Analysis on Healthcare Datasets using Machine Learning Algorithm

The project report submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

In

Computer Science and Engineering

By

Ishant Goyal(181500282), Deepak Awasthi(181500202)
Rajan Sharma(181500543)

Under the Guidance of
Dr.Saroj Kumar Pandey
(Assistant Professor)

Department of Computer Engineering & Applications
Institute of Engineering & Technology

GLA UNIVERSITY

MATHURA- 281406, INDIA

07 / May / 2022

Department of Computer Engineering and Applications
GLA University, 17 km Stone, NH#2, Mathura-Delhi Road, P.O. Chaumuhan, Mathura-281406 (U.P.)

DECLARATION

I hereby declare that the work which is being presented in the B.Tech. Project **"Analysis on healthcare datasets using Machine learning algorithm "**, in partial fulfillment of the requirements for the award of the ***Bachelor of Technology*** in Computer Science and Engineering and submitted to the Department of Computer Engineering and Applications of GLA University, Mathura, is an authentic record of my own work carried under the supervision of **Dr.Saroj Kumar Pandey (Assistant professor)**

The contents of this project report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree.

Sign _____

Name of Candidate: Rajan Sharma
University Roll No.: 181500543

Sign _____

Name of Candidate:Deepak Awasthi
University Roll No.:181500202

Sign _____

Name of Candidate: Ishant Goyal
University Roll No.: 181500282

CERTIFICATE

This is to certify that the above statements made by the candidate are correct to the best of my/our knowledge and belief.

_____

Supervisor
(Dr.SarojKumar Pandey)
Designation of Supervisor
Dept. of Computer Engg, & App.

_____ _____

Project Co-ordinator
(Dr. Mayank Srivastava)
Associate Professor
Dept. of Computer Engg, & App.
Date:

Program Co-ordinator
(**Dr. Rakesh Kumar Galav**)
Assistant Professor
Dept. of Computer Engg, App.

ACKNOWLEDGEMENT

We would like to express my sincere gratitude to several individuals and organizations for supporting us throughout my Graduate study. First, we wish to express my sincere gratitude to my supervisor, Professor Collins, for his enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my research and writing of this thesis. His immense knowledge, profound experience and professional expertise in Machine learning has enabled me to complete this research successfully. Without his support and guidance, this project would not have been possible. We could not have imagined having a better supervisor in my study.

We also wish to express my sincere thanks to the GLA UNIVERSITY for accepting me into the graduate program. In addition.

Finally, last but by no means least; also to everyone in the GLA University it was great sharing premises with all of you during last four years.

Thanks for all your encouragement!

Sign _____
Name of Candidate:Deepak Awasthi
University Roll No.:181500202

Sign _____
Name of Candidate:Rajan Sharma
University Roll No.:181500543

Sign _____
Name of Candidate:Ishant Goyal
University Roll No.:181500282

ABSTRACT

Nowdays, Machine learning algorithms (CAD System) construct a remarkable contribution to make computer add diagnosis system .The generic purpose of this work is to help the researches and practitioner to choose appropriate Machine learning algorithm in healthcare.The various study have shown that Machine learning algorithm (CAD System) provide best performance in diagnosing diseases but the performance of the machine learning algorithm and other related issues are hardly available in single diagnosis.The output of this work produces a list of the best Machine learning algorithm with performance for detecting disease.

In the study of machine learning we can save the live of cancer patients .Only the Breast cancer in India accounts that one women is diagnosed every two minutes every nine minutes,one women dies. Diabetes is first recorded in English,in the form diabetes,in a medical text written around 1425.Thomas Willis added the word "mellitus"'to the word diabetes in 1675. Diabetes prevalence was estimated from studies in 91 countries. Population estimates and health expenditures were from the United Nations and the World Health Organization.John Browne(1642–1700)was an English surgeon who first described cirrhosis,in 1685, as Kidney disease means your Kidney are damaged and can't filter blood the way they should .You are at greater risk for kidney disease if you have diabetes or high pressure .If you experience kidney failure, treatment include kidney transplant or dialysis.these are the different dataset we used here like, kidney,Diabetes,Heart Diseases patients.These diseases are silent killer that's why we proposed model is that we can use Multiple Classification algorithm for better accuracy like,SVM Support Vector Machine,Naive Bayes,KNN,Random Forest , so it take minimum time to finding good result.The proposed method has produced highly accurate and efficient result when compared to the existing methods.

Our result on four different datasets using four different supervised algorithm algorithm SVM **96.19% on breast cancer dataset** when we use the other like Knn **76% on diabetes ,** we use the other like Random Forest 98**% on Kidney diabetes** method, we use the other like KNN **87% on Heart diseases** .

