To improve Data Mining based Human Disease Detection System

Sanskriti Pandey
Electronics and Telecommunication Department, Thakur College of Engineering and Technology, Mumbai University

Pranit Oza
Electronics and Telecommunication Department, Thakur College of Engineering and Technology, Mumbai University

Divesh Gupta
Electronics and Telecommunication Department, Thakur College of Engineering and Technology, Mumbai University

Niket Amoda
Electronics and Telecommunication Department, Thakur College of Engineering and Technology, Mumbai University

Abstract- Analytics based on predictions for healthcare with the use of data-mining is challenging job in helping doctors evaluate the appropriate treatments for retaining lives. In this project, we present data-mining methods for prediction of the chronic kidney disease by use of clinical data and choose the one having highest accuracy while providing interface for user. Different data-mining techniques are evaluated including Neural Network (NN), support vector machine (SVM) and Random Forest (RF). These models that are predictive in nature have been constructed earlier using chronic renal disease dataset and their performances have been compared in order to choose the best classifier for the prediction of the chronic disease. The classifier having the best efficiency is selected after various comparisons. It is implemented and a user interface is provided for the user for entering their medical data. Output predicting the occurrence of chronic kidney disease is obtained.

Keywords- analytics, classifier, user-interface

I. INTRODUCTION

Towards the lower back of a human's abdomen there are pair of organs known as kidneys positioned. It removes toxins by the use of bladders through urination and thus helps in purification of the bloodstream. However, kidneys are sometimes incapable of filtering wastes which leads to our body becoming clogged with toxins, causing failure of the kidney and consequently leading to demise of the individual. Problems of the kidney can be broken down into either acute or chronic. Chronic kidney disease consists of situations that harm the kidneys and decrease its ability of keeping us perfectly fit. If the renal disease worsens, wastes might mount up so much so as to build enhanced waste levels in the blood and cause severe complications like anaemia Also, kidney diseases increase the probabilities of suffering from a heart as well as blood vessel disease. Chronic kidney disease might be caused due to diabetes, high blood pressure, hypertension, Coronary artery diseases, lupus, albumin in urine, or even several unwanted situations from some medications, and can even be hereditary and many such issues. Early realization and treatment might help the prevention of chronic kidney disease. When the renal disease advances, it might cause kidney failure, which might require for the retainment of life, a kidney transplant or dialysis. Data-mining techniques have proved successful in predicting and performing diagnosis of various diseases. Human experts are restricted to a certain limit in finding hidden pattern from data. As a result, the alternative is to make use of methods that are computational to mine the data in the correct manner. [5]

II. PROBLEM DEFINITION

Currently, the maintenance of databases from the clinics has become an important task in the medical field. The patient data which consists of various features as well as diagnostics related to disease must be entered with utmost care to provide good quality services.
The data stored in medical databases might contain missing values and thus, redundant data-mining of the medical data becomes cumbersome.

III. NEED OF PROJECT

Chronic Kidney Disease prediction is one amongst most central problems in medical decision making because it is one of the leading causes of death. Early detection of this chronic disease helps in taking appropriate preventive actions and treatment that is effective at an initial stage has found to be helpful for patients. Machine learning techniques can help and provide prediction capabilities to handle these circumstances. [4]

