



# INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

## DISEASE PREDICTION USING MACHINE LEARNING

Pallav Verma, Sneha Verma

Department of Computer Science and Engineering  
SRMIST Delhi-NCR Campus, Ghaziabad, UP

### ABSTRACT

Information Mining may be a procedure that's performed on a huge database for extricating covered up designs by utilizing combinational techniques from measurable examination, machine learning and database innovation. Advanced, restorative information mining is an amazingly critical inquiry about the field due to its significance within the improvement of different applications in prospering healthcare spaces. Whereas outlining the passings happening around the world, the cardiac infection shows up to be the driving cause. The recognisable proof of the conceivable outcomes of cardiac illness in an individual is complicated tests for therapeutic professionals since it requires a long time of involvement and a serious therapeutic test to be conducted.

### INTRODUCTION

Information is basically a method of extricating secret data from expansive volumes of crude information. Data must be modern, non-obvious, and usable. Information can as it were characterized "the trifling extraction of already obscure, important and valuable data". It is "the science of extricating important data from expansive information sets". Infection expectation plays an vital part in information mining

The venture recognised heart malady, kidney illness, and diabetes. In case of Arbitrary timberland we fair utilize different choice trees prepare them with information collected at that point based on the yield of all the choice trees

Restorative records have the capacity to hunt for covered up designs in information in therapeutic records. This demonstration can be utilized to analyze heterogeneous crude clinical information commonly found in nature. This data ought to be composed in a frame. The collected information can be combined to create a clinic database. Information mining procedures give a user-oriented way to reveal modern and covered up designs in information.

According to the World Health Organisation, about 12 million people worldwide die from heart disease each year. Half of all deaths from heart disease occur in India and other developing countries. In the above discussion, it is considered a cause of death for adults. Every 34 seconds a person dies from a heart attack in India. This project uses machine learning techniques to evaluate predictors of heart disease, kidney disease and diabetes.

The information are basically to foresee and clarify. Expectation comprises particular factors and regions within the information to anticipate obscure or future results of other zones of intrigued. Clarification, on the other hand, centers on finding designs to clarify information that can be deciphered by people.

### Predicting Cardiovascular Disease

In,[7] Three diverse following machine learning calculations have been proposed. Gullible Bayes, KNN and Choice List calculations. Three calculations for cardiac information investigation. These categorical information were assessed utilizing 10-fold cross approval and the comes about were compared. The choice tree is one of the foremost well known and vital items that are simple to utilize. No enlistment or parameter setting required. It works with expansive volumes of information. It is essential for logical inquiry. Comes about from choice trees are simpler to decipher and perused.

Naive Bayes is an exclusionary statistic that cannot assign a probability to attributes. To determine the class, the posterior must be complete. The advantage is that Naive Bayes models can be used without using Bayesian methods.

The K-nearest neighbor calculation (K-NN) is an imperative strategy for classification based on understood information learning at a given area. It is the only of all machine learning calculations, but the exactness of the K-NN calculation can be diminished by the nearness of clamor. The examination was done utilizing preparing 3000 cases of 14 distinctive characteristics. Information is isolated into both test and preparation, for illustration 70 The creators concluded that the Credulous Bayes calculation performs well compared to other calculations.

Information mining finds vital data covered up in expansive records. Weka Apparatuses could be a collection of machine learning calculations for information mining forms. It has preprocessing, classification, relapse and affiliation rules, clustering and perception instruments. We utilize the Gullible Bayes approach to perform the mining and classification handle. We utilize 10x cross approval to diminish any Bayesianism within the process and improve the execution of the method. A choice tree may be a directed learning calculation for understanding classification issues. The main reason for utilizing choice trees in this inquiry about venture is to anticipate classes utilizing choice rules inferred from past information. Employment hubs and between hubs for forecast and classification.