Keywords:Machine learning,Classification algorithm,Supervised datasets,Breast Cancer,Diabetes Disease,Kidney Diseases,Heart Diseases.

List Of Tables

List Of Figure

Table of contents

Chapter 1
INTRODUCTION

1.1 **Motivation and Overview**.

Machine learning algorithms make a significant contribution to design and develop computer aided diagnosis system.The work's overarching goal is to assist academics and practitioners in selecting a suitable machine learning algorithm in healthcare databases.Various studies have indicated that machine learning algorithms work best in detecting and identifications of several diseases in healthcare datasets.The day by day machine learning algorithms are used by the various researchers and scholars to improve the performance of the computer aided diagnosis system.

In this study we have proposed four supervised learning based algorithm such as [1] SVM(Support vector machine),KNN,Random Forest,and Naive Bayes models .These machine learning technique have been applied on four different healthcare datasets (taken from UCI machine learning repository) to analyse the performance of the algorithms. classify into two categories benign and malignant.Till we have applied now SVM model has obtained the performance result Machine learning has probably limited social impacts in the health-care field.Machine learning provides the solution for decreasing the increasing price of health-care and serving to create an improved patient-clinician communication.ML solutions will be used for an inordinately of health-relevant uses;some include serving to clinicians identify a lot of customized prescriptions and therapy for patients and additionally serving to patients identify once and if they must record follow up appointments.

Breast Cancer's causes are multi factorial and involve family history,obesity hormones,radiation therapy,and even reproductive factors.Every year one million women are newly diagnosed with breast cancer,according to the report of the world health organization half of them would die,because it's usually late when doctors detect the cancer.Breast Cancer is caused by a typo or mutation in a single cell,which can be shut down by the system or causes a reckless cell division.If the problem is not fixed after a few months,masses are formed from cells containing wrong instructions.

Machine Learning is a modern and highly sophisticated technological application that has become a huge trend in the industry.Machine Learning is Omnipresent and is widely used in various applications.It is playing a vital role in many fields like finance,Medical science and in security.Machine learning is used to discover patterns from medical data sources and provide excellent capabilities to predict diseases.In this paper,we review various machine learning algorithms used for developing efficient decision support for healthcare applications.During their life,among 8%of women are diagnosed with Breast cancer(BC),after lung cancer,BC is the second most common cause of death in both developed and undeveloped worlds.BC is characterized by the mutation of genes,constant pain,changes in the size,color(redness),skin texture of breasts.Classification of breast cancer leads pathologists to find a systematic and objective prognostic,generally the most frequent classification is binary(benign cancer/malign cancer).Today,Machine Learning(ML)techniques are being broadly used in the breast cancer classification problem.They provide high classification accuracy and effective diagnostic capabilities.In this paper,we present two different classifiers:Naive Bayes(NB)classifier and k nearest neighbour(KNN)for breast cancer classification.We propose a comparison between the two new implementations and evaluate their accuracy using cross validation.Results show that KNN gives the highest accuracy(97.51%)with lowest error rate then the NB classifier(96.19%).

1.2 Objective:

The objective of the study is to build a computer diagnosis system for analysis of various healthcare dataset.Here we are using four supervised learning based algorithm such as [1] SVM(Support vector machine),KNN,Random Forest,and Naive Bayes models .These machine learning technique have been applied on four different healthcare datasets (taken from UCI machine learning repository) to analyse the performance of the algorithms.

1.3 Contribution:

In this project we have studied ten research papers and studied the implementation part of each research paper, and we formulated the resulting table and in that table we made the columns as methodology, strength, weakness and efficiency of each research paper and finally we analysed all the above parameters after that we have done all the related work of the research papers and accordingly we proposed our method in which we are using the methods like SVM, KNN and Random Forest and we found the different accuracy for different algorithm like for SVM 96.19% on breast cancer dataset when we use the other like Knn 76% on diabetes , we use the other like Random Forest 98% on Kidney diabetes method, we use the other like KNN 87% on Heart diseases.

Chapter 2

LITERATURE SURVEY

In the study of algorithms that learn from data,develop models from that data,and use those models to forecast,make decisions,or solve problems.With regard to some classes of tasks T and performance P,a computer programme is to learn from experience E.The learning module and the reasoning module are the two components of ML.The learner module creates a model using information such as previous experience and background knowledge.The reasoning module employs models,and the reasoning module develops a solution to the task as well as a performance measure.Machine learning algorithms can anticipate or make judgments by generating a mathematical model based on training data.

