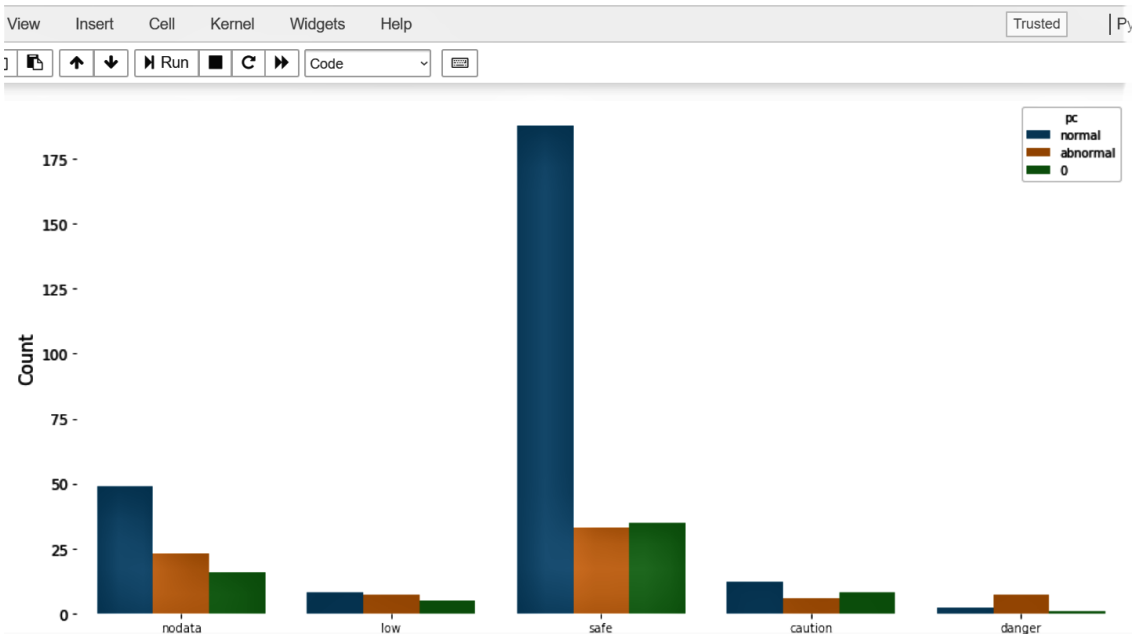

Figure 6.Comparision with pus cell

In this we have compared with pus cell levels and noted as normal, abnormal and 0 values. Here,

Y-axis – Count(Number of people)

X-axis – Outcome(No data, Safe , Danger and caution)

This graph gives the information about the appetite levels good or poor in the patient body.

6.

