# Medical Diseases Prediction using Machine Learning Algorithms

This Thesis is submitted in fulfilment of the requirement for the degree of
Bachelor of Science in Computer Science & Engineering.

**Submitted By**

Sosmita Akter
ID: 1105083
Khadija Begum
ID:1105085
K.S.M Ibrahim Shorif
ID:1105073

**Under the Supervision of**

Md. Fazle Rabbi
Lecturer
Department of Computer Science and Engineering
Bangladesh Army International University of Science and Technology

**Bachelor of Science in Computer Science & Engineering.**
**Bangladesh Army International University of Science and Technology**
**Spring 2021**

# Declaration of Own Work

This is to certify that the work presented in this thesis, titled, “Medical Diseases prediction using Machine Learning”, is the outcome of the investigation and research carried out by us under the supervision of Md. Fazle Rabbi

We thus announce that this accommodation is our claim work and to the finest of our information it contains no materials already distributed or composed by another individual, or significant extents of material that have been acknowledged for the grant of any other degree or confirmation.

**-----------------------------------**
**Signature of the Candidate**
**Date:**

**-----------------------------------**
**Signature of the Candidate**
**Date:**

**-----------------------------------**
**Signature of the Candidate**
**Date:**

# Approval

The Thesis titled “Medical Disease's prediction using Machine Learning” has been submitted to the following respected members of the Board of Examiners in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering in September 2021 by the following student and has been accepted as satisfactory.

**Author ID: 1105073 , 1105083 , 1105085**

**Supervisor: ________________________________________**

**Md. Fazle Rabbi**
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science and Technology

**External Board Member: ________________________________**

**Dr. Md. Saifur Rahman**
Associate Professor
Dept of ICT,
Cumilla University

# Acknowledgment

First and foremost, we would like to express our deepest sense of gratitude to Almighty Allah for giving us the strength and ability to finish this thesis work. Then to our revered parents who are the inspiration and blessings of our lives. The satisfaction that accompanies the successful completion of this thesis would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success

We would like to express our wholehearted gratitude to our honourable Supervisor **Md. Fazle Rabbi, Lecturer, Department of Computer Science & Engineering, BAIUST**, for giving us his support, guidance, and valuable suggestions throughout this entire research work.

We are also grateful to the Head of Department **Mohammad Asaduzzaman Khan and all the faculty members of Dept. of Computer Science & Technology, BAIUST**, for their valuable support, motivation, and time throughout our journey of thesis work. Without their support and time, it would not be possible to complete this work.

Finally, we would like to thank our batch mates and seniors for their valuable suggestions and assistance that has helped in the successful completion of the thesis.

**Author**

**K.S.M Ibrahim Sharif**
**Sosmita Akter**
**Khadija Begum**

# Abstract

In today's time, many diseases are on the rise in the world. Many people are dying due to a lack of proper treatment and right decisions. In 2019, the top 10 causes of death accounted for 55% of the 55.4 million deaths worldwide.

At a global level, 7 of the 10 leading causes of deaths in 2019 were no communicable diseases. These seven causes accounted for 44% of all deaths or 80% of the top 10. However, all non-communicable diseases together accounted for 74% of deaths globally in 2019.

The world’s biggest killer is heart disease, responsible for 16% of the world’s total deaths. Since 2000, the largest increase in deaths has been for this disease, rising by more than 2 million to 8.9 million deaths in 2019. Then different types of cancer, liver diseases, diabetes are also responsible for deaths.

Applying optimal machine learning models for early detection and accurate prediction of those diseases is essential to reducing deaths and treating patients with the best clinical decision support. This raises the motivation for this paper and this paper presents a survey of prediction derived and validation on different types of medical diseases.

Machine learning could function as a truly alternative diagnostic tool for prediction, giving patients information about their health status. Despite the efforts of researchers, uncertainty about the standardization of predictive models still remains and additional research on the optimal predictive model is needed.These machine learning methods take less time to more accurately predict diseases, thereby reducing precious lives worldwide**.**

# TABLE OF CONTENTS

# Figure

# Chapter 01: Introduction

## 1.1 Introduction

Medical data sets include a vast amount of medical data, various measurements, financial data, statistical data, demographics of specific populations, and insurance data, to name just a few, gathered from various healthcare data sources.

Data interchange in the US healthcare industry is strictly regulated both on national and federal levels. The Health Insurance Portability and Accountability Act (HIPAA), published in 1996, is the core set of healthcare IT data standards.

The Health Information Technology for Economic and Clinical Health Act (HITECH Act), adopted in 2009, is aimed to "improve healthcare quality, safety, and efficiency through the promotion of health IT, including electronic health records (EHR) and private and secure electronic health information exchange".

According to both laws, Medical data sets are to identify the data elements to be collected for each patient and to provide uniform definitions for common terms. [1]

Medical data sets are various kinds like Cancer, Heart disease, Heart Failure, Covid-19,Breast cancer, liver disease, Diabetes etc.

In this research, we collect five datasets like- Heart disease, Heart Failure, liver disease, Diabetes and cancer etc. Here, we try to show the difference Row dataset between a trained Model dataset

First, we trained every dataset under a Model. The model built by Data preprocessing like- Feature encoding, missing value handling, oversampling, feature scaling, feature selection. Then the datasets trained and applied machine learning algorithms.

Second, we applied machine learning algorithms to model the row datasets.

Finally, we try to show the actual difference between the model trained datasets and without model datasets' accuracy.

## 1.2 Problem Statement :

In this article our work is to find the prediction of different types of medical diseases using Machine Learning Algorithms to support vector machines, K-nearest algorithm. In this project, we collect several datasets from Kaggle and UCI. Datasets are heart diseases, heart failure, Wisconsin breast cancer diagnostic, liver disease, Hepatitis, diabetes, and breast cancer Coimbra. We have to predict the target of All these seven datasets. We need to do data preprocessing of all these seven datasets.

In this work, the encoding method is applied as a first step.

Second step, Over-sampling is applied.

Third step, the Feature selection method is applied.

Fourth step, stratified K-fold Cross-validation is applied as the Cross-validation method.

Therefore, the fifth step applies classification methods or different algorithms to diagnose different medical diseases and measures the classification accuracy to evaluate the performance of characteristic selection methods.

The main contributions of this paper are:

1. Extraction of classified accuracy useful for medical disease prediction and removing redundant and irrelevant features with feature selection method.

2. Use the over-sampling method and cross-validation method.

3. Comparison of different machine learning algorithms on disease datasets.

4. Comparison of differences with data pre-processing machine learning algorithms on medical disease datasets and without data pre-processing machine learning algorithms on medical disease datasets.

5. Identification of the best performance-based algorithm for medical disease prediction.

## 1.3. Motivation and background of the study:

At this time, most of us are getting busy in our lives and work so much that we are not even having the time to take care of ourselves. Most of the time we people experience anxiety, depression, stress, and many other things for our hectic lives. Considering these as the main factors, we're getting sick and having severe diseases. There are many diseases like cancer, heart condition, diabetes, liver, hepatitis, etc which also cause the death of the people per annum.

A vital part of the physical body is the heart. All the pumping of the blood is completed by memory to each part of the body and through that blood oxygen, and other nutrients are supplied to the body.

**Heart Disease/ Disorders (HD):** Heart Disease/ Disorders (HD) has been recognized together with the convoluted and fatal human illness within the world. For this disease, the guts function abnormally resulting in blocked blood vessels and suffering from angina, attack, and stroke. The foremost common sorts of heart diseases are Coronary Vascular Disease (CVD), arterial coronary Disease (CAD), Congestive coronary failure (CHF), and Abnormal Heart Rhythms.[1] There are many challenges in predicting such HD at the first stages thanks to the involvement of several conventional risk factors like age, sex, hypertension, high cholesterol, abnormal pulse, and lots of other factors. Despite wide diversity within the existence of cardiovascular risk factors across different sectors of society, CVD has been noticed to be one of the main causes of death everywhere in India, including economically backward states and rural areas. The worldwide statistics also showed that the premature mortality in terms of years of life lost due to CVD climbed to 37 million (2010) from 23.2 million (1990) with an incremental rise of 59% each year.[2]

**Heart Failure:** Heart failure means that the heart is incapable of pumping blood around the body correctly. It generally happens because the heart has become too weak or stiff. It is sometimes referred to as congestive heart failure. Heart failure does not mean that your heart has ceased to function. In 2011, non-hypertensive heart failure was one of the 10 most expensive conditions observed during hospital admissions in the United States, with total hospitalization costs of more than $10.5B.The current global prevalence of FH is 64.34 million cases, representing 9.91 million years lost due to disability and $346.17 billion in expenditures. About 64.3 million people worldwide suffer from heart failure. The known prevalence of heart failure in developed countries is generally estimated to be 1 to 2% of the general adult population.[3]