It benefits society in a variety of ways.

Machine learning is reshaping the world by transforming a wide range of industries,including healthcare,education,transportation,food,entertainment,and various assembly lines,to name a few.It will have an influence on practically every element of people's life,including homes,automobiles,shopping,food ordering,and so on.

About the Healthcare:

It benefits society in a variety of ways.Healthcare is the prevention, diagnosis, treatment, amelioration, or cure of disease, illness, injury, and other physical and mental impairments in individuals,as well as the maintenance or enhancement of their health.Health professionals and associated health areas provide healthcare.Healthcare includes medicine, dentistry,pharmacy,midwifery,nursing,optometry, audiology, psychology, occupational therapy, physical therapy, sports training, and other health professions.It encompasses work in primary care, secondary care, and tertiary care, as well as public health.

An efficient healthcare system can contribute to a significant part of a country's economy, development, and industrialization.Healthcare is conventionally

regarded as an important determinant in promoting the general physical and mental health and well-being of people around the world.An example of this was the worldwide eradication of smallpox in 1980, declared by the WHO as the first disease in human history to be eliminated by deliberate health care interventions.

Table.1Analysis of machine learning algorithms--

S.No.	Authors	Algorithms	Feature and Classification	Performance
1.	Y.lreneous Anna Rejani	SVM	Divide the problem to classify	96%
2.	K.Shailaj,B.Seetharamulu,M.A.Jabbar [11]	KNN	Decision support System	76%
3.	Alok ChauhanHarshwardhan karpate,yogesh Naekar,A]Sakhshi Gulhane,Tanvi virulkar,Yamini Hadeu	Naive Bayes	Utilizing Different dataset	83%
4.	Mariem ADrane,Aliha Oukid,Iktam Gagsous,Tolga Ensari	Random Forest	Pattern classification	87%
5.	Wechao Xing[2]	SVM,Naive Bayes	Class Based weigthing	79%
6.	Debarase Mitra,Sumanta Chatterjee	SVM,Reinforceme net	Better predicition	87%
7.	Nareeem O.M.Salim	SVM,Naive,KNN	Non trival predicition	95%
8.	Hartatik,Mohammad,B adri Tamam,Arief satyanto	KNN,Naive Bayes	computational and analytical tools	72%
9.	Jarar zaidi	Naive,Log isitic Regressio n	Yield the highest acuuracy	74%
10.	Hamid Ilyas,Sajid Ali,Mahvish Ponum	Random Forest	Classification	96%

Chapter 3
MATERIAL AND METHODS

3.1 Proposed algorithm:

As a result,we may expect machine learning algorithms to play a key role in the design and development of computer-aided diagnostic systems.The overall purpose of the project is to help academics and practitioners choose the best machine learning algorithm from healthcare resources. Here we are using four supervised learning based algorithm such as [1] SVM(Support vector machine),KNN,Random Forest,and Naive Bayes models .These machine learning technique have been applied on four different healthcare datasets (taken from UCI machine learning repository) to analyse the performance of the algorithms.Machine learning algorithms have been shown to be the most effective at detecting and diagnosing a variety of illnesses in healthcare datasets in a number of studies.Various academics and scholars employ machine learning algorithms on a daily basis to increase the effectiveness of computer aided diagnostic systems

Here as given fig how the proposed model will work.

