Health Application for Women using Decision Tree-Based Classifier

Chona B. Sabinay, Maria Visitacion N. Gumabay

Abstract*—Using the descriptive and developmental design, the study developed and evaluated an eHealth application for women using decision tree classifiers. It focuses on the development of an eHealth application system using open-access datasets from UCI Machine Learning Repository. This attempts to predict the onset of diabetes and chronic kidney diseases grounding from the generated predictive models. These decision models are created using C4.5, ID3 and CART algorithms with RapidMiner data science platform. Performance metrics are deployed such as accuracy, recall, precision and error rate to compare the reliability of each model. Models incurred the highest assessment are the bases of the developed system following Agile Software Development Life Cycle Model.

Easy access to healthcare workers through teleconsultation, diabetes and chronic kidney disease (CKD) online diagnosis, and maternal care videos are possible with this study.

The summary of the evaluation showed that the eHealth Application got an overall average weighted mean of 3.98, which is described as high extent. Based on the respondents' response, the strongest point of the system was its portability, which earned the highest average mean among categories of system evaluation. Thus, the system addresses the shortcomings of healthcare in terms of distance and timeliness of treatment fostering an equal access to healthcare.

Keywords: Classifier, Data Mining, eHealth, Prediction, Diagnosis.

I. INTRODUCTION

e-Health is the single most important revolution in healthcare since the advent of modern medicine, vaccines, or even public health measures like sanitation and clean water. It is an emerging field of medical informatics, referring to the organization and delivery of health services and information using the Internet and related technologies. According to Eysenbach, "the term embodies not only a technical development, but also a new way of working, an attitude, and an obligation for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology" [1].

Information Communication Technology has revolutionized healthcare in many developing countries by efficiently disseminating public health information and assisting consultation on health issues.

In the Philippines, the percentage of population living in urban areas is 48.8%. The rest of it is in rural and even remote areas having no access to state-of-the-art health institutions. For Biliran Province, out of 132 barangays, 20 of these are classified as urban areas and the remaining 113 barangays fall into rural classification areas. Additionally, in the entire province, there are 1,074 active Barangay Health Workers (BHWs) and 128 trained Birth Attendants assisting in the delivery of field health services. There is also one (1) physician per municipality with a total of eight (8) within the entire province [2].

As stated in the Biliran Provincial Development and Physical Framework Plan 2011-2016 (PDPFP) of the province, while access to health services is now easier, especially among the poor, public health delivery is still inundated by problems such as lack of health personnel, ill-equipped health centers and topographic location of the province.

Data mining is the process of pattern discovery and extraction where huge amount of data is involved. Consequently, both data mining and healthcare industry have emerged some of reliable early detection systems and other various healthcare related systems from the clinical and diagnosis data [3].

As a solution, the development of a software application that is purposive to this need is justifiable. With data science, the prediction of common chronic diseases is possible. Diabetes and chronic kidney disease data sets are utilized to create predictive models as basis for prediction using decision tree classifiers.

Conceptual Framework

The conceptual framework of this research was adopted based on the results of the study of Saw, et al. (2006). It deliberates the overlapping nature of eHealth. The study emphasized the distinct feature of eHealth along its boundaries [4].

This taxonomy categorizes telemedicine into three (3) domains: type of technology; the perspective of the individual, such as client or practitioner; and context in which e-Health is being applied to.

By adopting this taxonomy and using the Agile Software Development Model (ASDM), the researcher came up with the desired eHealth application.

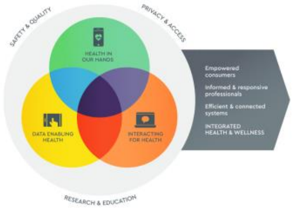


Fig. 1. A Conceptual Model for eHealth (Saw, et al., 2006)

Revised Manuscript Received on April 12, 2019.

Chona B. Sabinay, Biliran Province State University, Biliran, Philippines (E-mail: chonasabinay14@gmail.com)

Maria Visitacion N. Gumabay, St. Paul University Philippines, Cagayan 3500, Philippines. (E-mail: gumabayvc@gmail.com)

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

The overlapping characteristic of the model recognizes the complexity of eHealth while providing a practical way of understanding how eHealth is perceived and implemented. The model provides a conceptual framework that can assist individuals and organizations in developing and integrating eHealth initiatives and transforming current models of care. They proposed that interventions incorporating multiple domains have the greatest potential impact. For example, the author of this study who developed an application targeting self-management of a chronic kidney disease condition and diabetes considered how the user interacts with technology to monitor or manage their condition (i. e., health in our hands); how it offers opportunities for communication and interactions with healthcare professionals through chats, sending images and video call technology (interacting for health); and how gathered data are stored, managed, and analyzed for immediate decision support using classification techniques in data mining (i.e., data enabling health).