**Hepatitis:** Hepatitis means inflammation of the liver. The liver is a vital organ that processes nutrients, filters the blood, and fights infections. When the liver is inflamed or damaged, its function can be affected. Heavy alcohol use, toxins, some medications, and certain medical conditions can cause hepatitis.[4] World Health Organization (WHO) estimates for 2019 (1): 296 million people worldwide are living with hepatitis B infection. 58 million people worldwide are infected with hepatitis C. 1.5 million people are newly infected with chronic hepatitis B. 1.5 million new infections. Chronic hepatitis C infection Both hepatitis B and hepatitis C can lead to lifelong infection. WHO estimates that in 2019, 1.1 million people died from these infections and their consequences, including liver cancer, cirrhosis, and other conditions caused by chronic viral hepatitis. Hepatitis A and E infections do not lead to chronic infections, but can be serious and can cause liver damage and death. Outbreaks of these infections occur worldwide, especially in some areas with poor sanitation.[5]

**Diabetes:** Diabetes is a chronic metabolic disease characterized by high levels of glucose (or blood sugar) that over time cause serious damage to the heart, blood vessels, eyes, kidneys, and nerves. The most common type of diabetes is type 2, which usually occurs in adults and occurs when the body becomes insulin resistant or does not produce enough insulin. Over the past 30 years, the prevalence of type 2 diabetes has skyrocketed in countries of all income levels. Type 1 diabetes, once known as juvenile diabetes or insulin-dependent diabetes, is a chronic disease in which the pancreas makes little or no insulin on its own. For people with diabetes, access to affordable treatment, including insulin, is critical to survival. There is a globally agreed-upon goal to stop the rise in diabetes and obesity by 2025.[6]

About 422 million people worldwide have diabetes, most of them living in low- and middle-income countries, and 1.5 million people are directly related to diabetes each year. The number and prevalence of diabetes have increased steadily over the past few decades.[7]

**Breast cancer:** Breast cancer is a serious and incurable disease that mainly affects women. The disease is caused by abnormal mutations in genes in normal cells that lead to the development of cancer cells. Breast cancer treatment can be very effective, especially if the disease is detected early. Treatment for breast cancer often consists of a combination of surgical removal, radiation therapy, and drugs (hormonal therapy, chemotherapy, and/or targeted biological therapy) to treat microscopic cancer that has spread through the blood from the breast cancer. Treatments that can prevent cancer from growing and spreading can save lives. Globally in 2020, 2.3 million women were diagnosed with breast cancer and 685,000 died. As of the end of 2020, 7.8 million women worldwide were diagnosed with breast cancer in the past five years, making it the most common cancer in the world.[8]

**Liver disease:** Liver disease kills approximately 2 million people worldwide each year, 1 million as a complication of cirrhosis, and 1 million as a result of viral hepatitis and carcinoma of hepatocellular. Cirrhosis is currently the 11th leading cause of death in the world, and liver cancer is the 16th leading cause of death. Together, they account for 3.5% of all deaths worldwide. Cirrhosis is one of the top 20 causes of disability-adjusted life expectancy, accounting for 1.6% and 2.1% of the global burden. Worldwide, approximately 2 billion people use alcohol, and over 75 million people are diagnosed with an alcohol use disorder and are at risk for alcohol-related liver disease.[9]

About 2 billion adults are obese or overweight, and more than 400 million people have diabetes. Both are increased risk factors for non-alcoholic fatty liver disease and hepatocellular carcinoma. Worldwide, the prevalence of hepatitis B (3.5%) and hepatitis C (1%) is high.[9]

## 1.4 Objective of this Work

This research work has a main goal, which is to develop a model for 'Medical Diseases Prediction using Machine Learning Algorithms'. The proposed system can be divided into three main parts.Chapter Three represents the overview of the system. The main three parts of the system are given below:

### 1.4.1 Data Collection:

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. In this case a dataset collected from **UCI Machine Learning Repository**. The data collected are trained and tested further for evaluating outcomes**.[11]**

### 1.4.2 Data Preprocessing:

**Data preprocessing** can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance. The phrase "garbage This chapter discusses various machine learning classifiers and previous work on heart disease. Researchers have a long history of working on the identification and prediction of various diseases through machine learning.**[12]** In this section, we have presented various research works on cancer, hepatitis, heart diseases, heart failure, diabetes, liver diseases' prediction. Previous studies have looked at the application of machine learning techniques in predicting and classifying different diseases. These studies concentrate on the specific impacts of specific ml techniques and not on optimizing these techniques using optimized methods. The exploration of supervised learning, unsupervised learning and reinforcement learning, find which is best for machine learning. There are extensive works in the field of classification, involving many models. In, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, and missing values, etc. Analysing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running any analysis. Often, data preprocessing is the most important phase of a machine learning project, especially in computational biology.

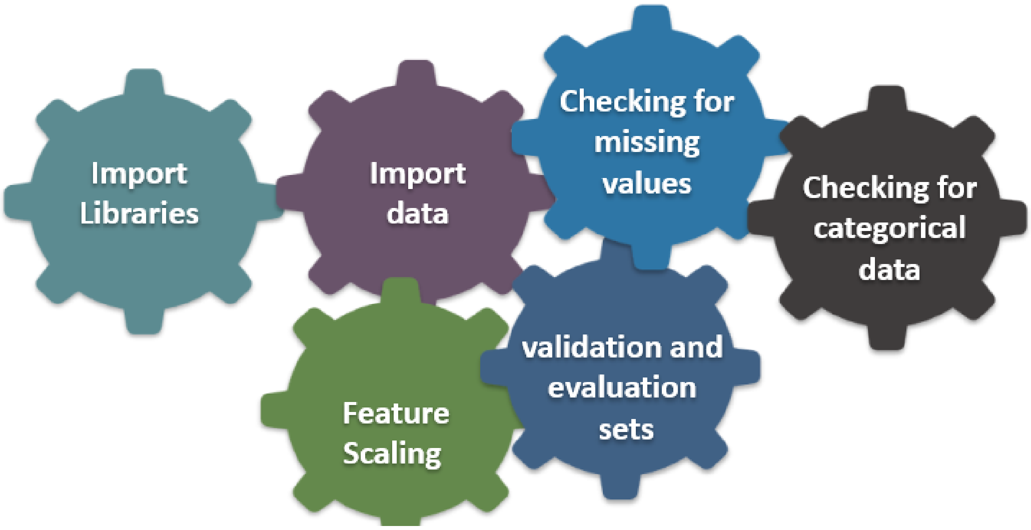
Import Libraries
Import data
Checking for missing values
Checking for categorical data
Feature Scaling
validation and evaluation sets

**Fig 1.1: Data Preprocessing Process**

# Chapter 02: Literature Review

## 2.2. Literature Review :

This chapter discusses various machine learning classifiers and previous work on heart disease. Researchers have a long history of working on the identification and prediction of various diseases through machine learning. In this section, we have presented various research works on cancer, hepatitis, heart diseases, heart failure, diabetes, liver diseases prediction. Previous studies have looked at the application of machine learning techniques in predicting and classifying different diseases. These studies concentrate on the specific impacts of specific ml techniques and not on optimizing these techniques using optimized methods. The exploration of supervised learning, unsupervised learning and reinforcement learning, find which is best for machine learning. There are extensive works in the field of classification involving many models.

Classifying is one of the most important tasks in ML.
From previous work we can see the accuracy of heart disease (Cleveland ) for Linear regression, Naïve Bayes, Support vector machine, Decision tree, Random forest and K-Nearest neighbor are respectively 93.40%,90.10%,92.30%,81.31%,95.60% and 71.42%.[13][14][15]

Respectively, The accuracy of Breast Cancer Coimbra for Naïve Bayes, Support vector machine, Decision tree, Random forest and K-Nearest neighbour are respectively 62.38%, 72.52%, 69.28%, 70.31% and 58.14%.[16][17][18]

The accuracy of Liver diseases for Linear regression ,Naïve Bayes, Support vector machine, Decision tree, Random forest and K-Nearest are respectively 75%,53%,64%,69%,74% and 62%.[19]

And For the breast cancer data set, Liu et al. used a Decision Table (DT) predictive model for breast cancer survival and concluded that the patient survival rate was 86.52%. Results (based on datasets with mean precision for breast cancer) showed that the naive Bayesian method was the best predictor, with an accuracy of 97.36% for the retained sample (this predictive accuracy is better than described in the literature).[20]

# Chapter 03: Methodology

## 3.1. Data analysis

Data analysis is defined as a process of cleaning, transforming, and modelling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and take the decision based upon the data analysis.