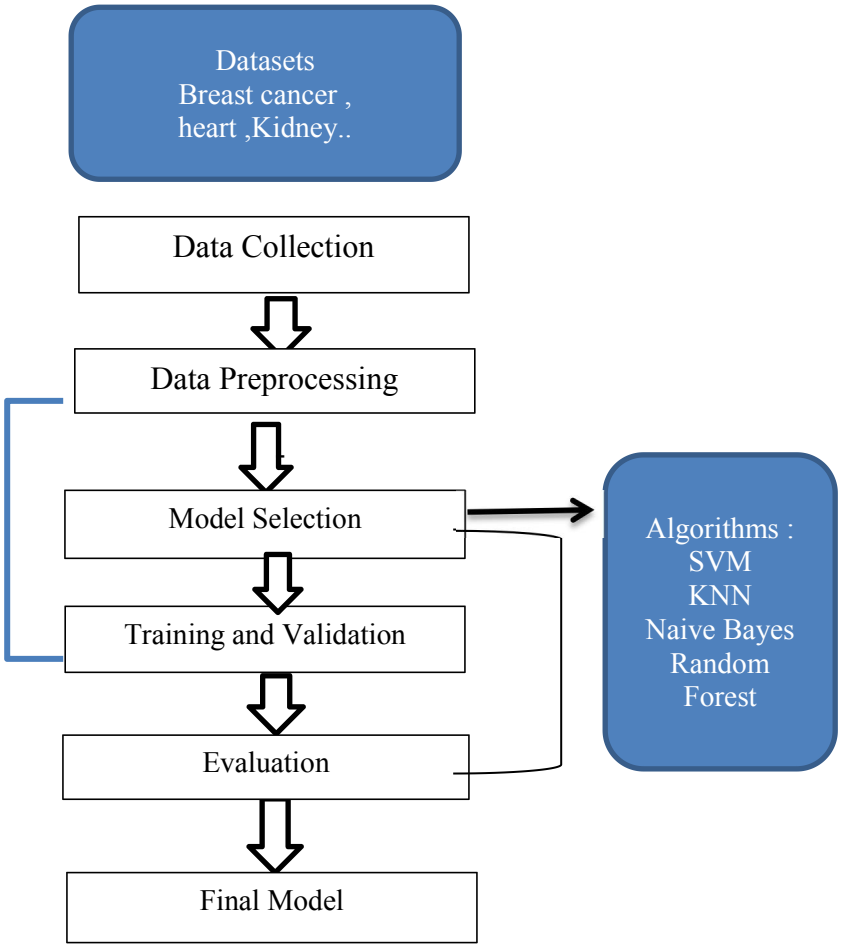


Fig.no.1 Proposed mode

3.2 Datasets

We have used 4 healthcare datasets in this analysis using different machine learning algorithms.These are the numerical datasets for supervised learning.All the datasets are from the UCI Machine Learning repository and Kaggle.

1. Breast Cancer-The numerical datasets is called the [1] Wisconsin Breast Cancer datasets.This datasets is taken from UCI Machine Learning repository.We have analyzed the datasets on different machine learning algorithms to predict two different types of breast cancers.
2. Diabetes-This numerical datasets is called [8] Pima Indians Diabetes

datasets.This datasets is taken from Kaggle repository.We have analyzed the datasets on different machine learning algorithms to predict the process of diabetes based on different factors present in the datasets.

3. Kidney Disease-The numerical datasets is calledKidney diseases datasets.This datasets is taken from Kaggle repository.We have analyzed this datasets on different machine learning algorithms to predict kidney diseases.
4. This numerical dataset is called Heart Disease dataset.This dataset is taken from kaggle repository.We have analyzed this dataset on different machine learning algorithms to predict heart diseases.

3.2.1 Breast Cancer Wisconsin(Original)Data Set using machine learning

Breast cancer can be caused by a number of factors that are both complex and numerous.Family history,obesity,hormones,and radiation are all factors.Treatment,as well as reproductive issues,must all be taken into account.One person is killed every year.A million women have been diagnosed with breast cancer for the first time.According to a World Health Organization estimate,half of them will die as a result of their tardiness.Doctors have the ability to detect cancer.The development of breast cancer is caused by a virus.a single cell that has made a mistake or has a mutation that can be switched off,or as a result of the system,leads to random cell division.The public will be upset if the matter is not fixed within a few months.They are made up of cells that have been wrongly coded.Tumors that are malignant can spread to neighbouring cells.

This can lead to metastasizing or spreading to other regions of the body.Because benign masses cannot spread to other tissues,their growth is confined to the benign mass alone.Because there are no symptoms in the early stages of BC,diagnosis might be challenging.Following a series of clinical testing,an appropriate diagnosis should be able to distinguish between benign and malignant tumour.

Classification of breast cancer;

BCC seeks to find the best therapy for the disease,which might be aggressive or less aggressive depending on the kind of cancer.Breast cancer categorization requires nine factors to produce a favourable prognosis.Recognize the several layers of structures.(clump diameter);

Take a look at the sample size and consistency(cell size uniformity).Estimate the equality of cell morphology and discover marginal variations since cancer cells vary in shape(Uniformity of Cell Shape.Cancer cells spread throughout the organ,whereas healthy cells create bonds(marginal adhesion).Uniformity measurement;larger epithelial cells suggest malignancy(Single Epithelial Cell Size).In benign tumour, the nuclei are not surrounded by cytoplasm(bare nuclei)..Describes the nucleus texture,which is homogenous in benign cells.In malignancies,the chromatin is coarser(Bland Chromatin).The nucleolus is normally tiny and inconspicuous in normal cells.There are more than one nucleoli in cancer cells,and they become much more conspicuous,(Normal Nucleoli);Estimation of the number of mitoses that have occurred.The higher the value,the higher the risk of cancer(Mitoses).Pathologists allocated to each of these diagnoses in order to categorized BC.