Statement of the Problem

This paper aimed to develop an e-Health system focusing on chronic diseases among women in rural and remote areas by implementing a designed decision-tree based classifier in machine learning.

Specifically, it answered the following questions:

1. What are the significant correlates of diabetes and chronic kidney disease?
2. What are the predictors of diabetes and chronic kidney disease?
3. What decision tree-based classifier algorithm is most appropriate to build predictive models for diabetes and chronic kidney disease in terms of:

a. Classification Accuracy Rate;
b. Classification Recall;
c. Classification Precision;

4. What proposed e-health application can be developed for women using a decision tree-based classifier?
5. What is the extent of compliance of the developed application to ISO 25010 Software Quality Assurance Standards in terms of:

a. Functional Sustainability,
b. Performance Efficiency,
c. Compatibility,
d. Usability,
e. Reliability,
f. Security,
g. Maintainability, and
h. Portability?

6. What are the strengths and limitations of the developed system?

II. METHODOLOGY

Several stages were strategized to come up the output, namely: (a) Data Correlation; (b) Data Preparation; (c) Data Modeling; (d) System Development; and (e) Software Quality Assessment.

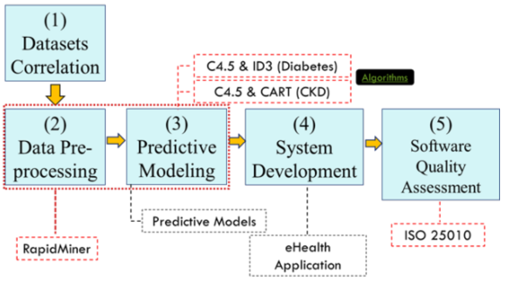


Fig. 2. Phases of the Research Development

Research Design

The study made use of descriptive survey and research system development to develop and evaluate an eHealth application to predict the onset of diabetes and chronic kidney disease.

Participants of the Study

The participants of this study (n=40), otherwise known as the "evaluators", are IT experts. This population is purposively selected to assess the software quality.

Additionally, diabetes and chronic kidney diseases datasets are taken from University of California Machine Learning Repository, School of Information and Computer Science.

Instrumentation

The researcher used a standardized questionnaire for software quality assessment. The questionnaire includes items to describe the system in terms of functional suitability, performance efficiency, compatibility, usability, reliability, security, and maintainability. This research instrument was developed by ISO/IEC as ISO 25010 - Software Product Quality Standards.

Data Gathering Procedures

The development of this research is undertaken by searching, sending letter for datasets utilization and downloading datasets online. The request for proper communication protocol was observed in the dissemination of questionnaires to system evaluators.

Data Analysis

The data collected were tabulated, analyzed, interpreted, and summarized using both descriptive and inferential statistics. The data were analyzed using the Statistical Package for Social Science for Windows (SPSS for Windows).

Mean was used to analyze the average rating of the IT experts with respect to the compliance of the application that was developed in this study in relation to the ISO standard.

Correlation was also used to determine the correlates of the two (2) datasets against its target outcome labels.

Kappa statistics was included as part of a predictive model's performance metrics to further test the classifiers predictors' reliability. Furthermore, different data mining algorithms were employed for data mining and data modeling. Selected performance metrics were used and analyzed to select the most accurate and best algorithms for implementation.

III. RESULTS AND DISCUSSIONS

The aim of this study was to develop and assess an eHealth application for basic diagnosis of diabetes and chronic kidney disease using selected decision tree classifiers.

A. Comparison of Classifiers using RapidMiner

Two (2) datasets were utilized in this study. For chronic kidney disease and diabetes prediction, related data available on the web are used for training and prediction. These datasets were downloaded from the UC Irvine Machine Learning Repository named Chronic Kidney Disease uploaded in 2015 and the Pima Indians Diabetes Database (Dua & Karra Taniskidou, 2017).

