The Journal of Social Sciences Research

ISSN(e): 2411-9458, ISSN(p): 2413-6670

Vol. 5, Issue. 2, pp: 219-232, 2019

URL: https://arpgweb.com/journal/journal/7

DOI: https://doi.org/10.32861/jssr.52.219.232

Academic Research Publishing Group

Original Research Open Access

Degree of Urgency and Progression Predictive Model for Dialysis using Hybrid System

August Anthony N. Balute*
School of Graduate Studies, AMA University, Project 8, Quezon City, 1116, Philippines

Melvin A. Ballera
School of Graduate Studies, AMA University, Project 8, Quezon City, 1116, Philippines

Shaneth C. Ambat
School of Graduate Studies, AMA University, Project 8, Quezon City, 1116, Philippines

Menchita F. Dumlao
School of Graduate Studies, AMA University, Project 8, Quezon City, 1116, Philippines

Dennis B. Gonzales
School of Graduate Studies, AMA University, Project 8, Quezon City, 1116, Philippines

Abstract

Machine-learning and data mining techniques using hybrid system were used to accurately predict the development of diseases such as Chronic Kidney Disease (CKD) and Acute Renal Failure (ARF). In this study, Random Forests Decision Algorithm, Autoregressive Integrated Moving Average (ARIMA) Model and K-means Clustering Algorithm were used to predict the degree of urgency and progression of dialysis from patients' electronic medical records. The use of such algorithms will provide a predictive model for forecasting the urgency level and CKD stages, clustering by gender, age, CKD stages and urgency level to anticipate adverse events that will help medical practitioners in the efficiency and accuracy of detecting the severity of the kidney disease. 20,000 instances were divided into training and testing data, wherein the data was able to label the urgency and progression of dialysis. Apart from this, the stages of CKD and urgency level were forecasted using ARIMA Model. The extracted pattern from the historical and current data predicted the urgency and progression of dialysis, thus a prototype software implementation was also proposed. The experimental results of the study show that 99 percent (%) of the prediction on the degree of urgency and progression of dialysis model deemed accurate, paving way to a better clinical decision-making process of nephrologists using a rule-based system from the important attributes of the patient's electronic medical records which will also help improve a patient's quality of life.

Keywords: Acute renal failure (ARF); Autoregressive integrated moving average (ARIMA)model; Chronic kidney disease (CKD); Cross-industry process for data mining (CRISP-DM); Electronic medical record (EMR); Knowledge discovery in databases (KDD); K-means clustering algorithm (KMCA); Orange data mining; Random forests decision algorithm (RFDA); R programming; Waikato environment for knowledge analysis (WEKA).

1. Introduction

Dialysis is a life-saving procedure for patients with kidney failure as it involves removing waste products and excess fluid from the body. The need for a dialysis treatment is usually for patients with end stage renal disease who developed acute illness and acute kidney injury from an acute illness or a procedure. The decision to start a dialysis treatment is based on the patient's signs and symptoms, although the optimal time to start a dialysis is still a topic of debate in both acute and chronic kidney failure. Kidney failure may initially be asymptomatic and may present differently from person to person, causing patients to only seek consult when the disease already developed signs of uremia requiring urgent dialysis (Scorecki, 2016). According to the Hemodialysis Adequacy Work Group, there is a lack of clear cut guidelines on when there is a need to start a dialysis, resulting to an early or late dialysis. The decision to start a dialysis and when it is done is usually based on the judgment of the attending physician affected by the factors: 1) patients may refuse to give consent for dialysis until symptoms of uremia appear and 2) patients tend to wait until they develop significant symptoms of uremia and fluid overload or malnutrition, a call to become forced to accept and commence with dialysis. In the Philippines, the number of chronic kidney disease has been steadily increasing with 23,000 patients undergoing dialysis treatment as of 2013 not including those suffering from kidney failure who were not able to get treatment. The annual mortality rate per 100,000 people from chronic kidney disease in the Philippines has increased by 16.2% since 1990 which is an average of 0.7% per year (Renal Disease Control Program). With the increase in number of patients needing dialysis and the resource limited setting of our health care system, it is often encountered that there are a lot of patients needing dialysis and there are only a few dialysis machines available. Since patients lined up for emergency dialysis are usually seen by different physicians,

*Corresponding Author

determining who to prioritize first once a slot is available becomes a challenge. The managers/supervisors at hemodialysis units are also unaware of the overall clinical status of the patient and only rely on the note of the attending physician. With the multitude of factors that needs to be considered in clinical decision making regarding dialysis, the use of machine learning algorithms applied to patient data sets may help clinicians and dialysis units become more efficient and increase cost effectiveness of treatments and quality of care.

1.1. Statement of the Problem

This study aims to create a prediction model to determine the degree of urgency and progression of dialysis using Random Forests Decision Algorithm, ARIMA Model and K-means Clustering Algorithm.

1. What are the important attributes that can help in preventing or recommending urgency and progression of dialysis?
2. How the Random Forests Decision Algorithm can help predict the degree of urgency of dialysis?
3. How the ARIMA Model can help forecast the progression of dialysis?
4. How the K- means Clustering Algorithm can cluster the urgency rate and progression level of dialysis?
5. How Random Forests Decision Algorithm, ARIMA Model and K-means Clustering Algorithm provide relevant information of the dialysis patient pertaining to patient medication and dialysis?

2. Review of Related Literature

The prompt treatment for a dialysis has several indications helping in predicting the urgency and progression of dialysis complications (Spinner, 2014). The AEIOU indications will help preserve long-term renal function:

A – Acidosis – metabolic acidosis with a pH <7.1

E – Electrolytes – refractory hyperkalemia with a serum potassium >6.5 mEq/L or rapidly rising potassium levels

I – Intoxications – use the mnemonic SLIME to remember the drugs and toxins that can be removed with dialysis: salicylates, lithium, isopropanol, methanol, ethylene glycol

O – Overload – volume overload refractory to diuresis

U – Uremia – elevated BUN with signs or symptoms of uremia, including pericarditis, neuropathy, uremic bleeding, or an otherwise unexplained decline in mental status (uremic encephalopathy)

2.1. The Use of Random forest Decision Algorithm

Random Forests Decision Algorithm is a learning algorithm that combines several randomized decision trees and aggregates the corresponding predictions by averaging (Scornet *et al.*, 2015). Its ability of generalization is higher than other multi-class classifiers due to the effect of bagging and feature selection (Louppe, 2014). As data analysis and machine learning is now an integral part of modern scientific methodology, the complexity of random forests shows good computation performance and scalability. The study of Louppe shows that induction of decision trees and construction of ensembles of randomized trees motivates design and purpose (Mishina *et al.*, 2015). In consequence of the same work, the result analysis variable importance' as computed from non-totally randomized trees (Random Forests) suffers from a combination of defects due to masking effects. This misestimates the node impurity due to the binary structure of decision trees (Sugiyatno *et al.*, 2015).

According to Fathi & Majari, a robust random forest regression model with good prediction outcomes can be employed for metabolic profiling to find out which metabolites in the serum have significance in the diagnosis of Crohn's disease (CD), a form of inflammatory bowel disease that may affect a part of the gastrointestinal tract, as it is difficult to diagnose using clinical tests. CD and healthy subjects were correctly classified using random forest methodology having a 94% correct classification. Moreover, good prediction outcomes were developed for correlating serum zinc level and metabolite concentrations. The regression model showed the correlation *(R(2))* and root mean square error values of 0.83 and 6.44, respectively. This model suggests valuable clues for understanding the mechanism of zinc deficiency in CD patients (Fathi *et al.*, 2014).