3.2.2 Diabetes Diseases:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.The objective of the dataset is to diagnostically predict whether or not a patient has diabetes,based on certain diagnostic measurements included in the dataset.Several constraints were placed on the selection of these instances from a larger database.In particular,all patients here are females at least 17 years old of Pima Indian heritage.

3.2.3 Kidney Diseases:

The data was taken over a 2-month period in India with 25 features(eg,red blood cell count,white blood cell count,etc).

The target is the'classification',which is either 'ckd' or' notckd '**-ckd=chronic kidney disease**.There are 400 rows The data needs cleaning:in that it has NaNs and the numeric features need to be forced to floats.Basically,we were instructed to get rid of ALL ROWS with Nans,with no threshold-meaning,any row that has even one NaN, get deleted.

3.2.4 Heart Diseases:

This is multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables,multivariate numerical data analysis.It is composed of 14 attributes which are age,sex,chest pain type,resting blood pressure,serum cholesterol,fasting blood sugar,resting electrocardiographic results,maximum heart rate achieved,exercise induced angina, oldpeak—ST depression induced by exercise relative to rest,the slope of the peak exercise ST segment,number of major vessels and Thalassemia.This database includes 76 attributes,but all published studies relate to the use of a subset of 14 of them.The Cleveland database is the only one used by ML researchers to date.One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has a heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

3.3 Data Gathering:

We will get the data from the **UCI Machine Learning Repository,**which is a collection of databases,domain theories,and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine.

So here we proposed the classification model(SVM Support vector machine),KNN,Naive Bayes and Random Forest on the datasets Breast Cancer Wisconsin(Original)Data Set using machine learning. We use Support vector Machine Classification.

3.4 Machine learning Algorithms

Machine Learning for Healthcare[3] Technologies offers algorithms with self-learning neural networks that can improve treatment quality by analysing external data on a patient's health,X-rays,CT scans,different tests,and screenings. We'll utilize SVM(support vector machine),random forest,KNN(kth closest neighbour),and Naive Bayes to classify the data.

Classification Algorithms can be further divided into the Mainly two category:

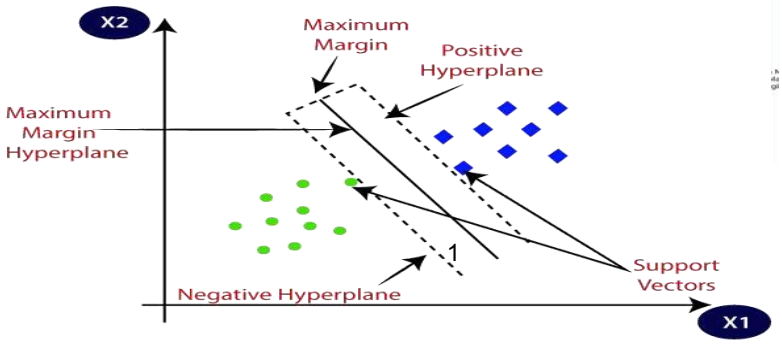
- **Linear Models**
  - Logistic Regression
  - Support Vector Machines
- **Non-linear Models**
  - K-Nearest Neighbour
  - Kernel SVM
  - Naive Bayes
  - Decision Tree Classification
  - Random Forest Classification

In the Above models we will work on the SVM,KNN,Random Forest,and Naive Bayes and as we discuss.

3.4.1 Support vector machine:

The Support Vector Machine,or [1]SVM,is a common Supervised Learning

technique that may be used to solve both classification and regression issues.However,it is mostly utilized in Machine Learning for Classification difficulties.The SVM algorithm's purpose is to find the optimum line or decision boundary for categorising n-dimensional space into classes so that additional data points may be readily placed in the proper category in the future.A hyperplane is the name for the optimal choice boundary.The extreme points/vectors that assist create the hyperplane are chosen via SVM.Support vectors are the extreme instances,and the method is called a Support Vector Machine.Consider the



.Fig.2 SVM(Support vector Machine)

picture below,which shows how a decision boundary or hyperplane is used to classify two separate groups.

3.4.2 K-Nearest Neighbour: Instatistics,the *k*-nearest neighbors algorithm(KNN)is anon-parametricclassificationmethod first developed by Evelyn Fixand Joseph Hodgesin 1951,and later expanded by Thomas Cover.