## DIABETES

Affront is one of the foremost vital hormones within the body. It makes a difference the body changes over sugars, starches and other supplements into vitality vital for standard of living. However, if the body cannot deliver or utilize affront accurately, an abundance of sugar from the affront is discharged within the pee. This infection is called diabetes. In spite of the fact that weight and physical dormancy appear to play a critical part, the cause of diabetes is obscure. According to November 2020 information from the American Diabetes Affiliation, 20.8 million children and grown-ups within the United States have been analyzed with diabetes. Early determination of diabetes plays a critical part within the treatment of the persistent.

In, [2] The creators foresee whether unused patients will be analyzed with diabetes. This article investigates a modern strategy called homogeneity-based calculation or (HBA) to decide the most excellent control for the overfitting and overgeneralization conduct of the conveyance of this information. Combined with the classification framework, the HBA is more productive than the current framework. The creators concluded from the tests that both diabetes forecast and information mining communities are vital.

In [6] Information mining calculations were utilized to assess the precision of the blood glucose estimation. Fluffy frameworks are utilized to create hereditary calculations to solve different issues and totally different applications. Fluffy frameworks permit the presentation of learning and adjustment. Neural systems are valuable for learning enrollment capacities. Diabetes happens around the world, but sort 2 is more common in creating nations. The creators utilized this calculation in this hereditary calculation. The steps included in this preparation are choice, competition, exchange, application and counting. The creators concluded that the utilization of hereditary qualities to attain chromosomal redress based on more seasoned diabetic patients may be constrained to accomplishing chromosomal precision in more youthful patients.

## KIDNEY DISEASE PREDICTION

The kidney's work is to channel the blood. All the blood in our body passes through the kidneys a few times a day. Kidneys evacuate squander items, keep up the body's liquid and electrolyte adjust. As the kidneys channel the blood, they deliver pee that collects within the renal pelvis, a funnel-shaped structure that streams into the bladder through vessels called the ureters. Each kidney has around one million units called nephrons, each with a channel or blood vessel. Up to 90% of kidney work will vanish without any indications or issues. There are numerous variables that increment the hazard of kidney illness.

- Diabetes
- Hypertension
- Smoking
- Obesity
- Heart Disease
- Family history of kidney disease • Alcohol intake

kidney symptoms

- Trouble or torment amid voiding
- Changes in your urinary work
  - Blood in the urine
  - Swelling or torment within the back or sides
  - Extraordinary weariness and summed up shortcoming

### LITERATURE SURVEY

Machine Learning techniques such as classification, clustering and association rules play an important role in extracting unknown information from data. Classification is a data mining technique used to estimate group membership of data. Partitioning is similar to clustering in that it divides data into different parts called classes. To predict the outcome, this is often called an objective or characteristic forecast.

Author & Year of Publication	Title of Paper	Methodology	Merits	DeMerits
C.Saravanabhan T.Saravana (2022)	Anticipating the impact of diabetes on kidney utilizing classification in Tanagara. [2]	SVM ranking with backward search Technique.	The proposed technique, enhanced the precise accuracy of Gullible Bayes Classifier and help healthcare calling for determination of type2 diabetes.	The proposed method increase the classification accuracy only by 1.88%
Kazban Alpan, G.S (2021)	A Patient Network-Based Machine Learning Model for Disease Prediction - The Case of Type 2 Diabetes [3]	Building a predictive model using Hybrid Approach	The Hybrid Approach gives the classification accuracy of 92.38%.	The system doesn't apply any feature technique to remove redundant feature.

senthilku mar mohan G.Srivastava (2021)	Predictive data  Kho Mob Mining Diagnostics Overview Cardiology prediction. [7]	Our different NN models are for Pima Indians, for Diabetes, is used to find the best method method out of these.	General Regression Neural Network functional best 80.21% accuracy.	Recommend d system limited to only NN find the best method separate diabetes patients.
---	---	---	---	--

