A Detailed Analysis on Kidney and Heart Disease Prediction using Machine Learning

Claire Salkar
Engineering Informatica, Methodist University of Angola, Luanda, Angola
clairesalkar89@outlook.com

Article Info
Journal of Computing and Natural Science (http://anapub.co.ke/journals/jcns/jcns.html)
Doi: https://doi.org/10.53759/181X/JCNS202101003
Received 25 October 2020; Revised form 25 November 2020; Accepted 28 December 2020.
Available online 05 January 2021.

Abstract - Detection of disease at earlier stages is the most challenging one. Datasets of different diseases are available online with different number of features corresponding to a particular disease. Many dimensionalities reduction and feature extraction techniques are used nowadays to reduce the number of features in dataset and finding the most appropriate ones. This paper explores the difference in performance of different machine learning models using Principal Component Analysis dimensionality reduction technique on the datasets of Chronic kidney disease and Cardiovascular disease. Further, the authors apply Logistic Regression, K Nearest Neighbour, Naïve Bayes, Support Vector Machine and Random Forest Model on the datasets and compare the performance of the model with and without PCA. A key challenge in the field of data mining and machine learning is building accurate and computationally efficient classifiers for medical applications. With an accuracy of 100% in chronic kidney disease and 85% for heart disease, KNN classifier and logistic regression were revealed to be the most optimal method of predictions for kidney and heart disease respectively.

Keywords - Kidney Disease, Cardiovascular Disease, Logistic Regression, Support Vector Machine, Random Forest.

I. INTRODUCTION

Kidney and heart are the main organs in the human body and require extra care and attention to remain healthy. In this era of modernization where humans are exposed to polluted air, bad lifestyle, consumption of packaged food high in transfat, and more interaction with the electronic gadgets rather than family members, friends and relatives, the prevalence of chronic kidney disease and cardiovascular disease is increasing tremendously. According to a report released in 2019 by National Kidney Foundation (NKF), 10% of the population worldwide is affected by chronic kidney disease (CKD), and it is ranked as the 6th deadliest disease worldwide causing 2.4 million deaths per year. According to World Bank [2] four out of 5 cardiovascular diseases (CVD) related deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. To predict the occurrence of these diseases at earlier stages, machine learning techniques have proved to be a boon to medical practitioners [1].

Machine learning techniques have become a hotspot in biomedical and health-care research with the development of electronic health records and medical information. Using informatics technology in medical, healthcare can significantly alter and subvert the conventional healthcare and medical services [2]. In order to eventually enhance human health, new models and methods for early diagnosis of illness, care, and prevention are being conceived [3]. Machine learning is the modern bedrock of artificial intelligence. Through repetition and adjustment, it is capable of exploiting huge amounts of information and determine complicated patterns which are otherwise time-consuming for humans. A study claimed that the modern bedrock of artificial intelligence is machine learning that could predict the occurrence of heart attack with an accuracy of more than 90 percent. The study was presented at The International Conference on Nuclear Cardiology and Cardiac CT (ICNC) 2019. An algorithm 'learned' how imaging data interacts by analyzing 85 variables in 950 patients with known six-year outcomes repeatedly. Identification of patterns was then done correlating the variables to heart attack with an accuracy of more than 90 percent .The use of risk scores is done by doctors to make decisions during treatment. But these scores are based on a subset of variables and generally have moderate accuracy in individual patients [4].

Heart diseases have emerged as one of the most prominent causes of death around the world, claiming 17.7 million lives every year and constituting to 31% of global deaths [5] . Kidney disease is another major health problem contributing to high mortality. Over 1 million people in 112 lower income countries die from kidney failure [6]. Medical organizations around the world collect data on health-related issues to predict the occurrence of a disease

at an early stage. Machine learning is used as a useful tool in such predictions. Machine learning techniques deliver more precise and efficient predictions based on the hidden pattern of disease in patients.

A. Dimensionality Reduction Technique

Dimensionality means the total number of features of any dataset. Dimensionality reduction means projecting a higher dimensional data into lower dimensional space retaining all the valuable features and the variance among them. Many machine learning algorithms perform better in terms of time required in executing the algorithm and accuracy with which the algorithm performs with reduced dimensions. Due to higher dimensions, sometimes the models overfit which is not a good sign of accuracy. Therefore, dimensionality reduction is a very important step that needs to be considered when building a model. Dimension reduction is usually achieved in two ways: Feature Extraction and Feature Selection.