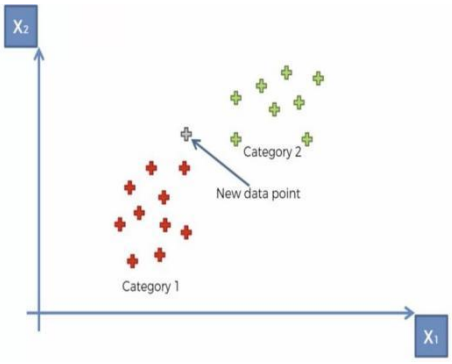


Fig.3 To find the k-th data point

Because of the rapid advancement of information technology,medical informatization has progressed in the direction of intelligence.Big data in medical health ensures a basic data resource for medical service intelligence and smart healthcare.For the intelligentization of medical information,categorization of medical health big data is critical.The KNN(K-Nearest Neighbor)classification technique is frequently used in various domains due to its simplicity.The efficiency of the KNN[3] method classification will be considerably lowered When the sample size is huge and the feature characteristics are large.This study compares and contrasts an upgraded KNN method with the standard KNN technique.The classification is done in the query instance neighbourhood of a standard KNN classifier,and each class is given a weight.To guarantee that the allocated weight does not have an unfavourable effect on the query instance,the algorithm evaluates the class distribution surrounding it.

3.4.3 Random forest: The general method of random decision forests[7] was first proposed by Ho in 1995.Random Forest is a well-known machine learning algorithm that uses supervised learning methods.In machine learning,it may be utilized for both classification and regression issues.

It's based on ensemble learning,which is the process of merging numerous classifiers to solve a complicated issue and enhance the model's performance."Random Forest is a classifier that contains a number of decision trees on various subsets of a given datasets and takes the average to enhance the predicted accuracy of that datasets,"according to the name.Instead than depending on a single decision tree,the random forest collects the forecasts from each tree and predicts the final output based on the majority vote of predictions.The bigger the number of trees in the forest,the more accurate it is and the problem of over fitting is avoided.

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.

- It predicts output with high accuracy,even for the large datasets it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.4.4 Naive Bayes:A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes'theorem with strong(naive)independence assumptions.Bayes'theorem was named after the Reverend Thomas Bayes(1702–61),who studied how to compute a distribution for the probability parameter of a binomial distribution.

- The Naive Bayes algorithm is a supervised learning algorithm,which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training datasets

The Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier,which means it predicts on the basis of the probability of an object.

The Naive Bayes algorithm is comprised of two words Naive and Bayes,Which can be described as:

- **Naive**:It is called Naive because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.Such as if the fruit is identified on the basis of color,shape,and taste,then red,spherical,and sweet fruit is recognized as an apple.Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes**:It is called Bayes because it depends on the principle of Bayes'Theorem:
- Bayes'theorem is also known as **Bayes'Rule** or **Bayes'law**,which is used to determine the probability of a hypothesis with prior knowledge.It

depends on the conditional

- The formula for Bayes'theorem is given as

Where,P(A|B)is Posterior probability:Probability of hypothesis A on the observed event B.P(B|A)is Likelihood probability:Probability of the evidence given that the probability of a hypothesis is true.P(A)is Prior Probability:Probability of hypothesis before observing the evidence.P(B)is Marginal Probability:Probability of Evidence .

P(A|B)=P(B|A)P(A)/P(B).

Chapter 4
RESULT

4.1 Performance evaluation of classification

Performance of classification is evaluated by calculating accuracy, sensitivity, specificity, F-Measure, and confusion matrix using the corresponding mathematical relationships, described below.

Accuracy

One of the most frequently used classification performance measures is accuracy. It is the ratio between the correctly classified samples to the total number of samples. The formula to calculate accuracy.

accuracy=TP+TN / TP+TN+FP+FN

Where, TP represents true positive values, TN represents true negative values, FP represents false positive values and FN represents false negative values.

Sensitivity

It is also called True Positive Rate (TPR), hit rate or recall. It represents the ratio of correctly classified positive instances to the total number of positive instances. The formula to calculate sensitivity, used in this study.

Sensitivity=TP / TP +FN

Specificity

It is also called True Negative Rate (TNR) or inverse recall. It measures the percentage of correctly classified negative instances to the total number of negative instances. The formula to calculate specificity.

Specificity=TN / TN + FN

F-Measure

F-Measure is calculated by taking the weighted average of sensitivity and precision values. The formula to calculate F-Measure.

F−Measure=2∗ sensitivity ∗ precision / sensitivity + precision

Measure uses the field of information retrieval for the estimation of classification performance.