The chronic kidney disease dataset has 25 attributes: 11 numeric and 14 nominal. A total of 400 instances of the dataset were used for the training to prediction model, out of which 250 has label chronic kidney disease (CKD) and 150 has label nonchronic kidney disease (NCKD). Likewise, the diabetes dataset has 768 instances with 9 attributes that are all numeric-valued.

These datasets were sujected to data pre-processing, data reduction, outlier removal, discretization, and data transformation.

After preprocessing and discretization, C4.5, ID3 and CART decision tree classification algorithms have been used to generate the diabetes predictive model. A cross-validation technique is used to estimate the statistical performance of the learning models being generated. A 10-fold cross-validation was used to prepare training and test data with a stratified sampling type. Both models are evaluated on the basis its perfomance metrics as reflected in Fig. 6 and 8.

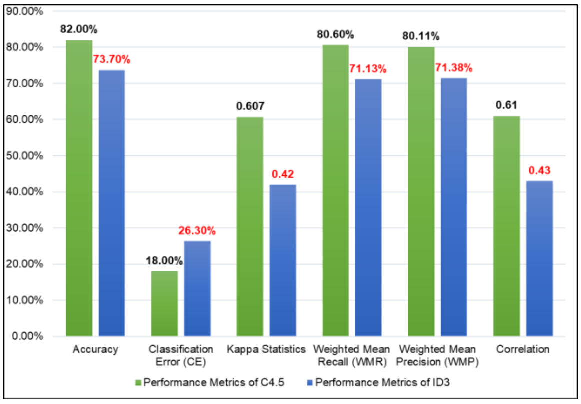


Fig. 3. Graphical Comparison of Performance Parameters of ID3 and C4.5 in Diabetes Dataset

It can be observed from Fig. 3 that C4.5 algorithm gives a classification rate accuracy of 82% against the 73.70% of ID3. It further shows the values derived in each algorithm based on the performance errors. The values of Kappa statistics, Weighted Mean Recall, and Weighted Mean Precision for C4.5 tends to decline when the processed dataset is applied with ID3. Naturally, the classification error value of ID3 is higher than C4.5. This result reveals that the C4.5 algorithm is suitable for the prediction of diabetes since the lesser the error value the better the prediction.

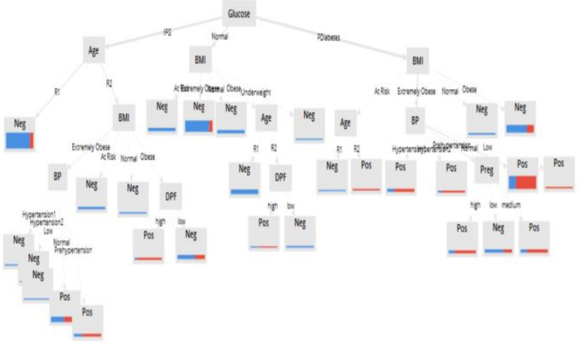


Fig. 4. Generated Pruned Tree Using C4.5 for Diabetes

Fig. 4 presents the decision tree using C4.5 algorithm. It implies that the Glucose attribute value has the highest influence in predicting diabetes as evident on the root node of the tree.

As shown in the figure, if an individual has IFG (Impaired Fasting Glucose), Age and BMI values will be tested and compared. The evaluation continues until a leaf node (positive/negative) is reached. These statements then were coded during the software development phase.

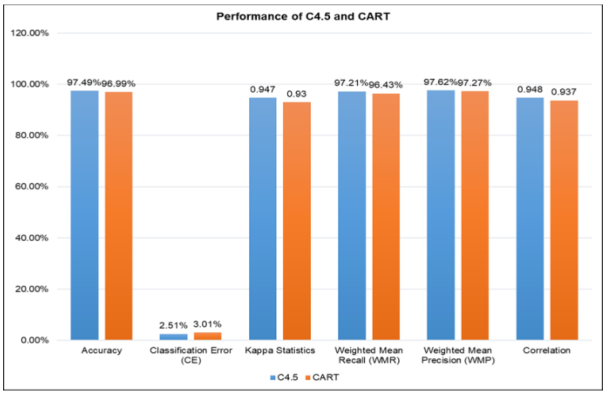


Fig. 5. Graphical Comparison of Performance Parameters of CART and C4.5 in CKD Dataset

As shown in Fig. 5, the percentage of accuracy is measured, it can be clearly seen that C4.5 classification tree provides the highest accuracy of 97.49%. It is also noted that CART algorithm has 96.99% accuracy.