Author & Year of Publication	Title of Paper	Methodology	Merits	DeMerits
C.Saravanabhavan T.Saravana (2022)	Anticipating the impact of diabetes on kidney utilizing classification in Tanagara. [2]	SVM ranking with backward search Technique.	The proposed technique, enhanced the prescie-nt accuracy of Gullible Bayes Classifier and help healthcare calling for determination of type2 diabetes.	The proposed method increase the classificati on accuracy only by 1.88%
Kazban Alpan, G.S (2021)	A Patient Network- Based Machine Learning Model for Disease Prediction - The Case of Type 2 Diabetes [3]	Building a predictive model using Hybrid Approach	The Hybrid Approach gives the classification accuracy of 92.38%.	The system doesn't apply any feature technique to remove redundant feature.
senthilkumar mohan G.Srivastava (2021)	Predictive data  Kho Mob Mining Diagnostics Overview Cardiology prediction. [7]	Our different NN models are for Pima Indians, for Diabetes, is used to find the best method method out of these.	General Regression Neural Network functional best 80.21% accuracy.	Recommend d system limited to only NN find the best method separate diabetes patients.

This literature review table particularly includes key information such as Arthur/s Name and Year of Publication, Title of Paper, Methodology Used, Dataset Used and Limitations from each Research Paper/Article that is reviewed. This table provides a visual overview of the relevant literature and can help identify gaps in this research.

### PROPOSED METHODOLOGY

Cardiovascular disease prediction can be done as a method to inform the study to build the classification model needed to predict the patient’s heart disease. This model creates a simple system for predicting heart disease using machine learning techniques.

Data Collection from various Government provided online sources.It is real time data which the program collects every time when it is executed and save it to a particular location and delete it after the prediction is done.Dataset has been created for analysis of heart disease, kidney disease and diabetic.

The quality is numerical information speaking to the patient’s age, and the extend is 29 to 65 a long time. cp is an trait to decide the condition and the most extreme esteem is 1 . Trestbps stands for resting blood weight between 92 and 100, fbs implies quick blood sugar, or 1, boolean implies genuine or untrue. restecg are resting electrocardiogram comes about between and 2 concurring to our conditions. thalach is the most noteworthy heart rate and ranges from 82 to 185. exang is work out including angina, which is boolean. The reason of the information is malady, with or 1 showing the nearness of cardiovascular illness.

First we need to filter out the data for which we will check if there are rows with missing values, if there are then we will drop these rows if they are not much affect the prediction otherwise try to fill it with mean or median values which ever is less greater.After that we will exclude all the unnecessary column Use coding to predict.

Here When we will train the model and validate it with the first filtered data we will try to make model and do prediction with more optimized data set that is we will exclude further non affecting data which is for particular state only we do not need name of state as it will be just repeated again and again and nothing else.

### SYSTEM ARCHITECTURE

Gullible Bayes is an exclusionary measurement that accepts no relationship between subjugation and conduct. Gullible Bayes is based on the The credulous Bayes classifier works like this:

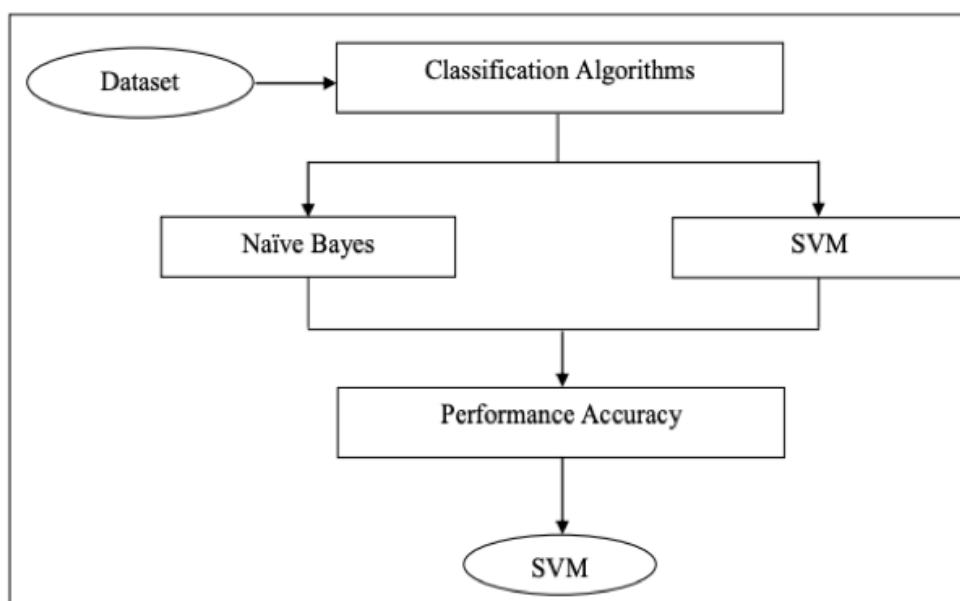


Figure 1: SYSTEM ARCHITECTURE

- Expectation Step:For obscure test information, the demonstration calculates the back likelihood that the dataset has a place in each lesson. The strategy at long last parts the test information agreeing to the biggest back likelihood within the preparation.

A decision tree is simple and easy to use individually. A few features entered in patient data are only available in the decision tree. Decision Tree Simplifies debugging and processing by creating distributed or repetitive models in tree models. Decision trees can be used for both categorical and numerical information. The calculation works by finding data around the properties and expelling the qualities utilized to part branches.

- Step 1: Decide the information pick up of the character within the information.
- Step2:Partitioned information increases from the heart malady dataset in address.
- Step 3:After deciding the information pick up, put the most excellent highlight of the dataset at the root of the tree.
- Step 4:The information increase is calculated utilizing the same equation.
- Step 5:Hubs are part concurring to the most noteworthy information rate.
- Step 6:This handle is rehashed until all objects are organized and cleared out in each department.

In Linear Regression we know that the equation of line is in this form  $Y = a + bX$ . Here, X can be said as an input variable whose value we will get by prediction. b is the slope of a line which we will get by taking the mean of both variables and dividing them, and can be said as an intercept. A scatter plot is used to determine the strength and correlation between these two variables which we are going to use to do our prediction. If the two variables are not related that we can see in the scatter plot it will not be straight increasing or decreasing.

## RESULT AND ANALYSIS

This chapter is all about testing code to perform better prediction. Here we are going to see how well our data is fitted in the model, how accurate is our data, and the output that we are getting.

The screenshot displays a web interface for a 'Disease Prediction System'. At the top, there are navigation links for 'Home' and 'About'. Below the navigation is a red header bar with the text 'HEART DISEASE PREDICTION'. The main content area is titled 'CHECK YOUR HEART DISEASE' and contains a series of input fields for patient data. The fields are: Age (text), Gender (text), Chest Pain Type (text), TRESTEPS (text), Cholesterol (text), Fasting Blood Sugar (text), Resting electrocardiographic (text), Thalach (text), Exang (text), OldPeak (text), Slope (text), Ca (text), and Thal (maximum heart rate) (text). At the bottom of the form is a red button labeled 'CHECK AND SHOW RESULT'.

Figure 2: HEART DISEASE INPUT

YOUR HEART DISEASE PREDICTION RESULT

**There is 30% chances that you are Suffering from Heart Disease**

Label	Value
Age	63
Chest Pain Type	3
TRETBPS	145
CHOLESTROL	233
Fasting Blood Sugar	1
Restecg	0
Thalach	150
Exang	0
Oldpeak	2
Slope	3
Ca	0
Thal	0

Figure 3: HEART DISEASE OUTPUT

CHECK YOUR DIABETES

PREGNANCIES

6

GLUCOSE

148

BLOOD PRESSURE

72

SKIN THICKNESS

35

INSULIN

0

BMI

33

DiabetesPedigreeFunction

0

AGE

50

CHECK AND SHOW RESULT

Figure 4: DIABETES INPUT

**YOUR DIABETES RESULT**

**There is 90% chances  
that you are Suffering  
from Diabetes**

Label	Value
Pregnancies	6
Glucose	148
BloodPressure	72
SkinThickness	35
Insulin	0
BMI	33
DiabetesPedigreeFunction	0
Age	50