On the use of electronic health records (EHR), there is challenge to translate eligibility criteria from free text into decision rules that are compatible with the data from the EHR. The possibility of predictive modeling to assess the eligibility of patients for clinical trials and report prototype's performance for different system configurations were evaluated in the study of Köpcke *et al.* (2013). The prototype worked by using existing patient data of manually assessed eligible and ineligible patients to induce prediction models. Performance was measured for three (3) clinical trials by plotting receiver operating characteristic curves and comparing the area under the curve *(ROC-AUC)* for different prediction algorithms, sizes of learning set, different number and aggregation levels of patient's attributes. It was found that Random Forests were generally the best performing model. The full potential of this algorithm indicated that predictive modeling is a feasible approach to support patient clinical trials, with a major advantage over the commonly applied rule-based systems.

2.2. The Use of Autoregressive Integrated Moving Average Model (ARIMA)

Time series data in business, economics, environment, medicine, and other scientific fields tend to exhibit patters such as trends, seasonal fluctuations, irregular cycles, and occasional shifts in level or variability (Tiao, 2015). According to Anguera *et al.* (2016), Jeewantha *et al.* (2017), the analysis of time series for knowledge discovery requires application of special-purpose tools such as the key information of interest to the expert concentrated within a particular time series region, known as events. The concept of data mining to analyze volumes of stored medical data to discover knowledge has a huge potential.

On the other hand, classification of multivariate time series data is considered challenging but necessary for medical care and research. According to Moskovitch & Shahar in their study on the classification-driven temporal discretization of multivariate time series, a series of raw-data time points can be abstracted into a set of time intervals which can be used for the classification of multivariate time series (Moskovitch and Shahar, 2015). Ordon *et al.* (2014) also stated that ARIMA models are useful to assess trends over time, like evaluating a population based on trends in the use of medical treatments (extracorporeal shock wave lithotripsy, ureteroscopy and percutaneous nephrolithotomy) as well as to assess the re-treatment rate and morbidity from treatment over time.

In the study of Tao & XingYu, ARIMA model was utilized together with Kalman Filter Algorithm to forecast the incidence of gonorrhea (Tao *et al.*, 2013). The results show that Kalman Filter Algorithm based on ARIMA model could perform the prediction of the disease more precisely and accurately, comparing the range of absolute error (AE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). Furthermore, ARIMA also accurately predicts the time of death, institutionalization, and need for full-time care of Alzheimer's disease patients serving as a clinical, research, and public health need (Akram *et al.*, 2018; Razlighi *et al.*, 2014).

2.3. The Use of K-Means Clustering Algorithm

As clustering and classification are two important techniques of data mining, categorization can also be provided when it comes to categorization of healthcare patients. First, K-means clustering will identify and eliminate incorrectly classified instances then a fine-tuned classification is done using Naive Bayes by taking the correctly clustered instances of the first stage. Experimental results signify the cascaded K-means clustering and Naive Bayes has enhanced classification accuracy. The results of the experiment show that integration of clustering and classification gives promising results with utmost accuracy rate even when the dataset contains missing values (Pandeeswari and Rajeswari, 2015).

Several medical studies ventured on using K-means algorithm to aid clustering methods. Prompt detection and treatment of exacerbations of Chronic Obstructive Pulmonary Disease (COPD) may improve outcome and early detection of the disease as it is a critical issue. A mobile health system could enable the early detection of COPD on a day-to-day basis using K-means clustering algorithm. The algorithm was trained and validated and its accuracy in detecting acute exacerbation was assessed. Sensitivity and specificity were 74.6% and 89.7% respectively, and area under the receiver operating characteristics curve was 0.84. 31out of 33 acute exacerbation of respiratory symptoms were early identified with an average of 4.5 to 2.1 days prior to the onset of the exacerbation that was considered the day of medical attendance. The applied methodology using K-means could help early detect COPD exacerbations on a day-to-day basis and therefore could provide support to patients and physicians (Sanchez-Morillo *et al.*, 2015).

K-means clustering algorithm is highly sensitive to the initial placement of the cluster centers (Celebi *et al.*, 2013). Further, the dynamic nature of healthcare data makes this domain challenging to select adequate data mining algorithm for optimal results (Mahoto *et al.*, 2014). In electrocardiogram (ECG), the most common pieces of information that can be read from the ECG is the heart rate (HR) through the detection of its most prominent feature: the QRS complex. K-means algorithm classifies data into QRS and non-QRS. Results show that the algorithm has a low computational load, with no decision thresholds and does not require any additional parameter. Sensitivity, positive prediction and accuracy from results are over 99.7% (Merino *et al.*, 2015); (Aghamohseni and Ramezanian, 2015).

Hybrid of K-means algorithm and support vector machine algorithms were also used in breast cancer diagnosis where K-means particularly utilized to recognize hidden patterns of the benign and malignant tumors separately. Then, a support vector machine (SVM) was used to obtain the new classifier to differentiate the incoming tours. This proposed methodology improved the accuracy to 97.38% illustrating the capability of the proposed approach on breast cancer diagnosis, times savings during the training phase, and mined abstract tumor features by better understanding the properties of different types of tumors (Ha and Tsai, 2015; Su *et al.*, 2016; Sugiyatno *et al.*, 2015; Umer, 2017; Zheng *et al.*, 2014).

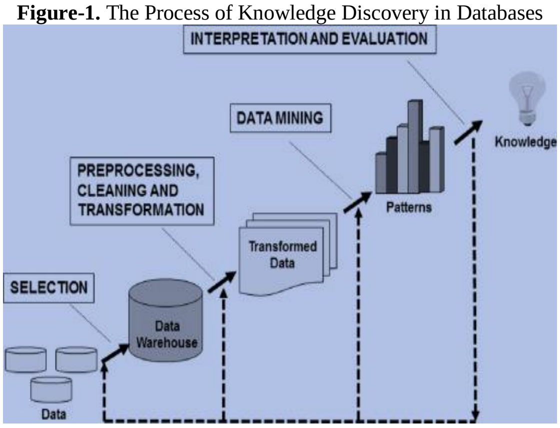
3. Research Methodology

This chapter describes the research approach and methodology of the study. Specifically, it presents the design of the study, methods and techniques, the respondents of the study, the instrument of the study, the developmental model, data processing and statistical treatment that was applied in the study.

3.1. Knowledge Discovery of Databases

The data mining of Knowledge Discovery of Databases (KDD) process aims at the discovery of useful information from large collections of data. The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data. In knowledge discovery cycle, the training data were cleansed, normalized and formatted in the pre-processing stage where the raw data were analyzed by sets of classifiers using a data mining tool.

Figure-1. The Process of Knowledge Discovery in Databases


INTERPRETATION AND EVALUATION
DATA MINING
PREPROCESSING, CLEANING AND TRANSFORMATION
SELECTION
Data
Data Warehouse
Transformed Data
Patterns
Knowledge

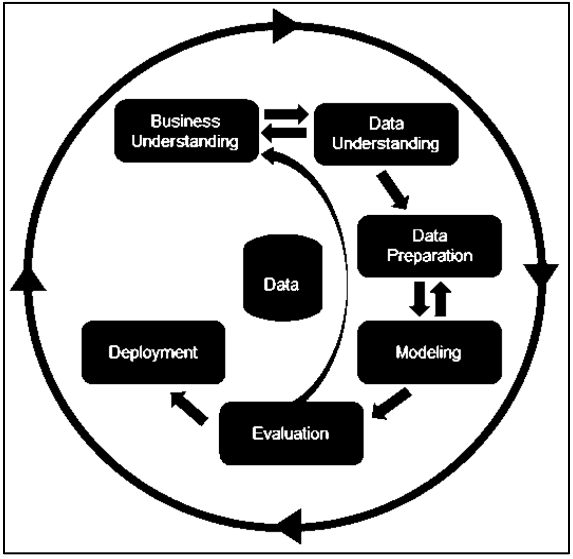
3.2. Development Methodology

The researcher utilized the steps of Knowledge Discovery in Databases and CRISP-DM methodologies for this study.