The Datasets are **Heart disease, Breast Cancer Coimbra, Hepatitis, Wisconsin Breast cancer diagnostic, Heart Failure disease, Diabetes, Liver disease.** All the datasets are collected from UCI Machine learning repository.

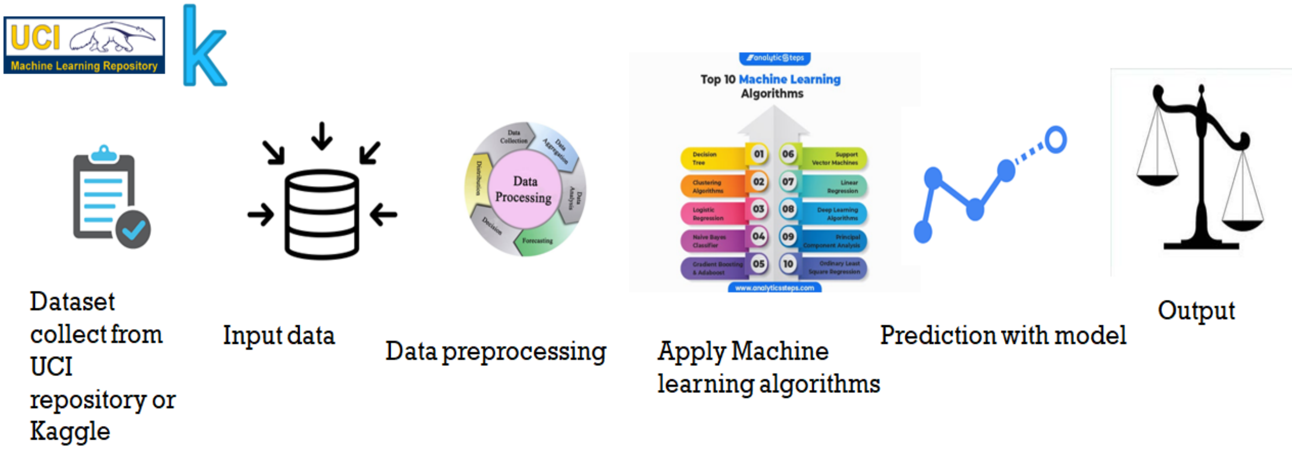
UCI Machine Learning Repository
Dataset collect from UCI repository or Kaggle
Input data
Data Processing
Data preprocessing
Top 10 Machine Learning Algorithms
Apply Machine learning algorithms
Prediction with model
Output

Fig 2.1: Overall Mechanism Process

**Heart diseases Dataset[22]** data classified if patients have heart disease or not according to features in it. We will try to use this data to create a model which tries to predict if a patient has this disease or not.[21]

Data contains;

- age - age in years
- sex - (1 = male; 0 = female)
- cp - chest pain type

- trestbps - resting blood pressure
- chol - serum cholesterol in mg/DL
- fbs - (fasting blood sugar > 120 mg/DL) (1 = true; 0 = false)
- restecg - resting electrocardiographic results
- thalach - maximum heart rate achieved
- exang - exercise induced angina (1 = yes; 0 = no)
- oldpeak - ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment
- ca - number of major vessels (0-3) coloured by fluoroscopy
- thal - 3 = normal; 6 = fixed defect; 7 = reversible defect
- target - have disease or not (1=yes, 0=no)

Data Exploration : target

| | |
|---|---|
| 1 | 165 |
| 0 | 138 |

**Breast Cancer Coimbra Dataset[23]** data, which classified if patients have a disease or not according to features in it. We will try to use this data to create a model which tries to predict if a patient has this or not.[23]

Data contains :

- Age
- Bmi
- Glucose
- Insulin
- Homo
- Leptin
- Adiponectin
- Resistin
- Classification

Data Exploration: classification

1=Healthy controls= 52

2=Patients=64

**Hepatitis Dataset[24]** data, which classified if patients have a disease or not according to features in it. We will try to use this data to create a model which tries to predict if a patient has this or not.[24]

Data contains:

- age
- sex
- steroid
- antiviral
- fatigue
- malaise
- anorexia
- liver_big
- liver_firm
- spleen_palpable
- spiders
- ascites
- varices
- bilirubin
- alk_phospet
- albumin
- protein
- histology
- class

Data Exploration: histology

| | |
|---|---|
| 1 | 123 |
| 0 | 32 |

**Wisconsin Breast cancer diagnostic[25]** data, which classified if patients have heart disease or not according to features in it. We will try to use this data to create a model which tries to predict if a patient has this death or not.

- Id
- Diagnosis
- Radius mean
- Texture mean
- Perimeter mean
- Area mean
- Smoothness mean
- Compactness_mean
- Concavity_mean
- concave points_mean
- Symmetry_mean
- Fractal_dimension_mean
- Radius_se
- Texture_se
- Perimeter_se
- Area_se
- Smoothness se
- Compactness_se
- Concavity_se
- concave points_se
- Symmetry_se
- Fractal_dimension_se
- Radius worst
- Texture worst
- Perimeter worst
- Area worst
- Smoothness worst
- Compactness_worst
- Concavity_worst
- concave points_worst
- Symmetry_worst
- Fractal_dimension_worst

Data Exploration: diagnosis

```
0   357
1   212
```

**Heart Failure clinical records dataset[26]** data, which classified if patients have heart disease or not according to features in it. We will try to use this data to create a model which tries to predict if a patient has this death or not.[26]

Data Contains:

- age
- anaemia
- creatinine_phosphokinase
- diabetes
- ejection_fraction
- high_blood_pressure
- platelets
- serum_creatinine
- serum_sodium
- sex
- smoking
- time

Data Exploration : DEATH_EVENT

```
0   203
1   96
```

**Diabetes Datasets[27]** data which classified if patients have disease or not according to features in it. We will try to use this data to create a model which tries to predict if a patient has this disease or not.[27]

Data Contains:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age
- sex
- Outcome

Data Exploration : Outcome

```
0   500
1   268
```

**Liver Disease Datasets[28]** This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patients (liver disease) or not (no disease)[28].

Data contains:

- Gender
- Total_Bilirubin
- Direct_Bilirubin
- Alkaline_Phosphotase
- Alamine_Aminotransferase
- Aspartate_Aminotransferase
- Total_Protiens
- Albumin
- Albumin_and_Globulin_Ratio
- Dataset

Data Exploration : Dataset

```
1   416
2   167
```

## 3.2. Data Preprocessing process

Data preprocessing is the process of transforming raw data into an understandable format. The quality of the data should be checked before applying machine learning or data mining algorithms[12]

### 3.2.1 Feature Encoding

Machine learning models can only work with numerical values. For this reason, it is necessary to **transform the categorical values of the relevant features into numerical** ones. This process is called feature encoding.

Categorical features are generally divided into 3 types:

1. **Binary: Either/or**

   ***Examples:***

   - **Yes, No**
   - **True, False**

2. **Ordinal: Specific ordered groups.**

   ***Examples:***

   - **low, medium, high**
   - **cold, hot, lava Hot**

3. **Nominal: Unordered Groups.**

   ***Examples***

   - **cat, dog, tiger**
   - pizza, burger, coke

Specifically: That categorical data is defined as variables with a finite set of label values. That most machine learning algorithms require numerical input and output variables. That an integer and one hot encoding is used to convert categorical data to integer data.[29]

### 3.2.2 Missing Value Handle

Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null as, the entire column can be dropped. The rows which have one or more column values as null can also be dropped.[30]

### 3.2.3 Over Sampling

Oversampling and undersampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented). These terms are used both in statistical sampling, survey design methodology and in machine learning.[31]

### 3.2.4 Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Feature scaling is essential for machine learning algorithms that calculate distances between data.[32]

Method of feature scaling are

**Rescaling (min-max normalization)**

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

**Mean normalization**

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

**Standardization (Z-score Normalization)**

$$x' = \frac{x - \bar{x}}{\sigma}$$

### 3.2.5 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.[33]

There are three types of feature selection:

- Wrapper methods (forward, backward, and stepwise selection),
- Filter methods (ANOVA, Pearson correlation, variance threshold ing),
- Embedded methods (Lasso, Ridge, Decision Tree).