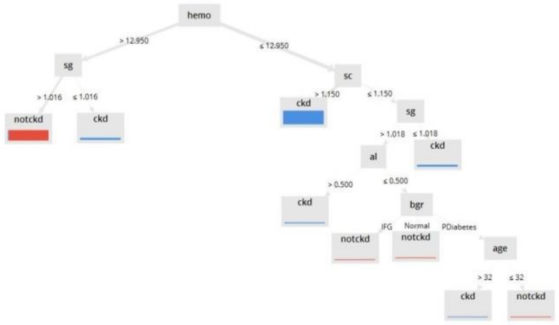
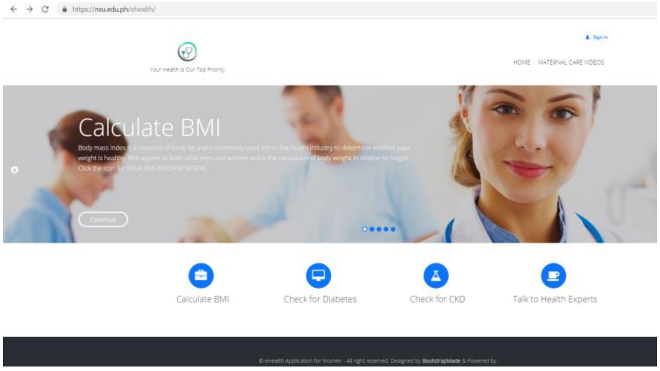


Fig. 6. Generated Decision Tree Using C4.5 for CKD

The generated pruned tree using C4.5 is presented in Fig. 6. It can be seen that the root node is the attribute hemoglobin. This indicates that hemoglobin value has the highest influence in predicting CKD. Additionally, specific gravity (sg), sugar count (sc), albumin, age, and blood glucose are considered also as good predictors of CKD.

B. eHealth Application Interface

The application was designed to operate on a web-based and mobile application that operate on desktop computer, laptop, tablet and android smart phones. The web-based application is designed using HTML, CSS, JQuery, Javascript and PHP was used for the front-end and MySql for the back-end. The notepad++ as the IDE and Xampp for local development were used. Also, Adobe Photoshop CS3 for web/graphic design. Its primary function is for basic diagnoses for diabetes and CKD, and teleconsultation.



https://nsu.edu.ph/ehealth/

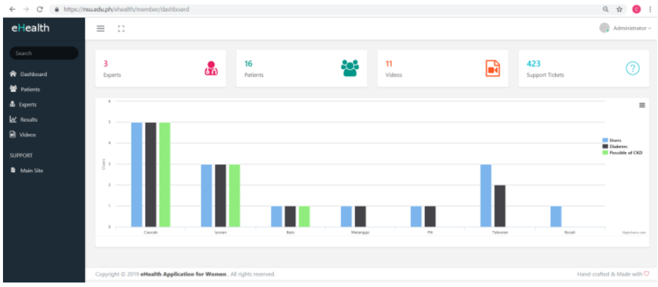
Your Health is Our Top Priority

Calculate BMI

Body mass index is a measure of body fat and is commonly used within the health industry to determine whether your weight is healthy. BMI applies to both adult men and women and is the calculation of body weight in relation to height. Click the icon for YOUR BMI INTERPRETATION.

© eHealth Application for Women - All right reserved. Designed by BootstrapMade & Powered by .

Fig. 7. eHealth Application Home Page



https://nsu.edu.ph/ehealth/member/dashboard

3 Experts

16 Patients

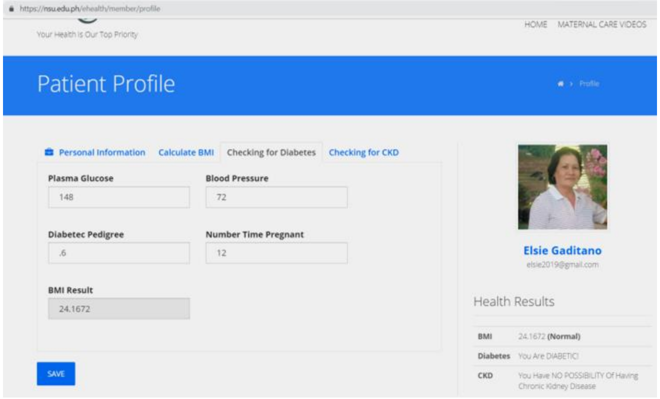
11 Videos

423 Support Tickets

Copyright © 2019 **eHealth Application for Women** . All rights reserved.