The data mining of Knowledge Discovery of Databases (KDD) process aims to discover useful information from a large collection of data. The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data. In the knowledge discovery cycle, the training data will be cleansed, normalized and formatted in the pre-processing stage where the raw data will be analyzed by sets of classifiers using a data mining tool. According to Sankar and Mitra (2004), the process model for data mining provides an overview of the life cycle of a data mining project.

Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth, describes CRISP-DM (Cross-Industry Standard Process for Data Mining) as a tool and application-neutral model to recognize better and faster results from data mining. It is a well-proven structured methodology in planning a data mining project (Chapman *et al.*, 2000).

Figure-2. Phases of Crisp-Dm Process

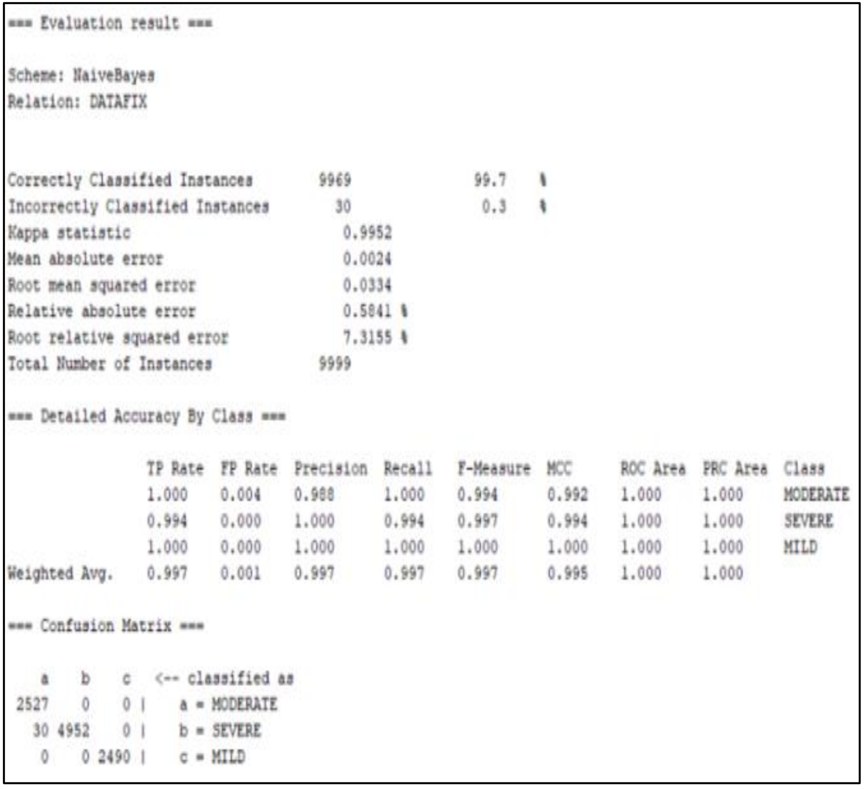


3.3. Algorithm Performance

This study utilized 3 data mining tools namely: WEKA, Orange and R Programming.

3.3.1. Performance Testing Using Weka

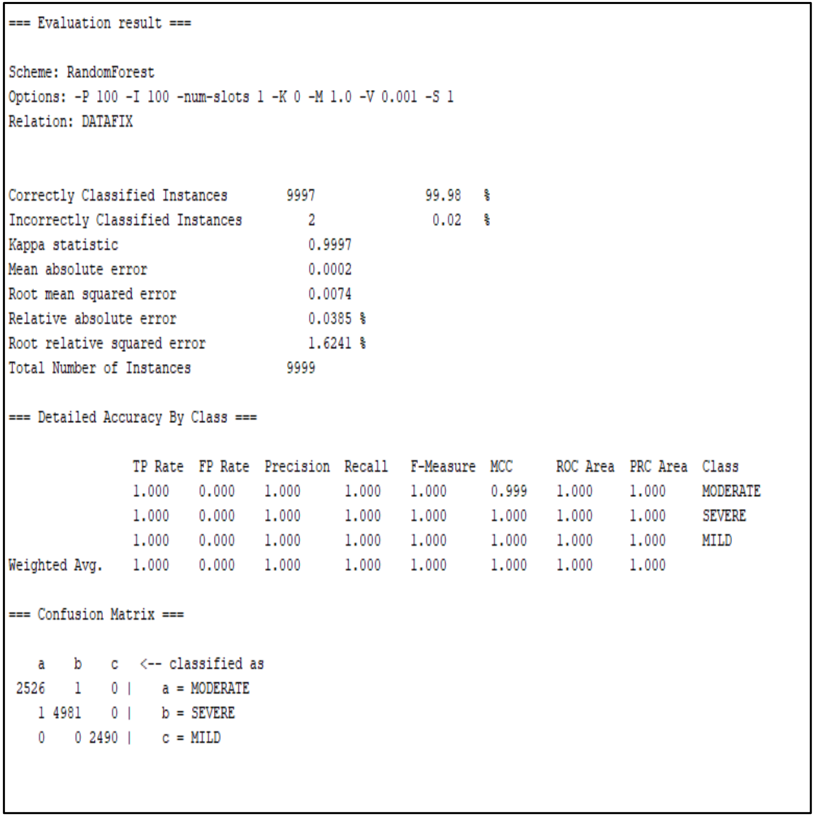
Figure-3. Naïve Bayes Algorithm



```
=== Evaluation result ===

Scheme: NaiveBayes
Relation: DATAFIX


Correctly Classified Instances        9969               99.7    %
Incorrectly Classified Instances        30                0.3    %
Kappa statistic                          0.9952
Mean absolute error                      0.0024
Root mean squared error                  0.0334
Relative absolute error                  0.5841 %
Root relative squared error              7.3155 %
Total Number of Instances             9999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.004    0.988      1.000    0.994      0.992    1.000     1.000     MODERATE
                 0.994    0.000    1.000      0.994    0.997      0.994    1.000     1.000     SEVERE
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     MILD
Weighted Avg.    0.997    0.001    0.997      0.997    0.997      0.995    1.000     1.000

=== Confusion Matrix ===

    a    b    c   <-- classified as
 2527    0    0 |    a = MODERATE
   30 4952    0 |    b = SEVERE
    0    0 2490 |    c = MILD
```

Figure 4 (above) shows that Naïve Bayes correctly classified 99.07 percent instances with 0.3 percent incorrectly classified instances.

Figure-4. Random Forests Decision Algorithm

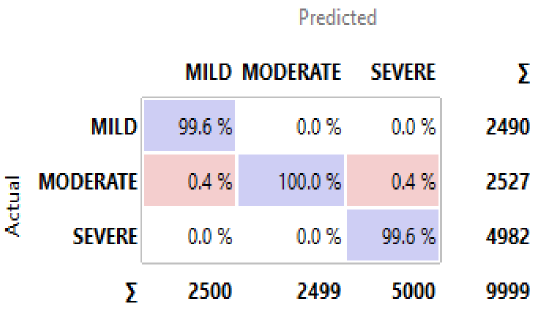


```
=== Evaluation result ===

Scheme: RandomForest
Options: -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation: DATAFIX


Correctly Classified Instances        9997               99.98   %
Incorrectly Classified Instances         2                0.02   %
Kappa statistic                          0.9997
Mean absolute error                      0.0002
Root mean squared error                  0.0074
Relative absolute error                  0.0385 %
Root relative squared error              1.6241 %
Total Number of Instances             9999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      0.999    1.000     1.000     MODERATE
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     SEVERE
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     MILD
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b    c   <-- classified as
 2526    1    0 |    a = MODERATE
    1 4981    0 |    b = SEVERE
    0    0 2490 |    c = MILD
```

Figure 5 Presents that Random Forests Decision Algorithm classified 99.98 percent correctly classified instances with 0.02 percent incorrectly classified instances. This clearly shows that Random Forests Decision Algorithm is the best fit for the important attributes.