## 3.3. Classification Algorithm

### 3.3.1 Linear Regression Algorithm:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.[34]

In simple words, the dependent variable is binary in nature, having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).[34]

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.[34]

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value of 0 and 1, it gives the probabilistic values which lie between 0 and 1.[34]

**Logistic Regression Equation:**

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y=b0 + b1x1 + b2x2 + b3x3 +.... + bnxn$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y)

$$y / 1-y \ ; 0 \text{ for } y=0 \text{ , and infinity for } y=1$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$\log[y / 1-y ] = b0 + b1x1 + b2x2 + b3x3 + ..... + bnxn$$

The above equation is the final equation for Logistic Regression.[34]

**Advantages and Disadvantages:**

**Advantages:**

1. Logistic regression is easier to implement, interpret, and very efficient to train.

2. It makes no assumptions about distributions of classes in feature space.

3. It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.

4. It is very fast at classifying unknown records.[35]

**Disadvantages:**

1. If the number of observations is lesser than the number of features, Logistic Regression should not be used otherwise, it may lead to overfitting.

2. It constructs linear boundaries

3. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

4. Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.[35 ]

### 3.3.2. Decision Tree Algorithm:

In general, Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decision trees are the most powerful algorithms that fall under the category of supervised algorithms.

They can be used for both classification and regression tasks. The two main entities of a tree are decision nodes, where the data is split and leaves, where we get the outcome. The example of a binary tree for predicting whether a person is fit or unfit providing various information like age, eating habits and exercise habits, is given below −

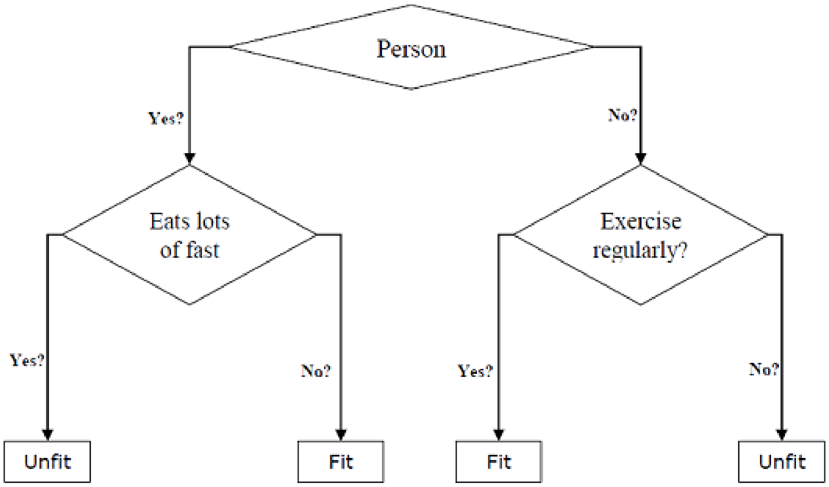

Fig 3.1 : decision tree

In the above decision tree, the questions are decision nodes and final outcomes are leaves. We have the following two types of decision trees −

- Classification decision trees − In this kind of decision trees, the decision variable is categorical. The above decision tree is an example of a classification decision tree.
- Regression decision trees − In this kind of decision trees, the decision variable is continuous.[36]

**Advantages:**

1. Compared to other algorithms, decision trees require less effort for data preparation during pre-processing.
2. A decision tree does not require normalization of data.
3. A decision tree does not require scaling of data as well.
4. Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
5. A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.[36]

**Disadvantage:**

1. A small change in the data can cause a large change in the structure of the decision tree, causing instability.
2. For a Decision tree, sometimes calculation can go far more complex compared to other algorithms.
3. Decision trees often involve higher time to train the model.
4. Decision tree training is relatively expensive as the complexity and time taken are more.
5. The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.[36]

### 3.3.3 Random Forest Algorithm:

Random forest is a supervised learning algorithm which is used for both classification and regression. But it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, a random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method, which is better than a single decision tree because it reduces the over-fitting by averaging the result.

### Working of Random Forest Algorithm:

We can understand the working of Random Forest algorithm with the help of following steps −

- Step 1 − First, start with the selection of random samples from a given dataset.
- Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- Step 3 − In this step, voting will be performed for every predicted result.
- Step 4 − At last, select the most voted prediction result as the final prediction result.[37]

The following diagram will illustrate its working –

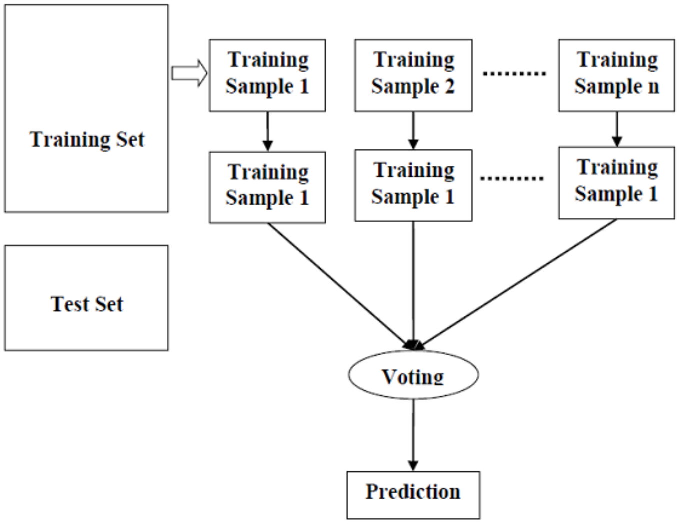

Fig 3.2: Random forest

### Pros and Cons of Random Forest

### Pros:

The following are the advantages of Random Forest algorithm −

1. It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
2. Random forests work better for a large range of data items than a single decision tree does.
3. Random forest has less variance than a single decision tree.
4. Random forests are very flexible and possess very high accuracy.

5. Scaling of data does not require a random forest algorithm. It maintains good accuracy even after providing data without scaling.
6. Random Forest algorithms maintain good accuracy even if a large proportion of the data is missing.[37]

**Cons:**

The following are the disadvantages of Random Forest algorithm −

1. Complexity is the main disadvantage of Random forest algorithms.
2. Construction of Random forests is much harder and time-consuming than decision trees.
3. More computational resources are required to implement the Random Forest algorithm.
4. It is less intuitive when we have a large collection of decision trees.
5. The prediction process using random forests is very time-consuming in comparison with other algorithms.[37]

### 3.3.4 Naive Bayes Algorithm

Naive Bayes algorithm is a classification technique based on applying Bayes' theorem with a strong assumption that all the predictors are independent to each other. In simple words, the assumption is that the presence of a feature in a class is independent of the presence of any other feature in the same class. For example, a phone may be considered smart if it has a touch screen, internet facility, good camera etc. Though all these features are dependent on each other, they contribute independently to the probability that the phone is a smartphone.

In Bayesian classification, the main interest is to find the posterior probabilities i.e. the probability of a label given some observed features, $P(L \mid features)$. With the help of Bayes theorem, we can express this in quantitative form as follows −

$$\mathrm{P(L|features)=P(L)P(features|L)P(features)P(L|features)=P(L)P(features|L)}P(features)$$

Here, $P(L \mid features)$ is the posterior probability of class.

$P(L)$ is the prior probability of class.

$P(features \mid L)$ is the likelihood, which is the probability of predictor given class.

$P(features)$ is the prior probability of predictor.[38]

Pros & Cons

## Pros

The followings are some pros of using Naive Bayes classifiers −

1. Naive Bayes classification is easy to implement and fast.
2. It will converge faster than discriminative models like logistic regression.
3. It requires less training data.
4. It is highly scalable in nature, or they scale linearly with the number of predictors and data points.
5. It can make probabilistic predictions and can handle continuous as well as discrete data.[39]

The Naïve Bayes classification algorithm can be used for binary as well as multi-class classification problems.

## Cons

The followings are some cons of using Naïve Bayes classifiers −

1. One of the most important cons of Naïve Bayes classification is its strong feature independence because in real life it is almost impossible to have a set of features which are completely independent of each other.
2. Another issue with Naïve Bayes classification is its ‘zero frequency’ which means that if a categorical variable has a category but not being observed in the training data set, then the Naïve Bayes model will assign a zero probability to it, and it will be unable to make a prediction.[39 ]

## Applications of Naïve Bayes classification

The following are some common applications of Naïve Bayes classification −

Real-time prediction − Due to its ease of implementation and fast computation, it can be used to do prediction in real-time.

Multi-class prediction − Naïve Bayes classification algorithm can be used to predict posterior probability of multiple classes of target variable.

Text classification − Due to the feature of multi-class prediction, Naïve Bayes classification algorithms are well suited for text classification. That is why it is also used to solve problems like spam-filtering and sentiment analysis.