B. Importance of Principal Component Analysis in Health science

Health science can be defined as the study of different aspects of health, disease and healthcare with the aim to develop interventions, knowledge and technology for use in medical treatments. For this purpose, humans make use of healthcare records which exist in large quantities, consisting a large number of variables frequently correlating with each other. This correlation introduces the problem of multicollinearity in the regression model. Despite effective prediction, its presence makes it difficult for analysts to determine the precise effect of each predictor. Moreover, frequent errors can occur in the results. In order to overcome these effects, many statistical methods are developed and among them mostly popularly used method is Principal Component Analysis (PCA).

Developed by Karl Pearson in 1901, PCA is a statistical technique which can be used to convert the set of correlated variables into uncorrelated form for having the effective results determination. It finds applicability in highly correlated environments. It summarizes the patterns of correlation among observed variables in order to reduce the large number of variables into small number of factors. This makes the data more manageable. Further, modern medical science relies on sophisticated imaging techniques such as Positron Emission Tomography (PET), Computed Tomography Imaging (CTI), and Magnetic Resonance Imaging (MRI). These data need to be pre-processed, with elimination of noise, discarding of redundancies, and gathering of different sources. Here, PCA works as a computationally efficient and reliable technique for producing the images, making better diagnosis and prediction of certain diseases like cancer and heart disease. Hence, PCA analysis helps in adding more accuracy and reliability to the results of health data. PCA technique is applied in many machine learning algorithms such as random forests, support vector machine (SVM), logistic regression, decision trees, KNN and naïve bayes. For each of these algorithms, the application of PCA helps reduce dimensionality and predict the occurrence of a disease better.

Several studies have been conducted on the use of PCA in accurately predicting occurrence of heart and kidney diseases independently. However, they do not cohesively take into consideration kidney and heart disease dataset, which delivers a more accurate assessment regarding the effectiveness of PCA in enabling better prediction of these diseases. Therefore, the aim of this research is to analyze and evaluate the performance of machine learning models in the absence and presence of dimensionality reduction technique on the datasets of kidney and heart disease.

II. BACKGROUND

A. Predictive analysis of Chronic Kidney Disease and Heart Disease

The [7] with the aim of accurate prediction of chronic kidney disease progression over time examined the dataset of newly diagnosed patients for over 10 clinical years. Having the threshold value of 15cc/kg/min/1.73 m^2 of GFR (Glomerular filtration rate), the author used Takagi-Sugeno type adaptive neurofuzzy inference system (ANFIS) model to predict GFR values. Considering the variables like age, weight, sex, diastolic blood pressure, underlying disease, phosphorus, calcium, creatinine, GFR, and uric acid; the prediction model was formulated which despite high uncertainties of human body and dynamic nature of disease progression, predicted the accurate GFR variations.

A study on the dataset of 400 patients (250 CKD patients) wherein 24 attributes were considered for analysis [8]. Using the classification algorithm K-nearest neighbours, neural networks, and random forest; and feature reduction methods i.e. wrapper and LASSO regularization method, the analysis was done. Results of the analysis reveal that random forest algorithm with 12 attribute can detect CKD with accuracy of 99.8% using F1-measure model and 0.107 root mean square error.

The [9] used common spatial pattern and linear discriminant analysis for identifying the attributes that dominantly contribute in chronic kidney disease. Having the most accurate result determination by common spatial pattern, the analysis worked on classifying the dataset into non-CKD and CKD. The analysis revealed that specific gravity, albumin, hypertension, hemoglobin, and diabetes are the most important attributes of determining CKD.

The [10] performed classification of dataset into non-chronic and chronic kidney disease using support vector machine neural networks. The study of 400 patients revealed that about 94.44% of accuracy is derived by having 6 attributes. The work [11] focused on the chronic kidney disease prevalence in United States by using the predictive analytics techniques like Logistic Regression, Random trees, Artificial Neural networks, Chi-square automatic

interaction detector, support vector machines, and Naïve Bayes for predicting the chronic kidney disease. Herein, having the preprocessing of data and imputing missing data, the analysis shows that random trees provide most accurate information for CKD prevalence.

In [5] identified the appropriate diet plan for CKD patients by using classification algorithms. Using the predictive potassium zone, the experiment was performed via algorithms like multiclass decision forest, multiclass decision jungle, multiclass logistic regression, and multiclass neural network. The analysis with 99.17% accuracy revealed that multiclass decision forest is the most effective algorithm for identifying diet plan for CKD patients.

