WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predictive Analysis on Diabetes, Liver and Kidney Diseases using Machine Learning

Hritik Bharadwaj[1], Rudraksh[2],Shubham Tyagi[3], Anil Gankotia[4]

[1,2,3,]Student (Final Year), Department of Computer Engineering,Raj kumar goel Institute of Technology and Management,Ghaziabad,Uttar Pradesh

[4,]Professor, Department of Computer Engineering,,Raj kumar goel Institute of Technology and Management,Ghaziabad,Uttar Pradesh

A B S T R A C T

Predictive analytics for healthcare using machine learning is a challenged task to help doctors decide the exact treatments for saving lives. Diabetes, Liver Disease and Chronic Kidney Disease combinedly affects a mass of world's population. The objective of this briefing is to develop an efficient decision support system to predict the possibility of a disease using the techniques of Machine Learning. Machine Learning is used to discover patterns in the data, detect and then make predictions with the help of algorithms. It provides methods, techniques and tools that can help in solving diagnostic problems in a variety of medical domains e.g., prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and for the overall patient management. It offers a principled approach for developing sophisticated, automatic, and objective algorithms for biomedical data. This paper mainly focuses on diagnostically predict the possibility of mainly three diseases: Diabetes, Liver Disease and Chronic Kidney Disease.

Keywords: Machine Learning (ML), Naïve Bayes Classification, Kernel Support Vector Classification, Decision Tree Classifier, Random ForestClassification, Logistic Regression Model, XGBoost Classifier, Diabetes, Liver Disease, Chronic Kidney Disease

1. INTRODUCTION

Diabetes, Liver Disease and Chronic Kidney Disease are affecting millions of people worldwide. But with early diagnosis and treatment, it's possible to slow or stop the progression of these diseases.

Diabetes mellitus, commonly known as diabetes, is a metabolic disease that causes high blood sugar. The hormone insulin moves sugar from the blood into your cells to be stored or used for energy. With diabetes, your body either doesn't make enough insulin or can't effectively use the insulin it does make.The global diabetes prevalence in 2020 is estimated to be 9.3% (463 million people), rising to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. One in two (50.1%) people living with diabetes do not know that they have diabetes Just under half a billion people are living with diabetes worldwide and the number is projected to increase by 25% in 2030 and 51% in 2045. [1]

The liver plays an important role in many bodily functions from protein production and blood clotting to cholestrol, glucose, and iron metabolism.Liver disease accounts for approximately 2 million deaths per year worldwide. Global prevalence of liver disease from autopsy studies ranges from 4.5% to 9.5% of the general population. Hence, it can be said that more than fifty million people in the world, taking the adult population, would be affected with chronic liver disease. [2]

Chronic kidney diseases (CKDs) are the most common forms of kidney disease all around the world. The incidence of CKD is rising is mainly driven by population aging as well as by a global rise in hypertension, metabolic syndrome, and metabolic risk factors. Chronic kidney disease is a worldwide health crisis. 10% of the population worldwide is affected by chronic kidney disease (CKD), and millions die each year because they do not have access to affordable treatment. [3]

* *Corresponding author.*
Rudrakshsaxena328@gmail.com

The objective of this study is to assess the efficiency and accuracy of the used ML models for the prediction of these three diseases. This study focuses on structured data and for the best possible output classification and prediction, advanced machine learning algorithms like Random Forest, Decision Tree, Naive Bayes Classification, XG Boost Classifier, Kernel Support Vector Classification and Logistic Regression has been used.

2. PROPOSED SYSTEM

2.1 MACHINE LEARNING

Machine learning is a subset of AI. It gives computers the ability to learn—usually by providing statistical data—without being programmed every step of the way.Machine learning is a method of data analysis that automates analytical model building. formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. Machine learning is sub-categorized to three types:

A. ***Supervised Learning***
In supervised learning, the system must "learn" inductively a function called target function, which is an expression of a model describing the data. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data In this paper, prediction using supervised learning models has been done.

B. ***Unsupervised Learning***
In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels.

C. ***Reinforcement Learning***
In Reinforcement Learning, system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward.