Recommendation system − Along with the algorithms like collaborative filtering, Naïve Bayes makes a Recommendation system which can be used to filter unseen information and to predict whether a user would like the given resource or not.[38]

### 3.3.5 Support Vector Machine Algorithm:

SVM is the foremost broadly utilized ML technique-based design classification strategy accessible these days. It is based on a factual learning hypothesis and was developed by Vapnik within the year 1995. The essential point of this strategy is to venture nonlinear distinguishable tests onto another higher dimensional space by utilizing diverse sorts of part capacities.[40]

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.[40]

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine. Consider the below diagram, in which there are two different categories that are classified using a decision boundary or hyperplane:[40]

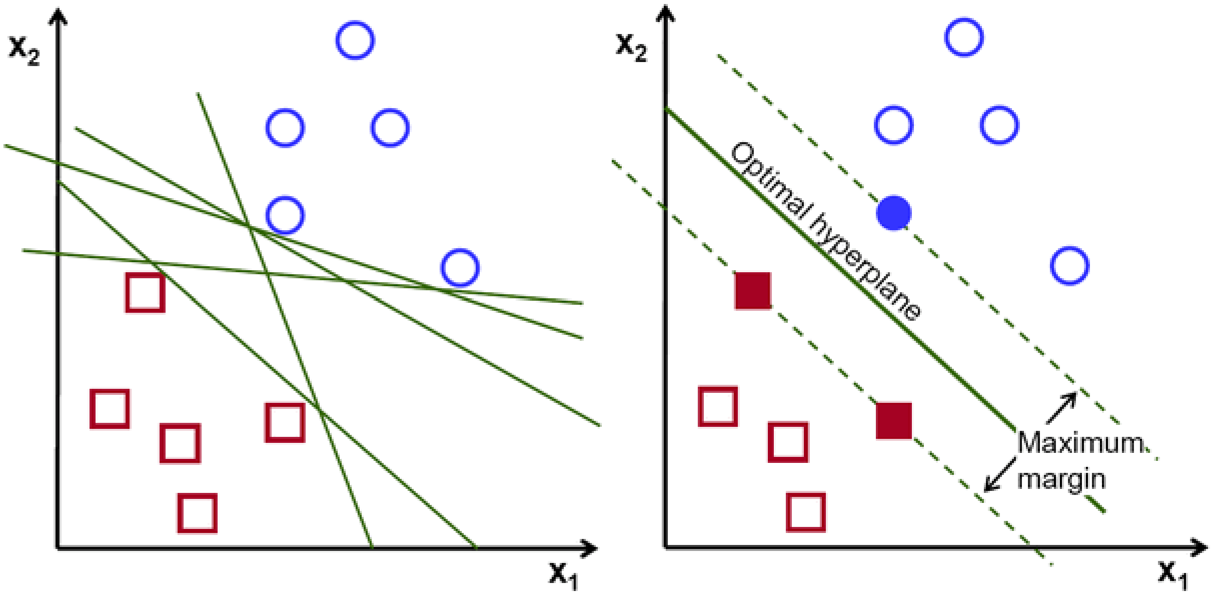

Fig 3.3: Support vector machine hyperplane

**Types of SVM:**

**SVM can be of two types:**

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and a classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and the classifier used is called a Non-linear SVM classifier. [40 ]

**Pros:**

1. It works really well with a clear margin of separation
2. It is effective in high-dimensional spaces.
3. It is effective in cases where the number of dimensions is greater than the number of samples.
4. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.[41]

**Cons:**

It doesn't perform well when we have large data set because the required training time is higher

It also doesn't perform very well, when the data set has more noise i.e. target classes are overlapping

SVM doesn't directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is included in the related SVC method of the Python Scikit-learn library.[41]

### 3.3.6 K-Nearest Neighbour Algorithm

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. It assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears, then it can be easily classified into a well-suited category by

using the K- NN algorithm. It can be used for Regression as well as for Classification, but mostly it is used for Classification problems. It is a non-parametric algorithm, which means it does not make any assumptions on underlying data.
It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
K-NN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. [42]

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbours
- **Step-2:** Calculate the Euclidean distance of **K number of neighbours**
- **Step-3:** Take the K nearest neighbours as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbours, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbour is maximum.
- **Step-6:** Our model is ready.[43 ]

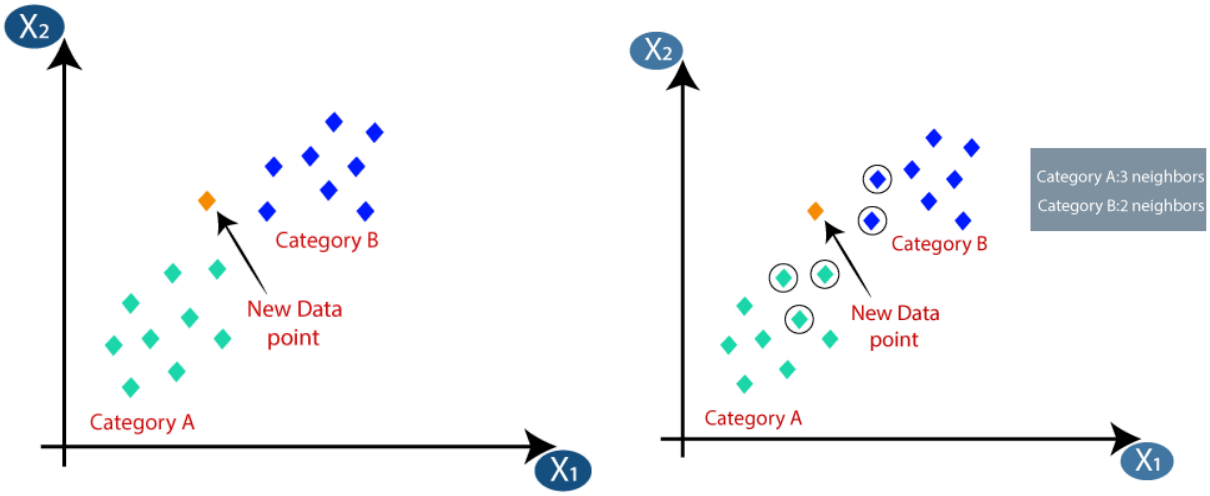

Fig 3.4: K-NN  model

**Advantages of KNN Algorithm:**

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.[43]

**Disadvantages of KNN Algorithm:**

- Always needs to determine the value of K, which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.[43]

## 3.4. Architecture of the model:

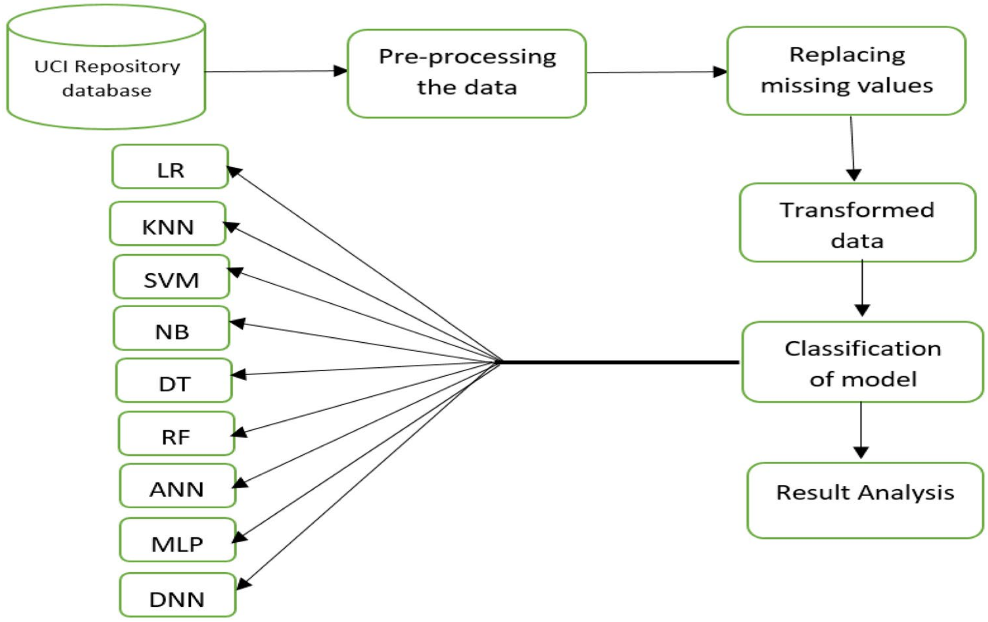

Fig 3.5: Architecture view

## 3.5. Evaluation Measure:

The model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data. Methods for evaluating a model's performance are divided into 2 categories: namely, holdout and Cross-validation. Both methods use a test set to evaluate model performance.