Figure 5: DIABETES OUTPUT



**CHECK YOUR KIDNEY DISEASE**

Age

Blood Pressure

SG

Aluminium toxicity

Sulfonylureas

Red Blood Cell

PC

pheochromocytoma

BA

BGR

Figure 6: KIDNEY DISEASE INPUT

BGR

BU

SC

SOD

POT

Hemoglobin

PCV

WC

RC

HyperTension

DM

Figure 7: KIDNEY DISEASE INPUT

CAD

Appet

PE

ANE

CHECK AND SHOW RESULT

Figure 8: KIDNEY DISEASE INPUT

**YOUR KIDNEY DISEASE PREDICTION RESULT**

**There is 90% chances that you are Suffering from Kidney Disease**

Label	Value
BP	80
SG	1
AL	1
su	0
RBC	0
PC	0
PCC	0




Figure 9: KIDNEY DISEASE OUTPUT

Our machine learning and data mining models are trained for prediction. Attribute is a numeric data from 29 to 65 years old representing the age of the patient. cp, 1 to . Trestbps is resting blood pressure between 92 and 100, fbs is a fasting blood sugar level off with 0 or 1 in boolean representing true or false.

Logistic Regression is a statistical method commonly used in machine learning for disease prediction. It is a type of regression analysis that is used to predict binary outcome such as presence or absence of diabetes. To use logistic regression for disease prediction, first a datasets is collected that includes information about patients with or without the

disease. These data were divided into two groups, one to train the logistic regression and model, and the other to test the accuracy

The code is written in python using Pandas ,numpy, sklearn for DataScience purpose and it is working efficiently without giving any error.there will not be any exception error will be thrown.We will have to provide the of the particular disease after filing all the input then it will show the output.Remember every time after using we have to clear the memory.

## CONCLUSION

Disease prediction using machine learning is a powerful technique that can provide valuable insights into patients health outcomes and inform clinical decision-making.By analyzing large datasets containing patients attributes and health outcomes, data mining techniques such as logistic regression can be used to identify risk factors for disease.predict disease outcomes, and develop personalized treatment plans.

The use of data mining in disease prediction has numerous benefits, including the ability to identify previously unknown risk factors and relationship between patients attributes and dis- ease, improve patients outcomes through early detection and intervention and optimize health- care resource utilization. However the accuracy and effectiveness of data mining techniques are heavily dependent on the quality and completeness of the input data.It is essential to ensure that the data is accurate, relevant, and up-to-data to obtain accurate predictions.

Overall, disease prediction using data mining holds great potential for improving patient out- comes and advancing healthcare research and it is essential for healthcare professionals and researchers to continue exploring and utilizing these techniques in the future.

## REFERENCES

- [1] aslam khan, F. and bukhari, S. A. C. (2021). "Detection and prediction of diabetes using data mining, 2012, 7, pp 20-25." IEEE.
- [2] C.Saravanabhavan, T.Saravana, D. B. M. "Kidney disease prediction using data mining volume 3, issue 4,april 2022." IEEE.
- [3] Kazban Alpan, G. S. (2021). "Diabetes data set with data mining technique by using weka approach." IEEE, 2021.
- [4] K.R.Lakshmi, Y. and M.veeraKrishna (March 2019). "Performance comparison of three data mining techniques for predicting kidney disease survivability." IEEE.
- [5] K.Srinivas, G.Raghavendra Rao, A. (2020). "Prediction of heart attacks in coal mine regions using data mining approach." IEEE, 19(2).
- [6] Manimekalai, K. . D. M. (January 2020). "Study of heart disease prediction using data mining." IJARCSSE, 4.
- [7] Senthilkumar Mohan, C. T. and Srivastava, G. (2021). "Effective heart disease prediction using machine learning and data mining techniques." IEEE Vol.3 No., 366(9498), 1744– 1749.