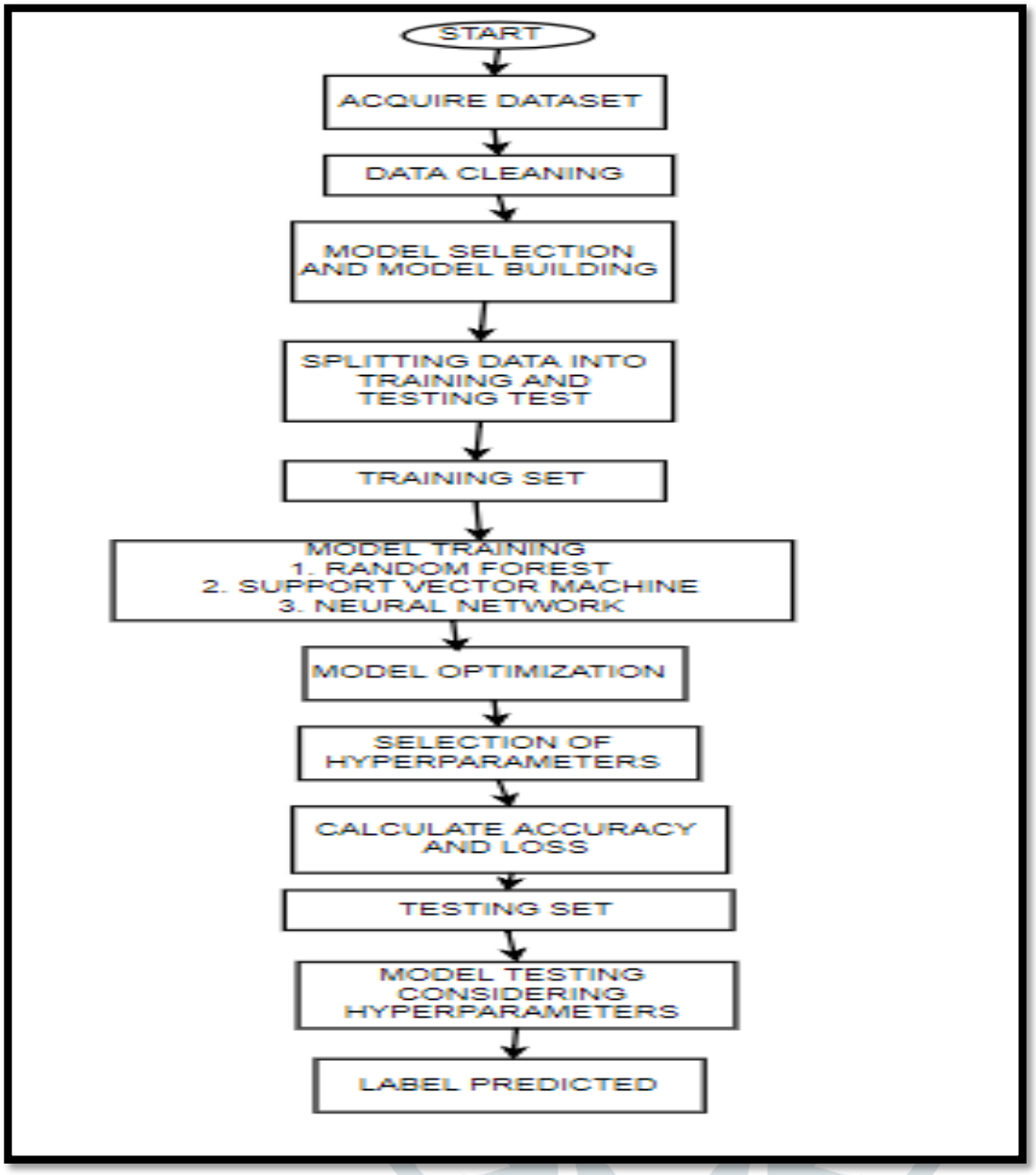
IV. PROPOSED WORK

Initially, study is done so as to how the current chronic renal disease prediction system functions. Then, planning is done to understand the improvements to be made to the current chronic kidney disease prediction system. After that is done, study of the different data mining classifiers such as SVM, Neural Network and Random Forest is completed. Once that is done successfully, finding accuracy of the different data mining classifiers plays a crucial role. After the accuracy determination of each classifier is done, selection of the ones with high accuracies takes place. Then there is a need to find the classifier with the highest accuracy. The classifier should be such that it reduces the feature dimensionalities simultaneously. Then this classifier is selected and implemented in our system. After that the input is given to the interface involving of various symptoms. Then, checking of the output resulting in prediction of a Chronic Kidney Disease takes place.

Classifier	Settings	Accuracy (%)
SVM	Complex Factor:1.5 Kernel: Polynomial Kerner	99.685
KNN	K:3	99.58
Decision Trees	seed: 1 confidence factor:0.25 number of folds:3	98.6
Naive Bayes	no settings	98.5
Random Forest	seed:1 number of trees:10	97.218

Method	overall accuracy (%)	kappa coefficient (κ)
Maximum likelihood (ML)	64.6	0.57
Minimum volume ellipsoid (MVE)	58.2	0.50
Naïve Bayes (NB)	67.2	0.60
Support vector machine (SVM)	75.0	0.68
Artificial neural networks (ANN)	71.7	0.64
Random forest (RF)	72.6	0.66
Nearest neighbour (NN)	65.4	0.57

V. PROPOSED METHODOLOGY



VI. SOFTWARES USED

Python: Python is a programming language which allows the same code to run on various platforms or Operating Systems without any need for recompilation. The main advantage of Python over other programming languages is its large and robust standard library. [1]

Data-mining techniques: It plays an important role in working on large datasets. It helps in analyzing data and converting the data into information which is meaningful. This helps in efficient retrieval of information and finding important solutions. [2]