Various ways to evaluate a machine learning model's performance

- **Confusion matrix.**
- **Accuracy.**
- **Precision.**
- **Recall.**
- **Specificity.**
- **F1 score.**

- **Precision-Recall or PR curve.**
- **ROC (Receiver Operating Characteristics) curve.**

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

The three main metrics used to evaluate a classification model are accuracy, precision, and recall.[44]

### Classification Accuracy

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of the number of correct predictions to the total number of input samples. It works well only if there are an equal number of samples belonging to each class.[45]

Here, we use only accuracy for prediction.

**Accuracy:** Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.[45]

Accuracy= correct predictions / all predictions

**Precision:** Precision is defined as the fraction of relevant examples (true positives) among examples which were predicted to belong to a certain class.[45]

Precision= true positives / (true positives + false positives)

**Recall :** Recall is defined as the fraction of examples which were predicted to belong to a class with respect to all the examples that truly belong in the class.[45]

Recall=true positives / (true positives + false negatives)

# Chapter 04: Result

## 4.1 Technology.

Machine learning technology is to optimize the performance of a system when handling new instances of data through user-defined programming logic for a given environment. To accomplish this goal effectively and efficiently, machine learning draws heavily on statistics and computer science.

Here, we used Pandas to visualize the data, and we used Python, Implemented the Algorithm on the Jupyter notebook. Our Python version was python 3.8, and we use Jupyter notebook 6.1.4. After selecting the dataset, we use pandas profiling for analysis of the dataset. In our datasets, there were Some have missing value and no missing values and no duplicate rows. Again, we analyse and visualize our dataset in Jupyter notebook. In Jupyter notebook we process our dataset, implement and find the result according to the machine learning algorithm of different methods.

Implementation Tools:

- **Python 3.8 64-bit version**
- **Pycharm**
- **Jupyter Notebook**

### Table 4.2 Accuracy comparison in different datasets & algorithms with Model

| Algorithm/ Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| Heart disease | 80.5977 | 81.2183 | 80.2874 | 80.8953 | 79.1223 | 81.5159 |
| Breast Cancer Coimbra | 54.5948 | 59.0642 | 60.0668 | 57.7276 | 58.6466 | 58.6883 |
| Hepatitis | 75.1927 | 78.4467 | 82.9365 | 83.3333 | 91.0884 | 91.8707 |
| Wisconsin Breast cancer diagnostic | 96.7787 | 93.5574 | 93.9775 | 96.7787 | 95.2380 | 97.7591 |
| Heart Failure disease | 77.0935 | 73.3990 | 71.1822 | 79.5566 | 82.5123 | 83.7438 |
| Diabetes | 71.7887 | 71.887 | 75.9951 | 82.7172 | 82.7172 | 86.7321 |
| Liver disease | 64.3091 | 67.3418 | 66.963 | 65.8870 | 84.2737 | 82.8331 |

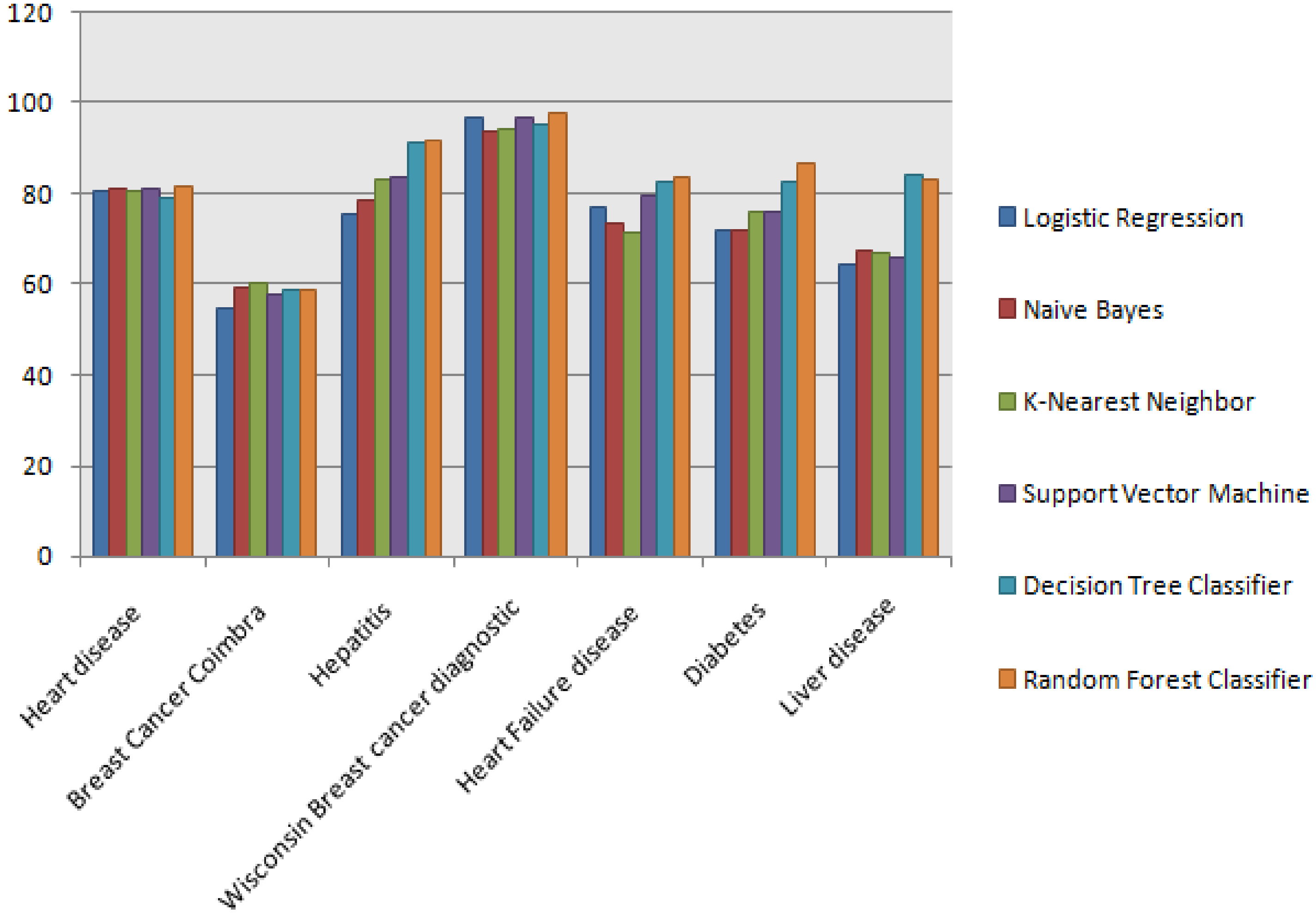

Fig 4.1: Chart column comparison in different datasets & algorithms with Model

**Table 4.3: Accuracy comparison in different datasets & algorithms without Model**

| Algorithm/ Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| Heart disease | 77.00 | 79.00 | 59.00 | 77.00 | 74.00 | 80.00 |
| Breast Cancer Coimbra | 34.00 | 64.00 | 73.00 | 64.00 | 89.00 | 98.00 |
| Hepatitis | 77.00 | 65.00 | 71.00 | 77.00 | 81.00 | 87.00 |
| Wisconsin Breast cancer diagnostic | 79.00 | 83.00 | 58.00 | 83.00 | 70.00 | 79.00 |
| Heart Failure disease | 90.00 | 90.00 | 75.00 | 85.00 | 85.00 | 85.00 |
| Diabetes | 70.00 | 75.00 | 60.00 | 70.00 | 70.00 | 75.00 |
| Liver disease | 70.00 | 60.00 | 65.00 | 70.00 | 65.00 | 75.00 |