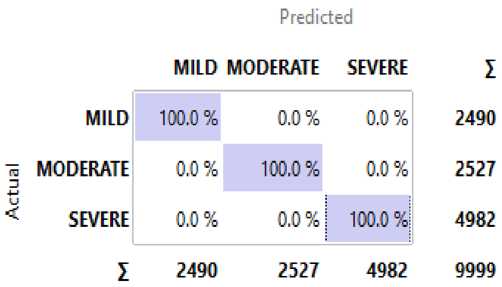
3.3.2. Performance Testing Using Orange

Figure-5. Naïve Bayes using Orange Data Mining



		Predicted			
		MILD	MODERATE	SEVERE	∑
Actual	MILD	99.6 %	0.0 %	0.0 %	2490
	MODERATE	0.4 %	100.0 %	0.4 %	2527
	SEVERE	0.0 %	0.0 %	99.6 %	4982
	∑	2500	2499	5000	9999

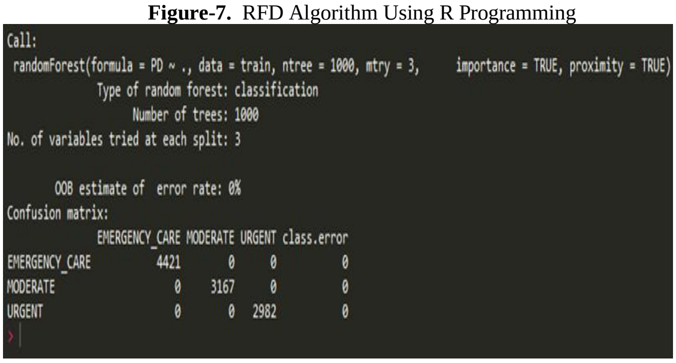
Figure-6. RFD Algorithm using Orange Data Mining



		Predicted			
		MILD	MODERATE	SEVERE	∑
Actual	MILD	100.0 %	0.0 %	0.0 %	2490
	MODERATE	0.0 %	100.0 %	0.0 %	2527
	SEVERE	0.0 %	0.0 %	100.0 %	4982
	∑	2490	2527	4982	9999

Figure 6 and Figure 7 shows that Random Forests Decision Algorithm predicted 100 percent accuracy compared to Naïve Bayes Algorithm which misclassified 0.40 percent instances. This shows that the Orange Data Mining Tool recommends Random Forests Decision Algorithm as it is more accurate than Naïve Bayes Algorithm.

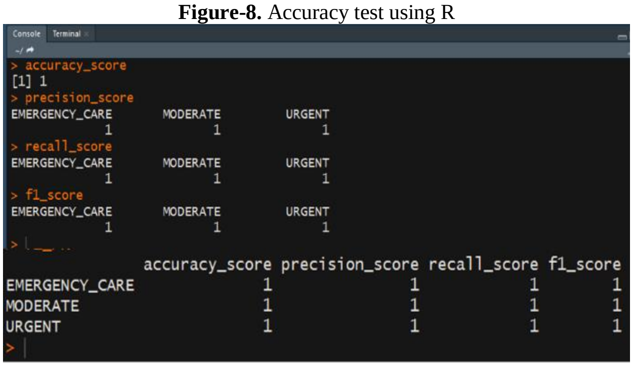
Figure-7. RFD Algorithm Using R Programming



```
Call:
 randomForest(formula = PD ~ ., data = train, ntree = 1000, mtry = 3,      importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0%
Confusion matrix:
               EMERGENCY_CARE MODERATE URGENT class.error
EMERGENCY_CARE           4421        0      0           0
MODERATE                    0     3167      0           0
URGENT                      0        0   2982           0
>
```

3.3.3. Performance Testing Using R

Figure-8. Accuracy test using R



```
> accuracy_score
[1] 1
> precision_score
EMERGENCY_CARE       MODERATE         URGENT
             1              1              1
> recall_score
EMERGENCY_CARE       MODERATE         URGENT
             1              1              1
> f1_score
EMERGENCY_CARE       MODERATE         URGENT
             1              1              1
>
               accuracy_score precision_score recall_score f1_score
EMERGENCY_CARE              1               1            1        1
MODERATE                    1               1            1        1
URGENT                      1               1            1        1
>
```

Figure 8. shows the urgency levels as Emergency Care, Moderate and Urgent which help indicate how soon does a patient need a dialysis treatment. Accuracy, precision, recall and F1 scores show 1.000 as a result, meaning the system can predict the Urgency Level of the patient accurately at 100%.

Figure-9. Important Attributes Selected by RFDA

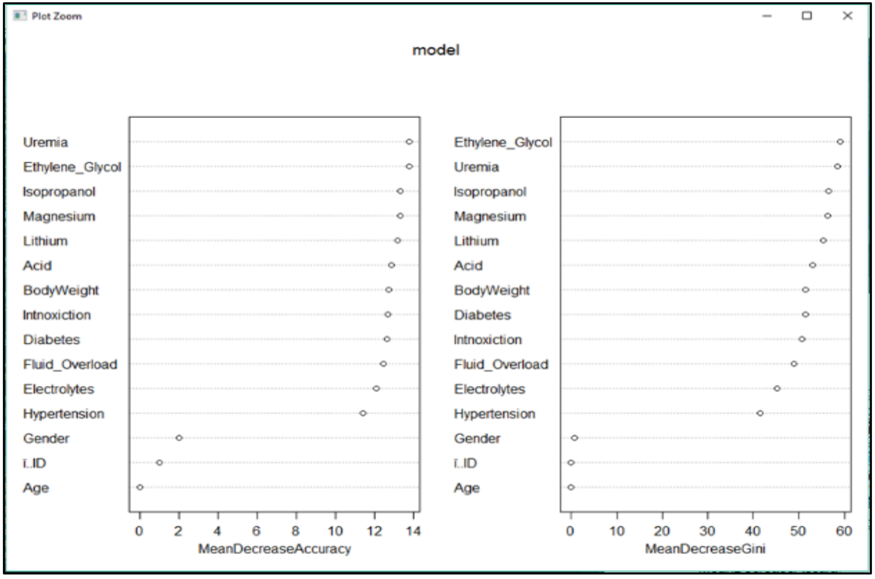


Figure 9. illustrates the important attributes to predict the urgency of dialysis using patients' data. According to Random Forests Decision Algorithm, uremia, ethylene glycol, magnesium and lithium are the most needed variables for proper prediction of urgency of dialysis.

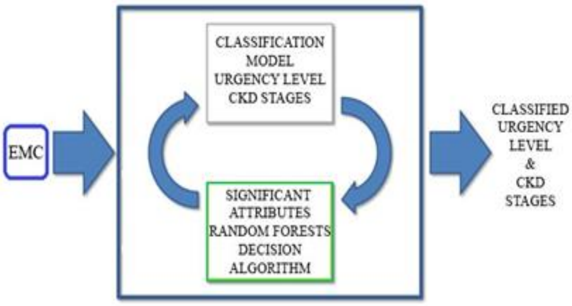
The study used Random Forests Decision Algorithm, ARIMA Model and K-means Clustering Algorithm in creating a rule-based system.