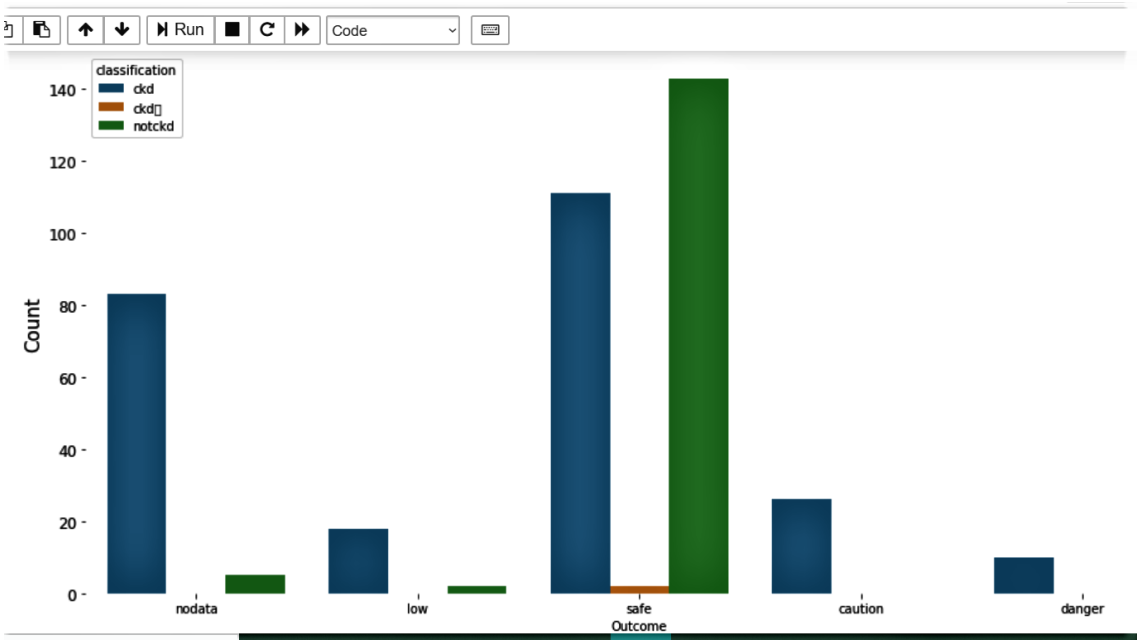

Figure 7. Classification of CKD, Not CKD

In the above graph we have classified into CKD, Not CKD. Here,

X-axis- Count(Number of People)

Y-axis-Outcome(No data, Low, Safe, Caution, Danger)

7.

age	41	6	55	41	44	53	61	18	45	46	...	45	29	50	36	43	48	35	9	12	51
bp	4	1	4	3	4	5	3	0	6	5	...	4	4	4	2	4	4	3	4	2	4
sg	4	4	2	1	2	3	2	3	3	4	...	5	5	4	5	4	4	5	4	5	5
al	1	4	2	4	2	3	0	2	3	2	...	0	0	0	0	0	0	0	0	0	0
su	0	0	3	0	0	0	0	4	0	0	...	0	0	0	0	0	0	0	0	0	0
rbc	0	0	2	2	2	0	0	2	2	1	...	2	2	2	2	2	2	2	2	2	2
pc	2	2	2	1	2	0	2	1	1	1	...	2	2	2	2	2	2	2	2	2	2
pcc	1	1	1	2	1	1	1	1	2	2	...	1	1	1	1	1	1	1	1	1	1
bgr	49	0	141	45	34	3	28	139	63	2	...	27	13	60	45	62	65	4	28	42	58
bu	24	6	41	44	14	13	42	19	47	80	...	13	4	35	32	33	36	19	14	37	6
sc	9	5	15	33	11	8	81	8	16	52	...	5	8	9	4	5	2	9	3	7	8
sod	0	0	0	3	0	27	2	0	0	5	...	20	27	32	26	24	33	26	22	20	26
pot	0	0	0	1	0	6	14	0	0	11	...	11	15	17	18	24	23	9	18	23	9
hemo	91	50	33	49	53	59	61	61	45	32	...	87	93	85	67	78	94	102	95	79	95
pcv	33	27	20	21	24	28	25	33	22	18	...	41	33	35	43	34	36	43	38	40	42
wc	73	57	71	63	69	73	0	65	89	19	...	59	55	62	70	88	63	73	62	68	64
rc	35	0	0	20	28	26	0	32	22	18	...	36	46	38	37	28	31	45	37	42	44
appet	1	1	2	2	1	1	1	1	1	2	...	1	1	1	1	1	1	1	1	1	1

Figure 8. Encryption of 25 attributes

In this the data has been encrypted based on 25 attributes. For further prediction.

8.