The [6] added more information to chronic kidney diseases by proposing a genetic algorithm trained neural network-based model (NN-GA) for assessing newest threats of this disease in underdeveloped and developing nations. This proposed model was compared to the multilayer perception feed forward network, random forest, and neural networks which revealed that in terms of accuracy, recall, F-measure, and precision NN-GA model is more efficient than other existing models.

The [7] studied 224 chronic kidney disease patients' records on UCI machine learning repository. Predicting the model based on deep neural network, the absence of presence of CKD could be predicted by 97% accuracy. Reducing the chances of overfitting, the built in model provided better results compared to any other algorithms.

] also worked with UCI machine learning repository wherein using the stacked autoencoder model and softmax In [2] classifier, the dataset of 400 patients and their 25 attributes has been examined. The analysis by the authors on the recall, precision, F1-score, and specificity revealed that with the accuracy of 100%, multimodal model has performed better than other conventional classifiers. The [1] for the CKD patients proposed a hybrid intelligent model by using logistic regression and neural network techniques. Having the accuracy of 97.8%, the analysis revealed that hybrid intelligent model was superior to models by 64%.

B. Logistic Regression

Logistic regression is a machine learning model targeting the binary classification. Working in similar form as linear regression, this method could be used for providing information about dataset and building in the relationship between one dependent variable and other variables. Predicting the probability, threshold value could be derived. Herein, considering t as linear function in univariate regression model, the logistic equation is expressed as per equation (1).

$$P = \frac{1}{1+ e^{-value\ of\ t}} \qquad (1)$$

The threshold value defined by the model could be affected by the precision or recall values.

C. K- Nearest Neighbor

The KNN method is the simplest algorithm for prediction using the Euclidean distance, thus also called lazy learning. K defines the neighbor number by considering nearest neighbor votes for prediction. The case wherein x and y define independent attribute, D defines their distance and K the nearest neighbor, output value could be predicted as per equation (2):

$$D = \sum_{i=1}^{k}(x_i - y_i)^2 \qquad (2)$$

Although the good value of K is selected via heuristic techniques, presence of irrelevant or noisy features could degrade the accuracy of this method.

D. Naïve Bayes

Naïve Bayes being a predictive machine learning algorithm is used as a method for regression and classification. Based on the bayes theorem, Guassian Naïve bayes is an extension to real-values attributes wherein working on Gaussian distribution predictions with the dataset is made. With requirement of small dataset for estimation and fast computation compared to other sophisticated methods, the Naïve bayes classifier works efficiently with real world situations.

E. Support Vector Machine

Support vector machine (SVM) is the algorithm used for classification. Herein, each feature data is plotted in n-dimensional space as a point and the n represents the features number. Using the hyperplane, the dataset is classified into different attributes. As SVM helps in separating non-separable classes from separable one by conversion of low dimensional input into high dimensional one, it eases the process of working even with complex dataset. However, as SVM is non-probabilistic model.

F. Random Forest

Being an ensemble learning algorithm, random forest (random decision forest) is used for handling classification and regression problem. Suitable for the dataset wherein decision tree is used for training, this method use bootstrap

aggregation technique. Thus, instead of having output derivation by just one decision tree, the method works with combined decision trees. As while splitting node, this method helps in best or most vital feature, thus good and effective results could be derived using this algorithm in data mining [1].

III. METHODOLOGY

Initially the dataset was collected from an open-source dataset platform. The initial step involved data pre-processing i.e. examination of the nature of data via mean and median in order to fill in missing data or have the normality assessment in dataset. Further, in the second phase the analysis commenced with data-classification. Herein, the entire dataset was randomly divided into two categories i.e. training data and testing data (80% and 20%). Then, a soft learning algorithm (Logistic Regression, KNN, Naïve Bayes, SVM, and Random forest) was built for the training data in order to have the identification of classification technique. Lastly, evaluating the accuracy, precision, recall, F1-measure, and support with the testing data based on the soft learning algorithm, the results were drawn. Hence, providing the final information about the performance of all models for the heart disease and kidney disease, the analysis helped address the aim of the research.

The dataset selected for the analysis consist of 500 chronic kidney disease patients' data and 5245 heart disease patients' data.

- Chronic Kidney Disease: Herein, data for the chronic kidney disease is collected from Mayo Clinic, Hospitals USA and the patient reflected with age, blood pressure, specific gravity, sugar, albumin, red blood cells, puss cells, pus cells clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, and class.