EMR data were loaded in R programming language to create a model using random forests decision algorithm.

RFDA C

```
rf_U <- randomForest(small_data.URGENCY~.,
data = attribs_train,
mtry = 3 , ntree = 1000,
proximity = TRUE,
importance = TRUE)
```

Figure-10. Simplified RFDA Framework


CLASSIFICATION MODEL URGENCY LEVEL CKD STAGES
EMC
SIGNIFICANT ATTRIBUTES RANDOM FORESTS DECISION ALGORITHM
CLASSIFIED URGENCY LEVEL & CKD STAGES

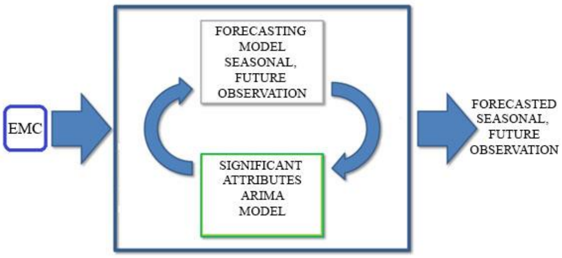
3.3.4. Using ARIMA Model

ARIMA (autoregressive integrated moving average) is a commonly used technique utilized to fit time series data and forecasting. It is a generalized version of ARMA (autoregressive moving average) process, where the ARMA process is applied for a differenced version of the data rather than original.

ARIMA Model Code

```
arima_ts <- auto.arima(dummy_ts_data)
#Arima model  forecast_ts <- forecast
(arima_ts, level = 95, h = 12)
```

Figure-11. Simplified ARIMA Model Framework

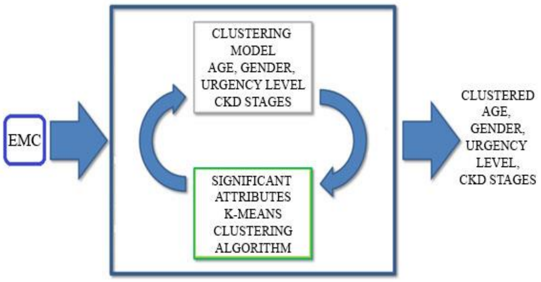

FORECASTING MODEL SEASONAL, FUTURE OBSERVATION
EMC
SIGNIFICANT ATTRIBUTES ARIMA MODEL
FORECASTED SEASONAL, FUTURE OBSERVATION

3.3.5. Using K-means Clustering Algorithm

KMCA Code

```
set.seed(123)
dataCluster <- kmeans(k_data_ckd_stage
[, 1:2], 5, nstart = 25)
table(dataCluster$cluster,
k_data_ckd_stage$small_data.STAGES)
dataCluster$cluster
<- as.factor(dataCluster$cluster)
```

Figure-12. Simplified K-Means Algorithm Framework


EMC
CLUSTERING MODEL AGE, GENDER, URGENCY LEVEL CKD STAGES
SIGNIFICANT ATTRIBUTES K-MEANS CLUSTERING ALGORITHM
CLUSTERED AGE, GENDER, URGENCY LEVEL, CKD STAGES

Data clustering has been an effective method for discovering structure in medical datasets. Clustering in a technique used to provide a structure to unstructured data so valuable information can be extracted. However, most medical datasets have overlapping information which could be best explained by clustering methods such as k-means (Khanmohammadi *et al.*, 2017). K-means algorithm is a partition-based algorithm which is used for clustering datasets into a number of clusters, most common for clustering due to its simplicity and efficiency (Chauhan and Shukla, 2015). Although the main challenge for using the algorithm is the need to recompute the nearest centroid for every data point at every iteration which has a prohibitive cost when the number of clusters is large the algorithm can be very useful to analyze unstructured datasets to derive meaningful information (Anchalia *et al.*, 2013).

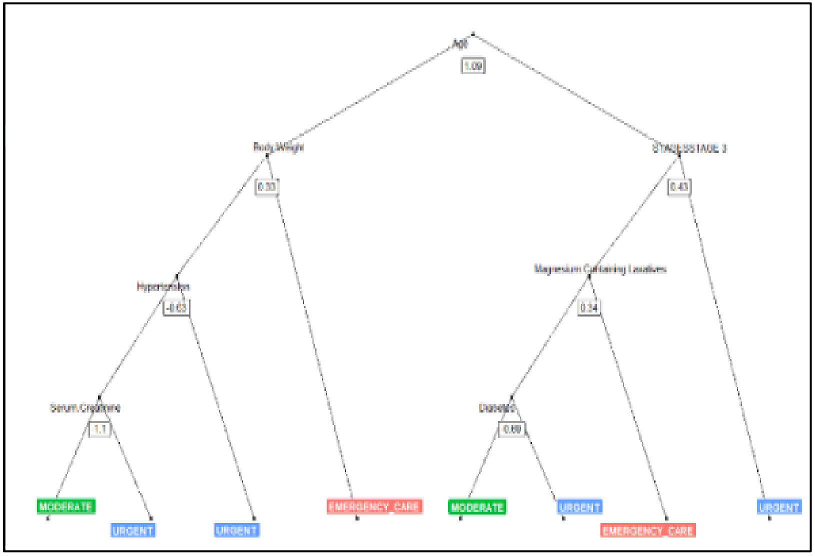
4. Results and Analysis

The aim of this chapter is to present the classification, progression and cluster of patient's diagnosis based on the features and values that were extracted from the electronic medical records, as well as the prototype software implementation of Degree of Urgency and Progression Predictive Model using Hybrid System

4.1. Applying Random Forests Decision Algorithm

Figure 13 Illustrates how the Random Forests Decision Algorithm was created according to important attributes selected. important attributes were used to predict the Urgency of Dialysis.

Figure-13. Predicting Urgency Level



This shows how Random Forests Decision Algorithm used the important attributes to predict the urgency of dialysis. Patients age, bodyweight, hypertension, magnesium containing laxatives, diabetes and serum creatinine were used by RFD Algorithm to predict the level of urgency. The aggregated results show that two (2) votes for moderate, two votes (2) for emergency care. The RFD Algorithm predicted the level of urgency with four (4) votes for urgent dialysis. This shows the majority of the decision of the n-trees predicted the male patient in emergency care for dialysis. The prediction was based on the attributes of the patient's age, bodyweight, electrolytes, fluid overload volume and ethylene glycol. The two other important attributes are serum creatinine and magnesium containing laxatives. The patient is hypertensive and diabetic affecting the urgency of dialysis. The patient also has Stage 3 chronic kidney disease

Figure-14. RFDA Classification Model

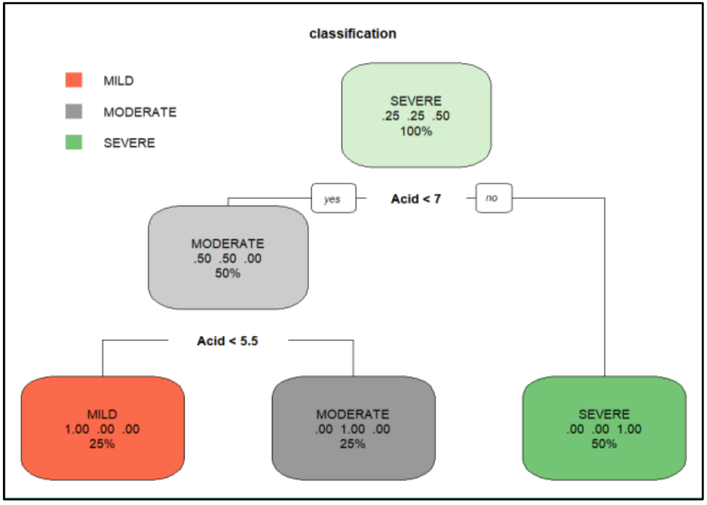


Figure 14 illustrates 20, 000 training data on the attribute of acid level, where the model predicted 25 percent of the patients having moderate urgency for dialysis, and another 25 percent as urgent. The model has also predicted that 50 percent of the training data is under emergency care condition based on acid level.