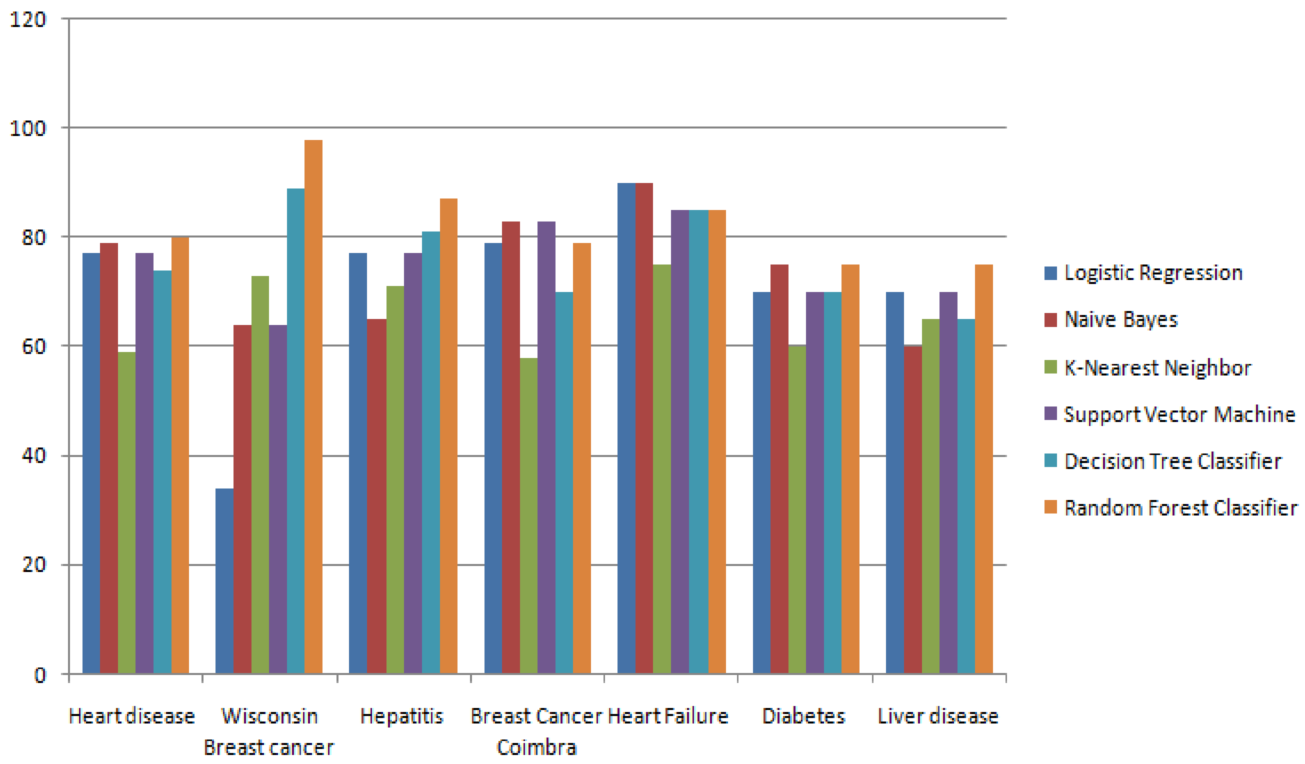

Fig 4.2: Chart column comparison in different datasets & algorithms without Model

## 4.4:Comparison between Row And Model

In Heart disease Dataset Row data set accuracy and model trained data set both of these the best outcome comes from Random forest Algorithm.

**Heart Disease Accuracy for Model :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Heart disease** | **80.5977** | **81.2183** | **80.2874** | **80.8953** | **79.1223** | **81.5159** |

**Heart Disease Accuracy for Raw :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Heart disease** | **77.00** | **79.00** | **59.00** | **77.00** | **74.00** | **80.00** |

Model dataset accuracy : 81.5159
Row dataset accuracy : 80.00

In **Breast Cancer Coimbra Dataset** Row data set the best accuracy given **Random forest** and model trained data set the best outcome comes from **K-NN Algorithm**

**Breast Cancer Coimbra Accuracy for Model :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Breast Cancer Coimbra** | **54.5948** | **59.0642** | **60.0668** | **57.7276** | **58.6466** | **58.6883** |

### Breast Cancer Coimbra Accuracy for Raw :

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| Breast Cancer Coimbra | 34.00 | 64.00 | 73.00 | 64.00 | 89.00 | 98.00 |

Model dataset accuracy : 60.0668
Row dataset accuracy : 98.00

In **Hepatitis Dataset** Row data set accuracy and model trained data set both of this the best outcome comes from **Random forest Algorithm**

### Hepatitis Accuracy for Model :

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| Hepatitis | 75.1927 | 78.4467 | 82.9365 | 83.3333 | 91.0884 | 91.8707 |

### Hepatitis Accuracy for Raw :

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| Hepatitis | 77 | 65 | 71 | 77 | 81 | 87 |

Model dataset accuracy : 91.8707
Row dataset accuracy : 87.00

In **Wisconsin Breast cancer diagnostic Dataset** Row data set given the best accuracy from **Gaussian Naive Bayes and Support Vector Machine** and model trained data set the best outcome comes from **Random forest Algorithm**

**Wisconsin Breast cancer Accuracy for Model :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Wisconsin Breast cancer** | **96.7787** | **93.5574** | **93.9775** | **96.7787** | **95.2380** | **97.7591** |

**Wisconsin Breast cancer Accuracy for Raw :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Wisconsin Breast cancer diagnostic** | **79.00** | **83.00** | **58.00** | **83.00** | **70.00** | **79.00** |

Model dataset accuracy : 97.7591
Row dataset accuracy : 83.00

In **Heart Failure disease Dataset** Row data set the best accuracy from **Logistic Regression and Gaussian Naive Bayes** and model trained data set the best outcome comes from **Random forest Algorithm**

**Heart Failure disease Dataset Model :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Heart Failure disease** | **77.0935** | **73.3990** | **71.1822** | **79.5566** | **82.5123** | **83.7438** |

**Heart Failure disease Dataset Raw:**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Heart Failure disease** | **90.00** | **90.00** | **75.00** | **85.00** | **85.00** | **85.00** |

Model dataset accuracy : 83.7438
Row dataset accuracy : 90.00

In **Diabetes Dataset** Row data set the best accuracy given **Random forest** and model trained data set the best outcome comes from **Random Forest Algorithm**

**Diabetes Dataset Model :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Diabetes** | **71.7887** | **71.887** | **75.9951** | **82.7172** | **82.7172** | **86.7321** |

**Diabetes Dataset Raw :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Diabetes** | **70.00** | **75.00** | **60.00** | **70.00** | **70.00** | **75.00** |

Model dataset accuracy : 86.7321
Row dataset accuracy : 75.00

In **Liver Disease Dataset** Row data set the best accuracy given **Random forest** and model trained data set the best outcome comes from **Decision Tree Classifier Algorithm**

**Liver Disease Dataset Model :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Liver disease** | **64.3091** | **67.3418** | **66.963** | **65.8870** | **84.2737** | **82.8331** |

**Liver Disease Dataset Raw :**

| Algorithm / Datasets | Logistic Regression | Gaussian Naive Bayes | K-Nearest Neighbour | Support Vector Machine | Decision Tree Classifier | Random Forest Classifier |
|---|---|---|---|---|---|---|
| **Liver disease** | **70.00** | **60.00** | **65.00** | **70.00** | **65.00** | **75.00** |

Model dataset accuracy : 84.2737
Row dataset accuracy : 75.00

# Chapter 05: Discussion & Limitation

## 5.1 Discussion

The prediction of disease at an earlier stage becomes an important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctors. The correct prediction of disease is the most challenging task. To overcome this problem, data mining plays an important role in predicting disease. Medical science has a large amount of data growth per year. Due to an increased amount of data growth in the medical and healthcare field, the accurate analysis of medical data, which has been benefited from early patient care.[46]

The use of different Ml Algorithm enabled the early detection of many maladies such as heart, kidney, breast cancer, liver etc. Throughout the literature SVM, RF, LR and NV algorithms were the most widely used for prediction, while accuracy was the most used performance metric.

## 5.2 Limitation

The limitations of a study are its flaws or shortcomings which could be the result of unavailability of resources, small sample size, flawed methodology, etc. No study is completely flawless or inclusive of all possible aspects.

It is for sure that our work will have some limitations, and it is normal. There can be some limitations in every work. However, it is critically important for us to be striving to minimize the range of scope of limitations throughout the research process

In our work, we worked with different kinds of datasets and around seven algorithms. Some of them work properly. While working with Datasets, we got the most accurate prediction (almost 90%). But we have datasets whose accuracy is much lower.

# Chapter 06: Conclusion & Future work

## 6.1 Conclusion

The use of different ML algorithms enabled the early detection of many maladies such as heart, breast, diabetes and liver diseases. Throughout the literature, LR, NB, DT, SVM, RF and LR algorithms were the most widely used at prediction, while accuracy was the most used performance metric. The model proved to be the most adequate at predicting common diseases.

- The result achieved by the Model for classification on the Medical Diseases Prediction using Machine Learning Algorithms.
- The main objective of our thesis is to distinguish the medical diseases between the trained model and without model

## 6.2 Future work

As we can see from our work, we got a better result with Model trained datasets against raw datasets. We will work with this in the future to overcome the overfitting problem.

In future work, the creation of more complex Ml Algorithms is much needed to increase the efficiency of disease prediction. In addition, learning models should be calibrated more often after the training phase for potentially a better performance

Finally, more relevant feature selection methods should be used to enhance the performance of the learning models.