Hand-crafted & Made with ♡

Fig. 8. Administrator's Dashboard



https://nsu.edu.ph/ehealth/member/profile

Your Health is Our Top Priority

Patient Profile

Plasma Glucose: 148

Blood Pressure: 72

Diabetec Pedigree: .6

Number Time Pregnant: 12

BMI Result: 24.1672

Elsie Gaditano

elsie2019@gmail.com

Health Results

BMI 24.1672 **(Normal)**

Diabetes You Are DIABETIC!

CKD You Have NO POSSIBILITY Of Having Chronic Kidney Disease

Fig. 9. Patient's Window

C. Extent of Compliance of the eHealth Application System to ISO 25010 Software Quality Assurance Standards

Areas of Evaluation	Mean	Description
Functional Sustainability	4.17	High Extent
Performance Efficiency	3.97	High Extent
Compatibility	3.92	High Extent
Usability	3.85	High Extent
Reliability	3.79	High Extent
Security	3.75	High Extent
Maintainability	4.07	High Extent
Portability	4.34	Very High Extent
Overall Mean	**3.98**	**High Extent**

Table 1. Summary of Evaluation of the eHealth Application for Women

As revealed in Table 1, it can be inferred that out of the eight (8) areas of the application being evaluated, portability was marked as compliant to a very high extent which has a mean of 4.34. This implies that the high rating is due to the easy access of the system through mobile phones and PCs which means that the system works in remote areas provides that there is a stable internet connection.

It obtained an overall mean of 3.98 classified as high extent. This implies that the all the areas being evaluated surpassed the minimum requirements set by ISO.

IV. CONCLUSION

After a thorough analysis of the findings of the study, the succeeding conclusions are drawn:

The generated predictive models for diabetes and chronic kidney disease diagnoses and its implementation (as algorithms) in the eHealth application can address the barriers in delivering healthcare as evident on the summary of results of the system evaluation. It is an important real-world medical problem.

With the assistance of barangay health workers, the rural residents can receive expert diagnosis and treatment from distant medical centers provided that the internet connection is stable. Thus, it can lessen the cost and time in delivering health concerns.

In addition, health data constant monitoring and teleconsultation can also offer first-hand treatment for minor health problems since early stage diagnosis is the key for treatment.

V. RECOMMENDATION

Based on the results, findings and conclusions, the researcher hereby recommends the following:

1. The predictive models and the system itself should be validated by diabetes and kidney disease experts or doctors against actual medical cases to further affirm its reliability;
2. The eHealth application may be presented to the Department of Health (DoH) for its use.
3. Healthcare experts should facilitate in implementing this scheme in their respective assigned areas to trap possible error upon user's inputs;
4. Datasets from different locales should be utilized in the succeeding studies to further confirm the model's prediction's reliability and integrity.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

5. As further research, performance of other classifiers like ANN, Fuzzy logic or by using other data mining tools and platforms may be utilized for comparative study analysis.

REFERENCES

1. Eysenbach, G. (2018). Journal of Medical Internet Research Publications. Retrieved from http://www.jmir.org/2001/2/e20/
2. Biliran Provincial Office. (2016). Disaster Risk Reduction Climate Change Adaptation-Enhanced Provincial Development Physical Framework. Retrieved from Biliran Province Official Website: www.biliran.gov.ph/wp-content/uploads/pdf
3. Rashid, N. A., Husain, W., & Jothi, N. (2015). Data Mining in Healthcare – A Review. Procedia Computer Science, 306-313.
4. Saw, T. a., Brunner, M., Keep, M., Janssen, A., & Stewart, B. (2006). Development of a Conceptual Model for eHealth: Qualitative Study. Journal of Medical Internet Research.
5. Ting, S., C. Shum, C., K. Kwok, S., Tsang, A., & W.B., L. (2009). Data Mining in Biomedicine: Current Applications and Further Directions for Research. Journal in Software Engineering and Applications, 2(03):150-159.