Figure-15. Predicting CKD Stages

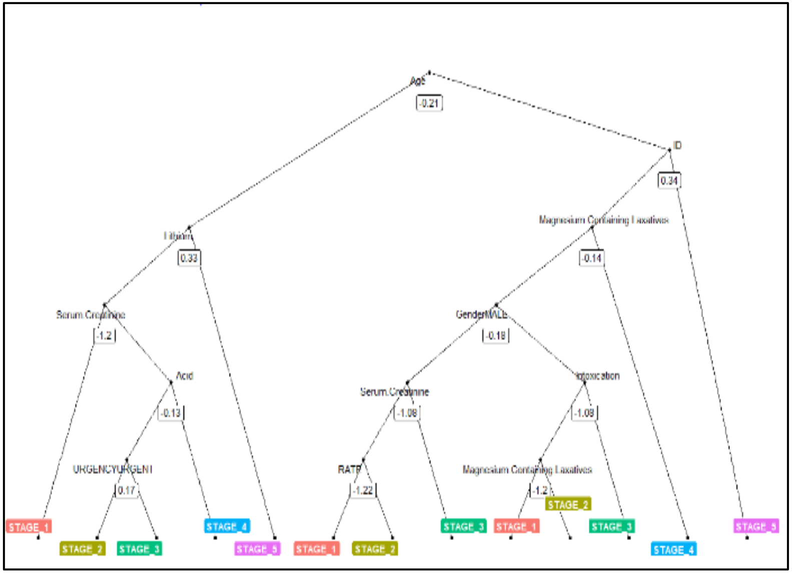


Figure 15. illustrates patients with CKD can be classified depending on their level of kidney function, or eGFR, and the amount of protein present in the urine. This information forms the basis of CKD staging into 5 stages which is useful for planning follow up and management. The higher the stage *(G1->G5)* and the greater the amount of protein present in the urine *(A1->A3)* the more "severe" the CKD.

4.2. Applying Arima Model

Figure 16. illustrates the future observation of the patient urgency rate and level. The system forecasted that the patient will be on 93 % percent of urgency rate in the next 6 days. The system forecasted that the patient's status will be on emergency care and CKD Stage 5.

Figure-16. Yearly Observation ARIMA Model

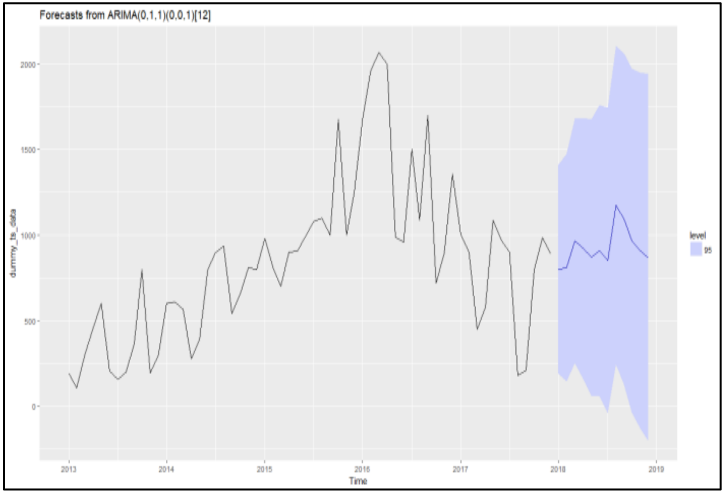


Figure 17. illustrates the patient's future observation of CKD Stages. The patient's creatinine level is gradually decreasing from 2.9 to 1.1, with unstable patterns and observations in between. The system has also forecasted that it will go below at 1.1.

Figure-17. Future Observation (CKD Stages)

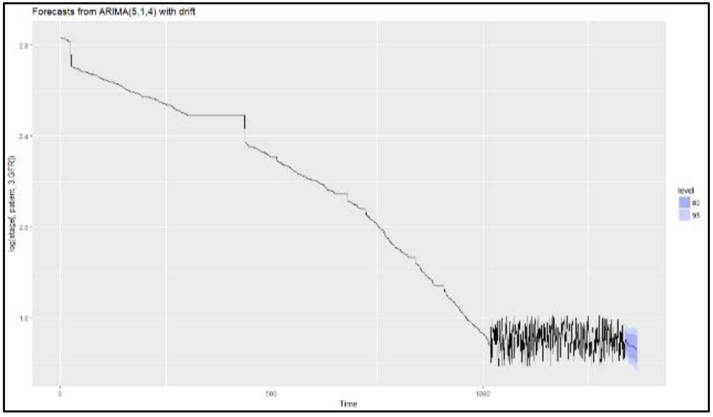


Figure-18. Future Observation (Urgency Level)

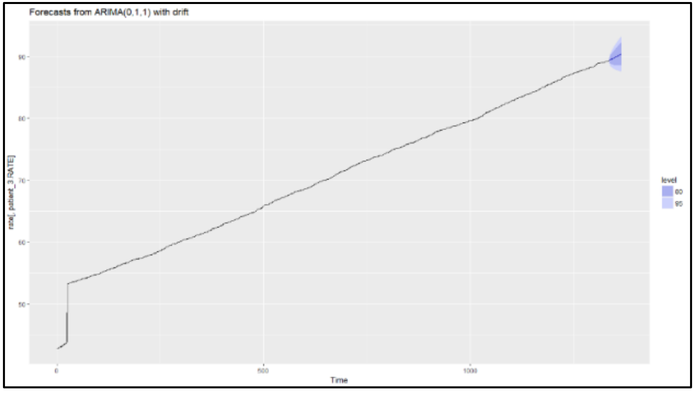


Figure 18. illustrates the future observation of the patient urgency rate and level. The system forecasted that the patient will be on 93% percent of urgency rate in the next 6 days and the patient's status will be on emergency care with a CKD of Stage 5.

4.3. Applying K-means Clustering Algorithm

K-means Clustering Algorithm main function is to cluster the number of patients according to age, gender, urgency level and CKD stages. The results will be a big factor in decision making of the Doctors, Hospital Administrators and future Researchers.

Figure-19. Cluster by Age

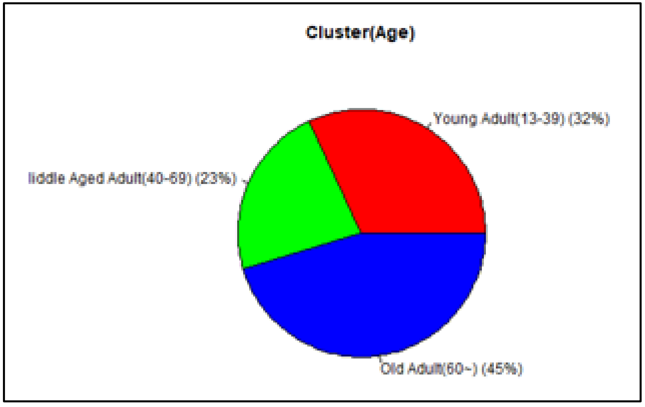


Figure 19. illustrates that twenty three percent (23%) were middle aged adult (age 40-69) clustered, while young adult (age 13-39) shows thirty two percent (32%) clustered, and old adult (60 and above) shows 45 percent (45%) clustered which is the highest in ranking. This clearly shows that old adult patients are on a more severe state in urgent and emergency care.