Precision

Precision is defined as what proportion of positive identifications was actually correct. The formula to calculate precision.

Precision = TP/TP+FP

Experimental setup and Performance:

4.2 On Diabetes datasets:

The result of this evalution here, we are discuss about the result table no.2 in which we use Diabetes dataset and we apply all algorithm which are used in the proposed model fig no .1and we get the accuracy 76% percent using KNN which was best fit for it. Confusion matrix in which we can find the predicted No and Predicted yes.

Table.2 Performance of the machine learning Models Diabetes datasets

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.76	0.92	0.83	150
1	0.76	0.46	0.57	32
ACC.	0.69	0.81	0.76	231
MACRO.AVG.	0.76	0.69	0.70	231
WEIGHTED AVG.	0.76	0.76	0.74	231

Confusion matrix:

No. OF features	Predict No	Predict Yes
Actual No	123	27
Actual Yes	40	41

4.3. On kidney Datasets :

Confusion matrix

No. Of features	Predicted NO	Prediected YES
Actual NO	23	0
Actual YES	0	9

Accuracy:[96%]

4.4 On Heart diseases:

Here in given result as we shown in the fig 4 for diffrent K values

Between the AXIS scores and Number of neighbours K.

On heart diseases datset we apply all algorithm but the KNN gives us highest accuracy is 87%.

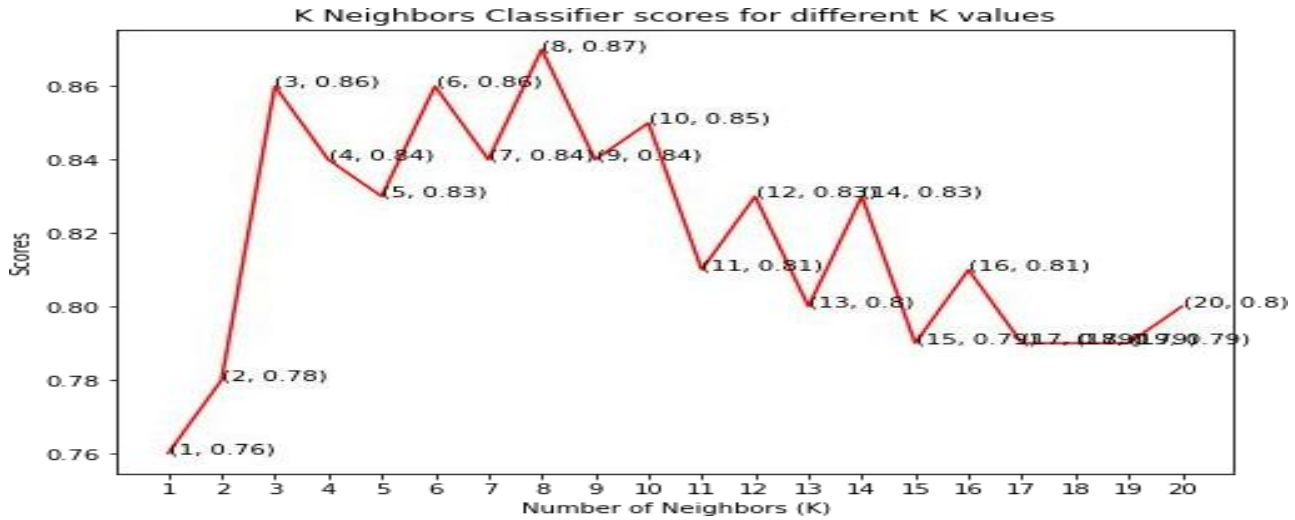


Fig.3 Result of heart disease using KNN classifier

Accuracy:[87%]

4.5 On Breast Cancer:

we are discuss about the result table no.3 in which we use Breast cancer dataset and we apply all algorithm which are used in the proposed model fig no .1and we get the accuracy 96% percent using SVM which was best fit for it.

Table.3 Performance of the machine learning Models On heart diseases datasets

	precision	recall	F1-score	support
0	1.00	0.98	0.99	48
1			0.99	66
Avg/Total	0.99	0.99	0.99	114

4.6 Result of all Algorithms:

As we know we are apply four algorithm SVM, KNN ,Naive bayes and Random forest on the dataset as given in our proposed model fig.no.1

Earlier we are discussed about the datasets like Breast cancer, Diabetes, Kidney and Heart diseases. Here we diagnosis and find the best algorithm for these given datasets as shown in the given table no.4 for which dataset is which algorithms gives us best result and accuracy like for Breast cancer SVM gives 96% accuracy and for diabetes KNN is best it gives us 76% and for Kidney disease Random forest is best it gives us 98% accuracy and for heart disease KNN it gives 87% accuracy .