Here, We used logistic and KNN algorithm and the results are according to that.

```
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.84	0.80	0.82	64
1	0.67	0.72	0.69	36
micro avg	0.77	0.77	0.77	100
macro avg	0.75	0.76	0.75	100
weighted avg	0.78	0.77	0.77	100

Figure 9. Results of taken data

9.

In this, we are showing the resultant date in visual form we got results as…

For KNN algorithms and Logistic Algorithms with accuracy of 89.17% and 85% .

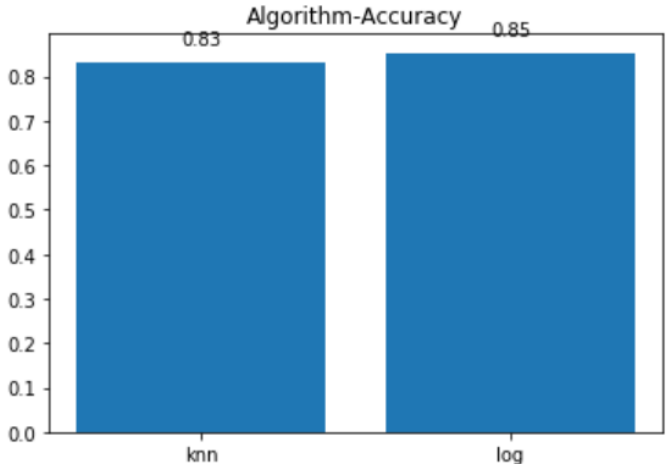

Figure 10. Visualization OF DATA

CONCLUSION

Prediction of diet plans for CKD patients is one of the essential topics in the medical area. The proposed study is to identify the different diet plans by predicting potassium zone of CKD patients according to the blood potassium level. But In results we have shown only on Logistic Regression and KNN algorithms. The classification algorithms that have been considered for predicting potassium zone are Multiclass Logistic Regression, and Multiclass Neural Network.

FUTURE WORK

There are so many possible ways to get the best results. In this project Multiclass Logistic Regression, and Multiclass Neural network are used to predict the Chronic Kidney Disease suitable using blood potassium level using potassium levels with help of machine learning algorithms. It helps to suggest suitable diet plan helps in the timely treatment of the patients suffering from CKD and also help patients to over come from the disease before it getting worse. As prevention better than curing. The new researchers can be used and their performance can be evaluated to find better solutions of the objective function in future work.

ACKNOWLEDGMENT

We would gratefully acknowledge parents and the other lecturers of Hindustan Institute of Technology and sciences for providing their valuable guidance and time.

We have taken efforts in this project. However, it would not have been possible without the kind support of Dr Dinah Punnoose and Dr Padmaveni K for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. I would like to thank my friends and my team members for being a constant source of encouragement in all my endeavors. Indeed it was your support that saw us through the many ups and downs of life.

REFERENCES

[1] A. Pujari, Data mining techniques. Hyderabad: Universities Press (India) Private Limited, 2013.

[2] M.A. Khaleel, S.K. Pradham, G.N. Dash, "A survey of data mining techniques on medical data for finding locally frequent diseases", Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 8, pp. 149-153, August 2013.

[3] 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.

[4] "Essential Guide to Kidney Disease", Western Hospital, 2017. [Online]. Available: http://www.westernhospital.lk/essential-guide-tokidney-disease. [Accessed: 22- Aug- 2017].

[5] 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 [Accessed: 22- Aug- 2017].

[6] S. Shah, A. Kusiak, B. Dixon, "Data Mining in Predicting Survival of Kidney Dialysis Patients", in Proceedings of Photonics West-Bios 2003, vol. 4949, pp. 1-8, 2003.

[7] S. Bala and K. Krishan, "A literature review on kidney disease prediction using data mining classification technique." International Journal of Computer Science and Mobile Computing 3.7, pp. 960-967, 2014.

[8] S. Vijayarani, S. Dhayanand, "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS", International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2015.

[9] Lambodar Jena, Narendra Ku. Kamila “Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney-Disease”, International Journal of Emerging Research in Management & Technology, ISSN: 2278-9359 Vol.4, Issue11, November 2015.

[10] V. Kunwar, K. Chandel, A. Sabitha and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016.

[11] M.A. Ameta and M.K. Jain, "Data Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review."

[12] Data Science Central, “CRISP-DM – a Standard Methodology to Ensure a Good Outcome”, datasciencecentral.com, 2016 [Online]. Available: http://www.datasciencecentral.com/profiles/blogs/crispdm-a-standard-methodology-to-ensure-a-good-outcome [Accessed: 23- Aug- 2017].

[13] "UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set", Archive.ics.uci.edu, 2015. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. [Accessed: 24- Aug - 2017].

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