.

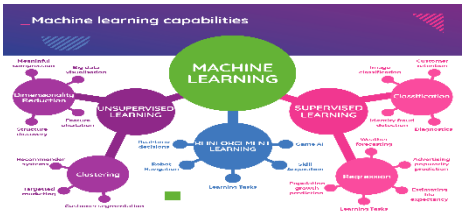


Figure 1 Approaches of different type of Machine Learning

.

3. MODEL ARCHITECTURE

To find hidden insights without needing explicit programming, Machine learning uses algorithms which learn from previous data to help produce reliable and repeatable decisions. Machine Learning at its core is just is a collection of algorithms. ML Algorithms leverage knowledge of statistics, probability, calculus, vector algebra, matrices, optimization techniques etc. Six classification algorithms has been used it this system which are as follows:

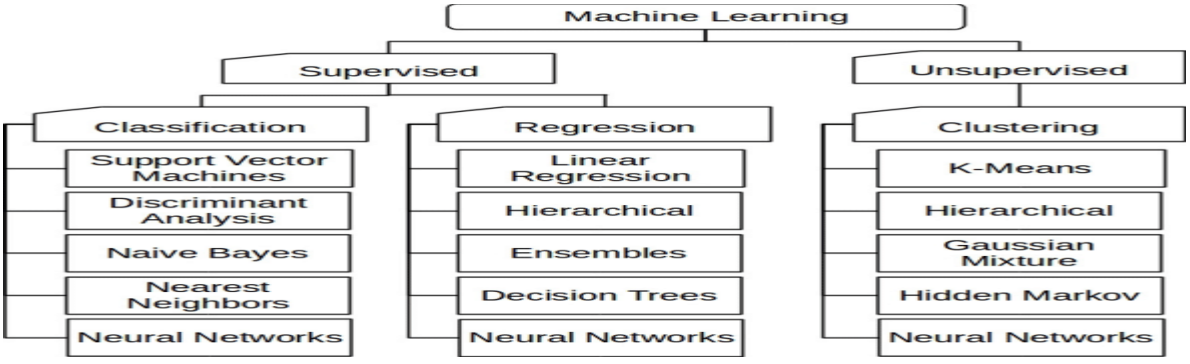


Figure 2.Different Algorithm used in this Project

3.1 Naïve Bayes Classifier

In statistics, naive Bayes classifiers are a family of simple "probabilistic classifier" based on applying Bayes theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network model, but coupled with kernel density estimation, they can achieve higher accuracy levels.The Naïve Bayesian classifier is based on Bayes' theorem with independence assumptions between

predictors. This model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Bayes theorem provides a way of calculating the posterior probability P(C/X) of class from P(C) is the prior probability of class.

3.2 Kernel Support Vector Classification

SVM are powerful yet flexible supervised machine learning algorithms. Generally, it is considered to be a classification approach, but it can be used both for classification and regression process. SVM is able to handle multiple continuous and categorical variables. SVM constructs a decision boundary or hyperplane that divide the datasets into classes to find a Maximum Marginal Hyperplane (MMH), where we can easily put the new data point in the correct category in the future."Kernel" is used due to set of mathematical functions used in Support Vector Machine provides the **window to manipulate the data**. So, Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transformed to a linear equation in a higher number of dimension spaces.

3.3 Decision Tree Classifier

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructtree like structure.

3.4 Logistic Regression Model

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1". Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function. Logistic Regression makes use of sigmoid function which takes solution of linear regression and output value between 0 and 1. It has Sshaped curve known as logistic curve.