Table no.4 Classification performance of the supervised machine learning algorithms on different dataset.

S.No.	Datasets	Algorithms	Accuracy	F1-score	Precision
1	**Breast Cancer**	**SVM**	**96%**	**0.99**	**0.99**
2	**Diabetes Diseases**	**KNN**	**76%**	**0.76**	**0.57**
3	**Kidney Diseases**	**Random Forest**	**98%**	**0.75**	**0.69**
4	**Heart Diseases**	**KNN**	**87%**	**0.78**	**0.59**

5.Conclusion and future work:

These are the few potential areas whereMachine Learningcan help the healthcare industry out of many scenarios.We see,with machine learning applications,the

healthcare and medicine segment can advance into a new realm and completely transform the healthcare operations. ML has an extensive experience in providing diagnostics,analysis,imaging,wearable and medicine solutions to healthcare organizations.We have used machine learning algorithms to classify datasets available on healthcare like breast cancer datasets by wisconsin on UCI repository.For now we have worked on SVM,KNN,Naive Bayes and Random Forest and were able to get the accuracy of 96%.

5.1 Future Scope:

Machine learning allows building models to quickly analyze data and dekidney results, leveraging historical and real-time data.

With machine learning, healthcare service providers can make better decisions on patients' diagnoses and treatment options, which lead to an overall improvement of healthcare services.

Previously, it was challenging for healthcare professionals to collect and analyze the huge volume of data for effective predictions and treatments since there were no technologies or tools available. Now with machine learning, it's been relatively easy, as big data technologies such as Hadoop are mature enough for wide-scale adoption.

In fact, 54% of organizations are using or considering Hadoop as big data processing tool to get important insights on healthcare according to the Ventana Research Survey. 94% of Hadoop users out of existing users perform analytics on voluminous data which they believe was not possible before.

Machine learning algorithms can also be helpful in providing vital statistics, real-time data and advanced analytics in terms of the patient's disease, lab test results, blood pressure, family history, clinical trial data, etc., to doctors.

Reference:

[1]. M.ADrane,S.Oukid,I.Gagaoua and T.Ensarİ,"Breast cancer classification using machine learning,"2018 Electric Electronics,Computer Science,Biomedical Engineerings'Meeting(EBBT),2018,pp.1-4,doi:10.1109/EBBT.2018.8391453.

[2]. Xing,Wenchao,and Yilin Bei."Medical health big data classification based on KNN classification algorithm."*IEEE Access* 8(2019):28808-28819.

[3]. Jain,Vishal,and Jyotir Moy Chatterjee."Machine learning with health care perspective."*Cham:Springer*(2020):1-415.

[4]. Chauhan,Alok,et al."Breast Cancer diagnosis and Prediction using Machine Learning."*2017 Third International Conference on Inventive Research in Computing Applications(ICIRCA)*.IEEE,2017.

[5]. ADrane,Meriem,et al."Breast cancer classification using machine learning."*2018 electric electronics,computer science,biomedical engineerings'meeting(EBBT)*.IEEE,2018.

[6]. Patel,Jaymin,Dr TejalUpadhyay,and Samir Patel."Heart disease prediction using machine learning and data mining technique."*Heart Disease* 7.1(2015):129-137.

[7]. Salim,Nareen OM,and Adnan Mohsin Abdulazeez."Human diseases diagnosis based on machine learning algorithms:A review."*International Journal of Science and Business* 5.2(2017):102-113.

[8]. Hartatik,Hartatik,Mohammad Badri Tamam,and Arief Setyanto."Prediction for diagnosing kidney disease in patients using KNN and Naïve Bayes algorithms."*2020 2nd International Conference on Cybernetics and Intelligent System(ICORIS)*.IEEE,2020.

[9]. Chang,V.,Bhavani,V.R.,Xu,A.Q.,&Hossain,M.A.(2022).An artificial intelligence model for heart disease diagnosis using machine learning algorithms.*Healthcare Analytics*,*2*,100016.

[10]. Ilyas,H.,Ali,S.,Ponum,M.,Hasan,O.,Mahmood,M.T.,Iftikhar,M.,&Malik,M.H.(2017).Chronic kidney disease diagnosis using decision tree algorithms.*BMC nephrology*,*22*(1),1-11.

[11]. Shailaja, K., B. Seetharamulu, and M. A. Jabbar. "Machine learning in healthcare: A review." *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 2018.

[12]. Ibrahim, I. and Abdulazeez, A., 2021.