Figure-20. Cluster by Gender

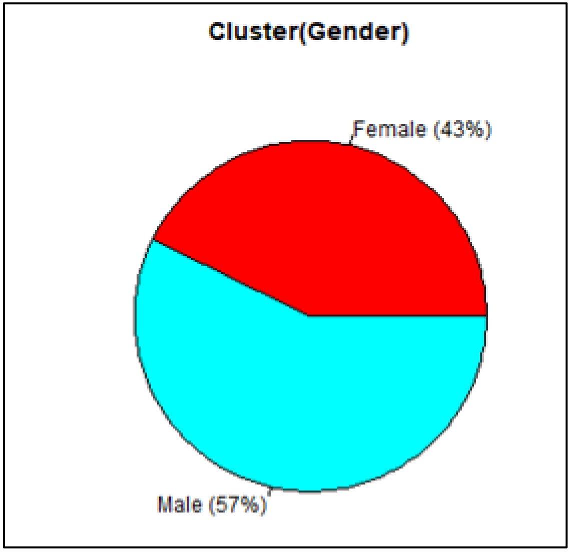


Figure 20. shows show that female patients are least admitted both in CKD and ARF with forty three percent (43%), while male patients have the most admitted both in CKD and ARF with fifty seven percent (57%). This indicates that male patients are admitted most both in CKD and ARF.

Figure 21. Cluster by Urgency Level

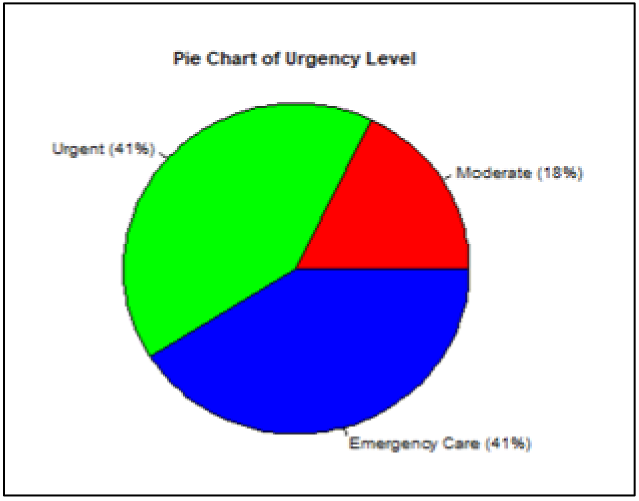


Figure 21. illustrates the clustered urgency level of a patients. Moderate Level has the least which is 18% (percent), while urgent and emergency care level tied with 41 % (percent). This indicates that most of the CKD and

ARF patients do not comply the recommended medication which leads to a critical stage both for urgency level and CKD stages.

Figure-22. Cluster by CKD Stages

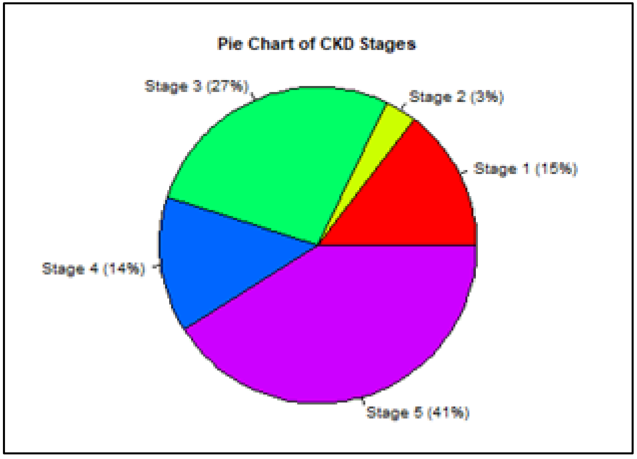


Figure 22. illustrates the clustered according to CKD stages of a patients. Stage two has 3 % (percent) clustered, stage four has 14 % (percent) clustered, stage one has 15 % (percent) clustered, stage three has 27% (percent) clustered, while stage five has 41 % (percent) clustered.

The result clearly shows that most of the patients developed CKD stage 5. These patients must do a frequent dialysis (3X) a day and recommend for kidney transplant.

Patients, especially those having chronic or renal failure, can receive better, more affordable healthcare services with appropriate identification, tracking and use of appropriate interventions and treatment protocols. This study can improve patient satisfaction, provide more patient-centered care, and decrease costs and increase operating efficiency of the dialysis center while maintaining high-quality care

5. Conclusion

The objective of this study was to develop a model to be used in predicting the degree of urgency and progression of dialysis patients. This research was intended to provide the dialysis centers in the Philippines to streamline and increase efficiency. A nephrologist would be assisted by the model prototype as patients will be informed about the status and possible stage of urgency and progression of dialysis. This study customizes the process using classification, forecast and cluster model.

The patient will be given an immediate recommendation of the nephrologist using the prediction model system. This way, the patient will be informed that needs and immediate dialysis based on patient hospital records. Dialysis Centers will be able to use the system for efficiency in helping at-risk patients.

The study can lead to a better decision-making capability of the hospital administrators, improve administrative services and reduced costs.

5.1. Implications of the Study

The study covered the years 2016 to 2018 where the urgency and the progression of dialysis of a particular patient were predicted. The study also showed the use of Random Forests Decision Algorithm in the classification of urgency level and CKD stages of dialysis with the accuracy rate of 99.07 percent (%) respectively which gives a precise recommendation and optimal treatment to start the dialysis of a patient. Meanwhile, the ARIMA Model presented an impressive result in forecasting the progression of dialysis with a result of 97 percent (%) accuracy rate. K -Means Clustering Algorithm, a partition-based algorithm which is used for clustering datasets into a number of clusters, was also utilized in the study and it showed clusters according to age and gender with the use of RFDA and ARIMA model.

Hybrid system was used to provide relevant information that can help patient medication and quality of life. A dashboard was used to visualize the output data.

The experiments revealed that a surprising amount of useful knowledge can be gathered from this type of models and structures. The specific examples reported in this research shows that is possible to discover knowledge from Hybrid system that is useful for medical experts. While medicine is possibly one of the richest domains for data mining engineers, it is definitely the toughest. To overcome this, we think that the scientific community needs to address the following challenges.

6. Recommendations

The aim of this study was to determine the urgency and progression of dialysis wherein model was created using Random Forests Decision Algorithm, ARIMA Model and K-means Clustering Algorithm. The insights and findings provided were able to predict the urgency and progression of a patient’s need for dialysis and the study has also developed a model and dashboard enhancing efficiency and accuracy of results based on patient’s medical record in which nephrologists, dialysis centers, hospital administrators and patients can highly benefit from. In line with this, the following recommendations for future research may be considered:

1. Build a new model with increased number of data sets and variables.
2. Explore on other potential attributes that may contribute to a more accurate classification, forecast and cluster of dataset.
3. Conduct a future study that can assist future researchers by giving an idea and enhancement of their predictive management and data mining topics.
4. The design of tools to automate some resource-consuming time series analysis tasks, such as preparation.
5. The specification of secure models for medical time series storage and publication with the aim of increasing efficient data reuse and processing.

6.1. Other Recommendations in the Field of Medicine include:

1. Hospital management or administrators may consider a venture on research on developing systems and approaches that can use algorithms to help in the efficiency of predicting the data gathered from patients along with the diagnosis of doctors, helping in streamlining processes of the current working operations.