3.5 XGBoost Classifier

It is short for extreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework by (Friedman, 2001) (Friedman et al., 2000). The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking. The package is made to be extendible, so that users are also allowed to define their own objectives easily

4. IMPLEMENTATION

The implementation phase has various steps of machine learning and the flow of implementation is shown

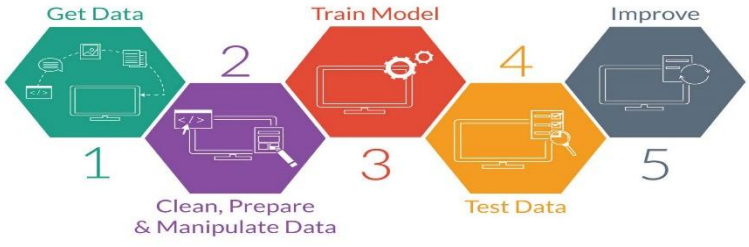


Figure 3. Implementation phases of machine learning

4.1 Data Collection All the datasets have been collected from Kaggle community

1) Diabetes:

The dataset includes diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetesaccording to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dlat any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona,USA.

	Number of times pregnant	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Diastolic blood pressure (mm Hg)	Triceps skin fold thickness (mm)	2-Hour serum insulin (mu U/ml)	Body mass index (weight in kg/(height in m)^2)	Diabetes pedigree function	Age (years)	Class variable (0 or 1)
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 4.Diabetes patients Dataset

2) Liver Disease

The data set has been elicited from UCI Machine Learning Repository. This data set contains 416 liver patientrecords and 167 non liver patient records. The data set was collected from test samples in North East of AndhraPradesh, India.

	age	gender	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	is_patient
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1

Figure 5. Liver Patient Dataset

3) Chronic Kidney Disease

This dataset is from UCI Machine Learning Repository. The objective of the dataset is todiagnostically predict whether a patient having chronic kidney disease or not, based on certaindiagnosticmeasurements included in the dataset.

	Bp	Sg	Al	Su	Rbc	Bu	Sc	Sod	Pot	Hemo	Wbcc	Rbcc	Htn	Class
0	80.0	1.020	1.0	0.0	1.0	36.0	1.2	137.53	4.63	15.4	7800.0	5.20	1.0	1
1	50.0	1.020	4.0	0.0	1.0	18.0	0.8	137.53	4.63	11.3	6000.0	4.71	0.0	1
2	80.0	1.010	2.0	3.0	1.0	53.0	1.8	137.53	4.63	9.6	7500.0	4.71	0.0	1
3	70.0	1.005	4.0	0.0	1.0	56.0	3.8	111.00	2.50	11.2	6700.0	3.90	1.0	1
4	80.0	1.010	2.0	0.0	1.0	26.0	1.4	137.53	4.63	11.6	7300.0	4.60	0.0	1

Figure 6. Chronic Kidney Disease Patient Dataset

4.2 Data Preparation

Data preparation may be one of the most difficult steps in any machine learning project. The reason is that each dataset is different and highly specific to the project. Nevertheless, there are enough commonalities across predictive modelling projects that we can define a loose sequence of steps and subtasks that you are likely to perform. As this paper deals with medical data, any changes have not been made in order to preserve the patient's data. Data correlation is the way in which one set of data may correspond or relate to another set. The set of correlation values between pairs of its attributes form a matrix which is called a correlation matrix.

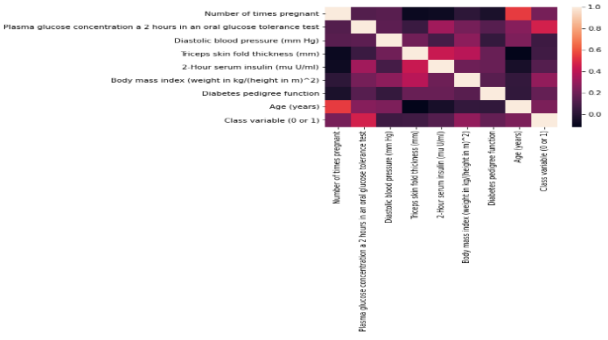
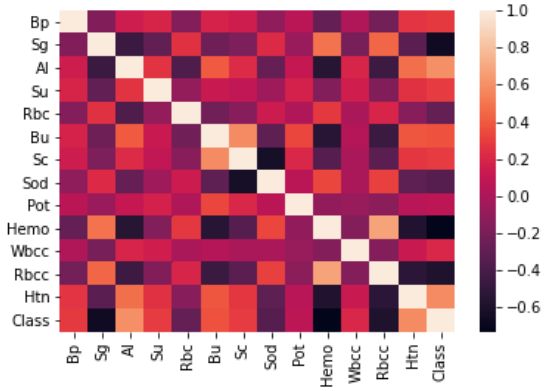