- Heart disease: For the heart disease, the database is of 5245 patients represent 16 attributes i.e. age, sex, education, current smoking status, cigarettes smoke per day, BP meds, prevalent smoke, prevalent Hyp, diabetes, total cholesterol, sys BP, diaBP, BMI, heart rate, glucose, and ten year CHD . Both the datasets are obtained from the Kaggle database.

IV. RESULT AND DISCUSSION

With the filtering of dataset and having the pre-processing to fill out the missing the data and understand the nature of dataset, final analysis was done to compare the efficiency of models and determine how the usage of principal component analysis affect the efficiency of model. Even the comparison of the soft-learning algorithm in the sub-section helped in determining the most efficient model that could predict the chronic kidney disease and heart disease. The analysis thus, based on the performance metrics is shown in below sub-sections.

A. Prediction model for chronic kidney disease

Table 1: Soft-Learning Algorithm assessment for chronic kidney disease

ML Models	Accuracy	Precision	Recall	F1-Score	Support
Results on Chronic Kidney Disease Dataset without PCA					
Logistic Regression	97.5	.97	.97	.97	80
KNN	99	.99	.99	.99	80
Naïve Bayes	99	.99	.99	.99	80
SVM	100	1	1	1	80
Random Forest	100	1	1	1	80
Results on Chronic Kidney Disease Dataset with PCA					
Logistic Regression	99	.99	.99	.99	80
KNN	100	1	1	1	80
Naïve Bayes	97	.98	.97	.97	80
SVM	99	.99	.99	.99	80
Random Forest	99	.99	.99	.99	80

Above PCA principal component analysis is 97.5% which is lowest compared to other models i.e. KNN has 99% accuracy, Naïve Bayes accuracy is also 99%, SVM has accuracy of 100% and the random forest accuracy level is 100%. Thus before PCA, SVM and Random forest has highest accuracy. With respect to the precision level, recall level, and F1-score too the situation is similar i.e. Logistic regression algorithm has least values with precision of

0.97, recall of 0.97, and F1-score of 0.97 while SVM and random forest has highest values with precision of 1, recall of 1, and F1-score of 1. Among all algorithms KNN and Naïve Bayes is at moderate level in their efficiency with precision, recall and F1-score of 99%. The support value for all the algorithms is same i.e. of 80.

With the usage of PCA for better results, the comparison of algorithms performance has changed. Though initially logistic regression was least effective and SVM and random forest provided more accurate results, after PCA all these three models have same performance level i.e. accuracy, precision, recall, and F1-score is 99%. Naïve Bayes has the least efficiency with accuracy, recall and F1-score of 97% and precision of 98% while KNN algorithm has the highest efficiency with accuracy, precision, recall and F1-score of 100%. Support value for all the algorithms are still same i.e. 80.

Hence, based on the examination of 425 chronic kidney patient's dataset, the analysis reveals that although SVM and random forest was most effective model before proper examination of dataset but after the Principal component analysis, effectiveness of the models has changed with KNN being the model having more accurate prediction about the chronic kidney disease.

B. Prediction model for heart disease

Table 2: Soft-Learning Algorithm assessment for heart disease

ML Models	Accuracy	Precision	Recall	F1-Score	Support
Results on Heart Dataset without PCA					
Logistic Regression	86	.85	.86	.80	750
KNN	85	.81	.85	.81	750
Naïve Bayes	84	.82	.84	.83	750
SVM	85	.72	.85	.78	750
Random Forest	85	.80	.85	.80	750
Results on Heart Dataset with PCA					
Logistic Regression	85	.83	.85	.80	750
KNN	84	.79	.84	.80	750
Naïve Bayes	85	.81	.85	.81	750
SVM	85	.72	.85	.78	750
Random Forest	85	.82	.85	.82	750

Table 2 shows the performance of each of learning algorithm before and after principal component analysis. Herein, the Naïve Bayes has the least level of accuracy i.e. 84% compared to other models wherein logistic regression has accuracy value of 86%, and random forest, SVM and KNN value for 85%. Precision and F1-score value for KNN is 81% and recall of 85%; for logistic regression precision level is 85%, recall is 86%, and F1-score is 80%; Naïve Bayes precision is 82%, recall is 84%, and F1-score of 83%; SVM model has precision of 72%, recall value of 85% and F1-score is 78%, and lastly the random forest precision value and F1-socre of 80% and recall of 85%. Thus, in terms of accuracy, precision, and recall Logistic regression have highest values while in terms F1-score Naïve Bayes has highest value. Support value for all the algorithms is same i.e. 750.