The increase of 0.7% annual mortality rate per year since 1990-2013 exhibits an alarming situation of the Philippines when it comes to treating chronic kidney disease alone. Efforts in improving the health care system of the country may be looked into as problems may be posited gradually with the use of technology, particularly machine-learning systems utilizing the use of data mining tools to better measure the status of patients with other chronic illnesses.

References

Aghamohseni, A. and Ramezanian, R. (2015). An efficient hybrid approach based on K-means and generalized fashion algorithms for cluster analysis. In 2015 AI and Robotics, IRANOPEN 2015 - 5th conference on artificial intelligence and robotics. Available: https://doi.org/10.1109/RIOS.2015.7270727

Akram, W., Ezzat, Ihab, A., Wahbi, Z. and Wahbi, A. (2018). Hybrid PCM and Transparent Solar Cells in Zero Energy Buildings. *International Journal of Technology and Engineering Studies,* 4(2): 102-11.

Anchalia, P. P., Koundinya, A. K. and Srinath, N. K., 2013. "Mapreduce design of K-means clustering algorithm." In *International Conference on Information Science and Applications, ICISA 2013.*

Anguera, A., Barreiro, J. M., Lara, J. A. and Lizcano, D. (2016). Applying data mining techniques to medical time series, An empirical case study in electroencephalography and stabilometry. *Computational and Structural Biotechnology Journal*: Available: https://doi.org/10.1016/j.csbj.2016.05.002

Celebi, M. E., Kingravi, H. A. and Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*: Available: https://doi.org/10.1016/j.eswa.2012.07.021

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). The CRISP-DM User Guide. NCR Systems Engineering Compenhagen.

Chauhan, P. and Shukla, M., 2015. "A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm." In *International Conference on Advances in Computer Engineering and Applications.*

Fathi, F., Majari-Kasmaee, L., Mani-Varnosfaderani, A., Kyani, A., Rostami-Nejad, M., Sohrabzadeh, K. and Arefi-Oskouie, A. (2014). 1H NMR based metabolic profiling in Crohn’s disease by random forest methodology. *Magnetic Resonance in Chemistry,* 52(7): 370–76. Available: https://doi.org/10.1002/mrc.4074

Ha, L. T. D. and Tsai, K. M. T. (2015). Numerical study on optimization of wooden-steel hybrid beams base on shape factor of steel component. *International Journal of Technology and Engineering Studies*, 1(2): 53-62. Available: https://doi.org/10.20469/ijtes.40004-2

Jeewantha, R. A., Halgamuge, M. N., Mohammad, A. and Ekici, G., 2017. "Classification performance analysis in medical science, Using kidney disease data." In *ACM International Conference Proceeding Series.*

Khanmohammadi, S., Adibeig, N. and Shanehbandy, S. (2017). An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications*: Available: https://doi.org/10.1016/j.eswa.2016.09.025

Köpcke, F., Lubgan, D., Fietkau, R., Scholler, A., Nau, C., Stürzl, M. and Toddenroth, D. (2013). Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. *BMC Medical Informatics and Decision Making*: Available: https://doi.org/10.1186/1472-6947-13-134

Louppe, G. (2014). *Understanding random forests.* Cornell University Library. https://doi.org/10.13140/2.1.1570.5928

Mahoto, N. A., Shaikh, F. K. and Ansari, A. Q. (2014). Exploitation of clustering techniques in transactional healthcare data. *Mehran University Research Journal of Engineering and Technology*:

Merino, M., Gómez, I. M. and Molina, A. J. (2015). Envelopment filter and K-means for the detection of QRS waveforms in electrocardiogram. *Medical Engineering and Physics*: Available: https://doi.org/10.1016/j.medengphy.2015.03.019

Mishina, Y., Murata, R., Yamauchi, Y., Yamashita, T. and Fujiyoshi, H. (2015). Boosted random forest. *IEICE Transactions on Information and Systems,* E98D(9): 1630–36. Available: https://doi.org/10.1587/transinf.2014OPP0004

Moskovitch, R. and Shahar, Y. (2015). Classification-driven temporal discretization of multivariate time series. *Data Mining and Knowledge Discovery*: Available: https://doi.org/10.1007/s10618-014-0380-z

Ordon, M., Urbach, D., Mamdani, M., Saskin, R., D'A Honey, R. J. and Pace, K. T. (2014). The surgical management of kidney stone disease, A population-based time series analysis. *The Journal of Urology*: Available: https://doi.org/10.1016/j.juro.2014.05.095

Pandeeswari, L. and Rajeswari, K. (2015). K-Means clustering and naive bayes classifier for categorization Of Diabetes Patients. *IJISET - International Journal of Innovative Science, Engineering & Technology*:

Razlighi, Q. R., Stallard, E., Brandt, J., Blacker, D., Albert, M., Scarmeas, N. and Stern, Y. (2014). A new algorithm for predicting time to disease endpoints in Alzheimer's disease patients. *Journal of Alzheimer's Disease*: Available: https://doi.org/10.3233/JAD-131142

Renal Disease Control Program, R. National Kidney and Transplant Institute. Available: http://www.nkti.gov.ph/index.php/services/specialty-centers/renal-disease-control-program-redcop

Sanchez-Morillo, D., Fernandez-Granero, M. A. and Jiménez, A. L. (2015). Detecting COPD exacerbations early using daily telemonitoring of symptoms and k-means clustering, A pilot study. *Medical and Biological Engineering and Computing*: Available: https://doi.org/10.1007/s11517-015-1252-4

Sankar, K. P. and Mitra, P. (2004). Pattern recognition algorithms for data mining. New York: Chapman & Hall/CRC. Randhawa, R., Sohal, J. S. and Kaler, R. S. 2009. Optimum algorithm for wdm channel allocation for reducing four-wave mixing effects. *Optik*, 120: 898–904.

Scorecki (2016). *Brenner and Rector's the Kidney*. 10th edn2: 127-46.

Scornet, E., Biau, G. and Vert, J. P. (2015). Consistency of random forests. *Annals of Statistics*, 43(4): 1716–41. Available: https://doi.org/10.1214/15-AOS1321

Spinner, M. (2014). Mnemonic Monday, AEIOU – Indications for Dialysis in Patients with Acute Kidney Injury. Available: https://firstaidteam.com/2014/04/07/mnemonic-monday-aeiou-indications-for-dialysis-in-patients-with-acute-kidney-injury/

Su, T. J., Tsou, T. Y., Wang, S. M., Hoang, V. M. and Pin, K. W. (2016). A hybrid control design of FOPID and FWA for inverted pendulum systems. *Journal of Applied and Physical Sciences*, 2(3): 89-95. Available: https://doi.org/10.20474/japs-2.3.4

Sugiyatno, Djunaedi, I. and Mahardiono, N. A. (2015). Modelling and simulation of hybrid control system in solar cell-battery-super capacitor. *International Journal of Technology and Engineering Studies*, 1(3): 74-80. Available: https://doi.org/10.20469/ijtes.40002-3

Tao, W. Z., XingYu, W. Z., YuanYuan, W. L. and XiaoSong, W. L. (2013). The application of the Kalman, filter algorithm based on ARIMA, Model in forecasting the incidence rate of gonorrhea in China. *Modern Preventive Medicine*:

Tiao, G. C. (2015). Time series, ARIMA Methods. *International Encyclopedia of the Social & Behavioral Sciences*: Available: https://doi.org/10.1016/B978-0-08-097086-8.42182-3

Umer, F. C., N. (2017). Transient analysis due to short circuit faults in wind hybrid systems. *Journal of Advances in Technology and Engineering Research* 3(3): 89-100. Available: https://doi.org/10.20474/jater-3.3.4

Zheng, B., Yoon, S. W. and Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*: Available: https://doi.org/10.1016/j.eswa.2013.08.044