Figure 7 Heatmap Of Diabets



/

Figure 8. Corelation Heatmap of Liver

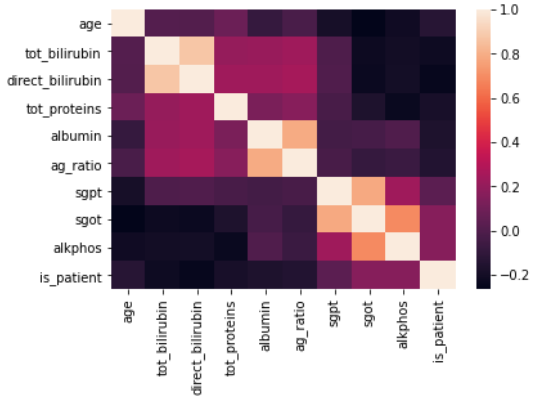


Figure 9. Corelation Heatmap of chronic kidney disease

4.3 Training Model

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output.The datasets are trained using different machine learning algorithms. Every algorithm has different working procedures resulting in varying accuracy. Training is basically the process of giving the machine capability to make further predictions after learning from the training dataset.

```
# XG Boost
from xgboost import XGBClassifier
xgboost_model = XGBClassifier()
xgboost_model.fit(x_train, y_train)
y_pred_xg = xgboost_model.predict(x_test)

xgboost_model.score(x_test , y_pred_xg)*100
```

```
#Logistic Regression Model

from sklearn.linear_model import LogisticRegression
log_model = LogisticRegression()
log_model.fit(x_train, y_train)
y_pred_log = log_model.predict(x_test)
```

```
#Naive Bayes Classification

from sklearn.naive_bayes import GaussianNB
classifier_naive = GaussianNB()
classifier_naive.fit(x_train, y_train)
y_naive_pred = classifier_naive.predict(x_test)
```

```
#Decision Tree Classifier

from sklearn.tree import DecisionTreeClassifier
classifier_dtc = DecisionTreeClassifier()
classifier_dtc.fit(x_train,y_train)
y_dtc_pred = classifier_dtc.predict(x_test)
```

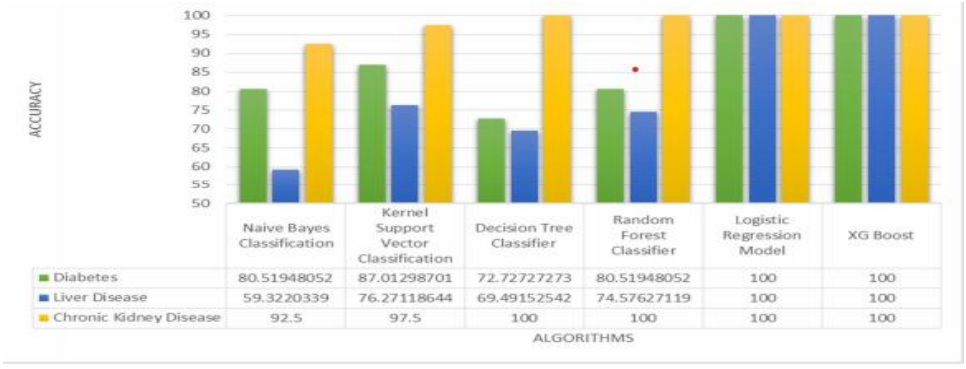
```
#Random Forest Classifier

from sklearn.ensemble import RandomForestClassifier

classifier_rfc = RandomForestClassifier(n_estimators=20, random_state=0)
classifier_rfc.fit(x_train, y_train)
y_rfc_pred = classifier_rfc.predict(x_test)
```

```
#Kernel Support Vector Classification

from sklearn.svm import SVC

classifier_svc = SVC(kernel='linear',random_state=0)
classifier_svc.fit(x_train,y_train)
y_svc_pred = classifier_svc.predict(x_test)
```

Figure 10. Training datasets using different algorithms

4.4 Testing Model

In machine learning, model testing is referred to as the process where the performance of a fully trained model is evaluated on a testing set.Testing of a model is done to check the performance of the algorithms in term of accuracy, precision etc. In testing whether the prediction is correct or not is checked using already predefined dataset. Higher the accuracy, better the results.