## References

[1] *10 Best Healthcare Data Sets (examples)*. Cprime Studios. (2021, October 22). Retrieved November 5, 2021, from https://cprimestudios.com/blog/10-best-healthcare-data-sets-examples.

[2] Centers for Disease Control and Prevention. (2021, July 19). *Coronary artery disease*. Centers for Disease Control and Prevention. Retrieved November 5, 2021, from https://www.cdc.gov/heartdisease/coronary_ad.htm.

[3] World Health Organization. (n.d.). Cardiovascular diseases (cvds). World Health Organization. Retrieved November 5, 2021, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[4] NHS. (n.d.). Heart failure - NHS. NHS choices. Retrieved November 5, 2021, from https://www.nhs.uk/conditions/heart-failure/#:~:text=Heart%20failure%20means%20that%20the,your%20heart%20has%20stopped%20working.

[5] Centers for Disease Control and Prevention. (2020, July 28). What is viral hepatitis? Centers for Disease Control and Prevention. Retrieved November 5, 2021, from https://www.cdc.gov/hepatitis/abc/index.htm#:~:text=Hepatitis%20means%20inflammation%20of%20the,medical%20conditions%20can%20cause%20hepatitis.

[6] Centers for Disease Control and Prevention. (2021, July 19). Global viral hepatitis: Millions of people are affected. Centers for Disease Control and Prevention. Retrieved November 5, 2021, from https://www.cdc.gov/hepatitis/global/index.htm.

[7]https://www.paho.org/en/topics/diabetes#:~:text=Diabetes%20is%20a%20chronic%2C%20metabolic,and%2For%20action%20of%20insulin.

[8]World Health Organization. (n.d.). Diabetes. World Health Organization. Retrieved November 5, 2021, from https://www.who.int/health-topics/diabetes.

[9]World Health Organization. (n.d.). Breast cancer. World Health Organization. Retrieved November 5, 2021, from https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[10]Asrani, S. K., Devarbhavi, H., Eaton, J., &amp; Kamath, P. S. (2019). Burden of liver diseases in the world. Journal of Hepatology, 70(1), 151–171.

[11] UCI Machine Learning Repository. (n.d.). Retrieved November 5, 2021, from https://archive.ics.uci.edu/ml/index.php.

[12] *Data preprocessing in data mining -A hands on guide*. Analytics Vidhya. (2021, August 10). Retrieved November 5, 2021, from https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/.

[13] Katarya, R., & Srinivas, P. (2020). Predicting heart disease at early stages using Machine Learning: A Survey. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. https://doi.org/10.1109/icesc48915.2020.9155586

[14] Gavhane, A., Kokkula, G., Pandya, I., &amp; Devadkar, K. (2018). Prediction of heart disease using machine learning. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). https://doi.org/10.1109/iceca.2018.8474922

[15]S. R. Tithi, A. Aktar, F. Aleem and A. Chakrabarty, "ECG data analysis and heart disease prediction using machine learning algorithms," *2019 IEEE Region 10 Symposium (TENSYMP)*, 2019, pp. 819-824

[16]*AnALysis of breast CANCER DETECTION ...* (n.d.). Retrieved November 5, 2021, from https://www.researchgate.net/profile/Aman-Jatain/publication/342640935_Analysis_of_Breast_Cancer_Detection_Techniques_Using_RapidMiner/links/60769bb7299bf1f56d563895/Analysis-of-Breast-Cancer-Detection-Techniques-Using-RapidMiner.pdf.

[17] *Comparison of machine learning algorithms in breast cancer ...* (n.d.). Retrieved November 5, 2021, from https://www.researchgate.net/publication/337193772_Comparison_of_Machine_Learning_Algorithms_in_Breast_Cancer_Prediction_Using_the_Coimbra_Dataset.

[18]Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*. June 2018:119-126.

[19] Rahman, A. K. M. & Shamrat, F.M. & Tasnim, Zarrin & Roy, Joy & Hossain, Syed. (2019). A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. 8. 419-422.

[20]Liu, Y.-Q., Wang, C., &amp; Zhang, L. (2009). Decision tree based predictive models for breast cancer survivability on imbalanced data. 2009 3rd International Conference on Bioinformatics and Biomedical Engineering.

[21] Ronit. (2018, June 25). *Heart disease UCI*. Kaggle. Retrieved November 5, 2021, from https://www.kaggle.com/ronitf/heart-disease-uci.

[22] Heart disease classification using machine learning algorithms. (2021). *Strad Research*, *8*(1). https://doi.org/10.37896/sr8.1/035

[23]  UCI Machine Learning Repository: Breast Cancer Coimbra Data Set. (n.d.). Retrieved November 5, 2021, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra?fbclid=IwAR23HHS9o2ZaRORviQPISnv4U6hPYENO4KQNYTRFjbdDRfOcIpgPOJnSGag.

[24] UCI Machine Learning Repository: Hepatitis Data Set. (n.d.). Retrieved November 5, 2021, from https://archive.ics.uci.edu/ml/datasets/hepatitis.

[25] UCI Machine Learning Repository: Breast Cancer wisconsin (diagnostic) data set. (n.d.). Retrieved November 5, 2021, from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic

[26] Larxel. (2020, June 20). *Heart failure prediction*. Kaggle. Retrieved November 5, 2021, from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data.

[27] Learning, U. C. I. M. (2016, October 6). *Pima Indians Diabetes Database*. Kaggle. Retrieved November 5, 2021, from https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[28] Sanjames. (2018, February 27). *Liver patients analysis, Prediction & Accuracy*. Kaggle. Retrieved November 5, 2021, from https://www.kaggle.com/sanjames/liver-patients-data.

[29] *Feature encoding techniques - machine learning*. GeeksforGeeks. (2021, August 6). Retrieved November 5, 2021, from https://www.geeksforgeeks.org/feature-encoding-techniques-machine-learning/.

[30] Kumar, S. (2021, September 28). *7 ways to handle missing values in machine learning*. Medium. Retrieved November 5, 2021, from https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.

[31] Wikimedia Foundation. (2021, October 26). *Oversampling and undersampling in data analysis*. Wikipedia. Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis.

[32] Wikimedia Foundation. (2021, September 21). *Feature scaling*. Wikipedia. Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Feature_scaling.

[33] singh, R. (2021, February 17). *Implementing feature selection methods for machine learning*. Medium. Retrieved November 5, 2021, from https://ranasinghiitkgp.medium.com/implementing-feature-selection-methods-for-machine-learning-bfa2e4b4e02.

[34]*Machine Learning - Logistic Regression - Tutorialspoint*. Machine learning - logistic regression. (n.d.). Retrieved November 3, 2021, from https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm.

[35] Rout, A. R. (2020, September 2). Advantages and disadvantages of logistic regression. GeeksforGeeks. Retrieved November 4, 2021, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/.

[36] K, D. (2020, December 26). Top 5 advantages and disadvantages of Decision Tree Algorithm. Medium. Retrieved November 5, 2021, from https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a.

[37] Machine Learning - Random Forest- Tutorialspoint. Machine learning - random forest. (n.d.). Retrieved November 3, 2021, from https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm.

[38] Learn naive Bayes algorithm: Naive Bayes classifier examples. Analytics Vidhya. (2021, August 26). Retrieved November 5, 2021, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[39] admin2, admin2. (1970, March 4). Classification algorithms - naïve bayes. Prutor Online Academy (developed at IIT Kanpur). Retrieved November 5, 2021, from https://prutor.ai/classification-algorithms-naive-bayes/.

[40]Support Vector Machine (SVM) algorithm - javatpoint. www.javatpoint.com. (n.d.). Retrieved November 5, 2021, from https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[41]SVM: Support Vector Machine Algorithm in machine learning. Analytics Vidhya. (2021, August 26). Retrieved November 5, 2021, from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

[42]K-Nearest Neighbor(KNN) algorithm for Machine Learning - Javatpoint. www.javatpoint.com. (n.d.). Retrieved November 5, 2021, from https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

[43]K nearest neighbor: Knn algorithm: KNN in Python & R. Analytics Vidhya. (2020, October 18). Retrieved November 5, 2021, from https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.

[44]evaluate-model-on-test-data. Loading... (n.d.). Retrieved November 5, 2021, from http://americanas.link/evaluate-model-on-test-data.

[45]Mishra, A. (2020, May 28). Metrics to evaluate your machine learning algorithm. Medium. Retrieved November 5, 2021, from https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

[46] Japkowicz, N., & Shah, M. (n.d.). Machine Learning and Statistics Overview. *Evaluating Learning Algorithms*, 23–73. https://doi.org/10.1017/cbo9780511921803.003