With PCA, there has been variation in the performance of each of the models and more appropriate results could be derived in terms of all performance metrics i.e. accuracy, precision, recall, and F1-score. Initially accuracy, precision, and recall was highest for logistic regression while Naïve Bayes was more optimal in terms of F1-score, and even after application of PCA Table 2 reveal that Logistic regression is the optimal model for prediction as herein accuracy and recall value is 85% which is highest among all models, and the precision is 83% while F1-score is only 80%. Further, the analysis reveals that KNN is least effective as the accuracy and recall level is 84%, precision is 79%, and the F1-score is 80%. For other models i.e. Naïve Bayes accuracy and recall is 85%, and precision and F1-score is 81%; SVM model shows accuracy and recall value of 85%, precision 72%, and F1-score of 78%; and lastly random forest model accuracy and recall value is 85% with precision and F1-score of 82%. Support value of all algorithms is still same as 750.

V. CONCLUSION

The proposed work implements different techniques viz. LR, KNN, Naïve Bayes, SVM and RF Classifier, which can be used to predict the possibility of occurrence of kidney or heart disease. The work also evaluates the accuracy of above classification techniques without and with using Principal Component Analysis, a dimensionality reduction technique. For lower dimensional data sets, it is observed that similar level of accuracy can be achieved without using PCA but the model over fitting can be a matter of concern in those cases. It is also concluded that KNN proves

to be a better classifier for kidney disease and logistic regression for heart disease datasets. KNN gives an accuracy of 100% with PCA for kidney disease on a dataset size of 425 patients which also overcomes the over fitting dilemma, and logistic regression accuracy of 85% for heart disease on a dataset of 5245 patients which is best among all the classifiers chosen.

References

[1]. M. E. Farooqui and D. J. Ahmad, "A Detailed Review On Disease Prediction Models That Uses Machine Learning," International Journal of Innovative Research in Computer Science & Technology, vol. 8, no. 4, Jul. 2020.

[2]. P. M. M. Bhajibhakare, "Heart Disease Prediction using Machine Learning," International Journal for Research in Applied Science and Engineering Technology, vol. 7, no. 12, pp. 455–460, Dec. 2019.

[3]. S. M. Pasha, "Diabetes and Heart Disease Prediction Using Machine Learning Algorithms," International Journal of Emerging Trends in Engineering Research, vol. 8, no. 7, pp. 3247–3252, Jul. 2020.

[4]. Dokare, A. Prithiani, H. Ochani, S. Kanjan, and D. Tarachandani, "Prediction of Having a Heart Disease Using Machine Learning," SSRN Electronic Journal, 2020.

[5]. Singh, "Prediction of Heart Disease using Machine Learning," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 150–166, Jul. 2020.

[6]. K. K. Y., Dr. Anurag Sharma, Dr. Abhishek Badholia, "Heart Disease Prediction Using Machine Learning Techniques," Information Technology In Industry, vol. 9, no. 1, pp. 207–214, Feb. 2021.

[7]. Dokare, A. Prithiani, H. Ochani, S. Kanjan, and D. Tarachandani, "Prediction of Having a Heart Disease Using Machine Learning," SSRN Electronic Journal, 2020.

[8]. P. K. Sahoo and P. Jeripothula, "Heart Failure Prediction Using Machine Learning Techniques," SSRN Electronic Journal, 2020.

[9]. K. Howe, N. James, P. Gladding, C. Prabhakar, A. Gavin, and L. Dawson, "Predicting CRT Response Using Machine Learning Analysis of Pre-Implant ECG Data," Heart, Lung and Circulation, vol. 26, p. S188, 2017.

[10]. R. Keniya, A. Khakharia, V. Shah, V. Gada, R. Manjalkar, T. Thaker, M. Warang, and N. Mehendale, "Disease Prediction From Various Symptoms Using Machine Learning," SSRN Electronic Journal, 2020

[11]. M. S. Devi, P. Swathi, S. S. Upadhyay, N. K. Sah, A. Budhia, vamsi chowdary, S. Srivastava, and M. Rohella, "Feature Predominance Ensemble Inquisition towards Liver Disease Prediction using Machine Learning," SSRN Electronic Journal, 2021.