	Naive Bayes Classification	Kernel Support Vector Classification	Decision Tree Classifier	Random Forest Classifier	Logistic Regression Model	XG Boost
Diabetes	80.51948052	87.01298701	72.72727273	80.51948052	100	100
Liver Disease	59.3220339	76.27118644	69.49152542	74.57627119	100	100
Chronic Kidney Disease	92.5	97.5	100	100	100	100

Figure 11. Algorithms with respect to their accuracy scores in respective disease

4.5 Prediction

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome.Prediction Explanations avoid the "black box" syndrome by describing which characteristics, or feature variables, have the greatest impact on a model's outcomes.The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be.

	Class	Predicted Class
0	1	1
1	1	1
2	1	1
3	1	1
4	1	1

Figure 12. Actual vs. Predicted values of Diabetes patient's dataset

	is_patient	Predicted is_patient
0	1	1
1	1	1
2	1	1
3	1	1
4	1	1
...	...	...

Figure 13. Actual vs. Predicted values of Liver patient's dataset

5. CONCLUSION

Machine learning has transpire as a field demanding for providing tools and methodologies for analyzing the datagenerated by the biomedical sciences. This review has provided a precipitate snapshot of applications of machinelearning for the detection of three diseases: Diabetes, Liver Disease and Chronic Kidney Disease using differentclassifiers of supervised machine learning.Fusion of contrasting multimodal and multi-scale biomedical data continues to be a challenge. Furtherimprovements in data can be made like having more features and least null values. Moreover, considerableimprovements in Python-based workflow can also be done. We have also made up a user interface for this researchwhich can be considered as a prototype.

6. FUTURE SCOPE

Machine learning includes a number of algorithms and techniques to examine and contraption to gain the benefits ofthem in different fields including healthcare. ML methods can help the amalgamation of computer-based systems inthe healthcare environment administer opportunities to expedite and enhance the work of medical experts andconclusively to improve the efficiency and quality of medical care. ML technologies can be used to diagnosepotential clinical trial candidates, access their medical history records, monitor the candidates throughout the trialprocess, select best testing samples, reduce data-based errors, and much more.In future with respect to this model, one can try to develop a system in which most plausible disease for a patient canbe predicted on the basis of symptoms and moreover test can also be advocated for the predicted disease.A potentialfuture development of the presented work is to apply ML models to other data with different features, concerningthe survival prognosis of the patients and early detection of the disease and it can also be developed in web-basedapplication with additional services.

.

REFERENCES

[1] D. Lowd, P. Domingos, “Naïve Bayes Models for Probability Estimation”.
[2] S. Kanchana, “Statistical Analysis Using Machine Learning Approach for Multiple Imputation of Missing Data”, in IJRASET, vol.6, February 2018, p.2091
[3] A Comparative Study of Machine Learning Classifiers for Medical Diagnosis, Bhavnath Thakur, Harshit Rohela, Kanishk Gupta, Chhaya Sharma. "A
[4] Comparative Study of Machine Learning Classifiers for Medical Diagnosis", Volume 8, Issue IV, International Journal for Research in Applied Science andEngineering Technology (IJRASET) Page No: 1748-1752, ISSN: 2321-9653.
[5] S.B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques”, Informatica 31 (2007) 249-268.
[6] T. Chen and T. He, “XGBoost: extreme gradient boosting”, R Package. Version 0.4-2, 2015.
[7][DataRobot] [Wiki][Prediction] https://www.datarobot.com/wiki/prediction/.
[8] [Upgrad] [Blog][Machine Learning Applications in Healthcare] https://www.upgrad.com/blog/machine-learning-applications-in-healthcare/.