VII. DATASET FOR PREDICTION

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Id	Age	Blood Pressure	Specific Gravity	Albumin	Sugar	Red Blood Cells	Pus Cell	Pus Cell 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	Class
2	1	48	70	1.005	4	0	normal	abnorma	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	1
3	2	53	90	1.02	2	0	abnormal	abnorma	present	notpresent	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	1
4	3	63	70	1.01	3	0	abnormal	abnorma	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	1
5	4	68	80	1.01	3	2	normal	abnorma	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	1
6	5	61	80	1.015	2	0	abnormal	abnorma	notpresent	notpresent	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	1
7	6	48	80	1.025	4	0	normal	abnorma	notpresent	notpresent	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	1
8	7	69	70	1.01	3	4	normal	abnorma	notpresent	notpresent	264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes	yes	good	yes	no	1
9	8	73	70	1.005	0	0	normal	normal	notpresent	notpresent	70	32	0.9	125	4	10	29	18900	3.5	yes	yes	no	good	yes	no	1
10	9	73	80	1.02	2	0	abnormal	abnorma	notpresent	notpresent	253	142	4.6	138	5.8	10.5	33	7200	4.3	yes	yes	yes	good	no	no	1
11	10	46	60	1.01	1	0	normal	normal	notpresent	notpresent	163	92	3.3	141	4	9.8	28	14600	3.2	yes	yes	no	good	no	no	1
12	11	56	90	1.015	2	0	abnormal	abnorma	notpresent	notpresent	129	107	6.7	131	4.8	9.1	29	6400	3.4	yes	no	no	good	no	no	1
13	12	48	80	1.005	4	0	abnormal	abnorma	notpresent	present	133	139	8.5	132	5.5	10.3	36	6200	4	no	yes	no	good	yes	no	1
14	13	59	70	1.01	3	0	normal	abnorma	notpresent	notpresent	76	186	15	135	7.6	7.1	22	3800	2.1	yes	no	no	poor	yes	yes	1
15	14	63	100	1.01	2	2	normal	normal	notpresent	present	280	35	3.2	143	3.5	13	40	9800	4.2	yes	no	yes	good	no	no	1
16	15	56	70	1.015	4	1	abnormal	normal	notpresent	notpresent	210	26	1.7	136	3.8	16.1	52	12500	5.6	no	no	no	good	no	no	1
17	16	71	70	1.01	3	0	normal	abnorma	present	present	219	82	3.6	133	4.4	10.4	33	5600	3.6	yes	yes	yes	good	no	no	1
18	17	73	100	1.01	3	2	abnormal	abnorma	present	notpresent	295	90	5.6	140	2.9	9.2	30	7000	3.2	yes	yes	yes	poor	no	no	1
19	18	71	60	1.015	4	0	normal	normal	notpresent	notpresent	118	125	5.3	136	4.9	11.4	35	15200	4.3	yes	yes	no	poor	yes	no	1
20	19	52	90	1.015	4	3	normal	abnorma	notpresent	notpresent	224	166	5.6	133	47	8.1	23	5000	2.9	yes	yes	no	good	no	yes	1
21	20	50	90	1.01	2	0	normal	abnorma	present	present	138	208	9.2	134	4.8	8.2	22	16300	2.7	no	no	no	poor	yes	yes	1
22	21	70	100	1.015	4	0	normal	normal	notpresent	notpresent	118	125	5.3	136	4.9	12	37	8400	8	yes	no	no	good	no	no	1
23	22	60	90	1.01	2	0	abnormal	normal	notpresent	notpresent	105	53	2.3	136	5.2	11.1	33	10500	4.1	no	no	no	good	no	no	1
24	23	60	60	1.01	3	1	normal	abnorma	present	notpresent	288	36	1.7	130	3	7.9	25	15200	3	yes	no	no	poor	no	yes	1
25	24	55	90	1.01	2	1	abnormal	abnorma	notpresent	notpresent	273	235	14.2	132	3.4	8.3	22	14600	2.9	yes	yes	no	poor	yes	yes	1
26	25	62	70	1.025	3	0	normal	abnorma	notpresent	notpresent	122	42	1.7	136	4.7	12.6	39	7900	3.9	yes	yes	no	good	no	no	1
27	26	59	80	1.01	1	0	abnormal	normal	notpresent	notpresent	303	35	1.3	122	3.5	10.4	35	10900	4.3	no	yes	no	poor	no	no	1

VIII. EXPECTED OUTCOMES

A lot of economical, emotional and physical burden is faced by patients all over the world.
A system like the one proposed by this project helps to eradicate such burdens.
Knowing anything beforehand always helps in facing and dealing with a situation in a better and efficient manner. Predicting the occurrence of a chronic disease of the kidney will help save lives of people all the world.
It will prove fruitful for the people belonging to a backward economic class as it will help save the money which otherwise is used for a lot of health-related tests and also during consultancy.
It will prove to be an interface which is readily available and easily accessible.

IX. CONCLUSION

We have done the literature survey on predictive models by using machine learning methods including Support vector machine (SVM), Random Forest (RF), and Neural Network (NN) classifiers to predict chronic kidney disease. From the experimental results, it can be seen that SVM classifier gives the highest accuracy. In addition, SVM has highest sensitivity after training and testing by the proposed method. Thus, we have concluded that SVM classifier is appropriate for predicting the chronic kidney disease. We have then provided an interface for the user to enter their input and the possibility of Chronic Kidney Disease is predicted. [3]

X. REFERENCES

[1]https://medium.com/@mindfiresolutions.usa/python-7- important-reasons-why-you-should-use-python-5801a98a0d0b.

[2]M. Thangamani, P. Thangaraj, Bannari, “Automatic medical disease treatment system using datamining,” International Conference on Information Communication and Embedded Systems, ICICES, 2013.

[3]Guneet Kaur, Ajay Sharma “Predict chronic kidney disease using data mining algorithms in hadoop” International Conference on Inventive Computing and Informatics, IEEE, 2017.

[4] https://en.wikipedia.org/wiki/Data_mining

[5]https://www.researchgate.net/figure/Classifiers-that-achieved-the-highest-accuracy-scores-along-with-their-corresponding_tbl1_272297364