

**Common Human Disease Prediction Using Machine Learning Based on Survey
Data**

BY

**Jabir Al Nahian
ID: 173-15-10414**

**Mehedi Mir Srabon
ID: 173-15-10413
AND**

**Mahedi Hasan
ID: 173-15-10391**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MD. Jueal Mia
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By
Mr. Aniruddha Rakshit
Senior Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

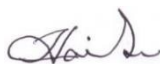
DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project titled “**Common Human Disease Prediction Using Machine Learning Based on Survey Data**”, submitted by **Jabir Al Nahian, ID NO: 173-15-10414**, **Mehedi Mir Srabon, ID NO: 173-15-10413** and **Mahedi Hasan, ID No: 173-15-10391** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 06 January, 2021.

BOARD OF EXAMINERS



Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



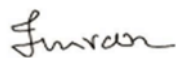
Dr. Sumit Kumar Banshal (SKBL)
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Abbas Ali Khan (AAK)
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Shah Md. Imran
Industry Promotion Expert
LICT Project, ICT Division, Bangladesh

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md.Jueal Mia, Senior Lecturer, Department of CSE and** co-supervision of **Mr. Aniruddha Rakshit, Senior Lecturer, Department of CSE, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

Supervised by:



Md.Jueal Mia

Senior Lecturer
Department of CSE
Daffodil International University

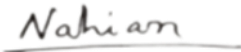
Co-Supervised by:



Mr.Aniruddha Rakshit

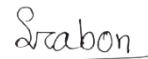
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Jabir Al Nahian

ID:173-15-10414
Department of CSE
Daffodil International University



Mehedi Mir Srabon

ID: 173-15-10413
Department of CSE
Daffodil International University



Mahedi Hasan

ID: 173-15-10391
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound to our supervisors **Md.Jueal Mia, Senior Lecturer and Mr. Aniruddha Rakshit, Senior Lecturer, Department of CSE, Daffodil International University**, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Machine Learning based research to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Professor, and Head, Department of CSE, Daffodil International University, Dhaka**. For his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

One of the main drivers of today's AI revolution is machine learning and data mining. The development of machine learning with huge data and the improvement of data mining sorting type results have already changed the Artificial Intelligence industry. So, as time is going on this field of machine learning and data mining is getting improved in the medical treatment sector. We are finding various problems and it will be solved to improve it later. In this era, the moment has arrived to move away from disease as the primary emphasis of medical treatment. Although impressive, multiple techniques have been developed to overcome the constraints of the disease approach. In the world, there are eleven types of diseases - Covid 19, Normal Flue, Migraine, Heart Disease, Lung Disease, Kidney Disease, Stomach Disease, Gastric, Diabetics, Bone Disease, Autism are the very Common diseases at this time. In this analysis, we looked at disease predictions and the factors that influence them. We studied a range of symptoms and took a survey from people in order to complete the task. This proposed work predicts the individual's symptoms and recognizes the disease. Several classification algorithms have been employed to train the model. This paper also presents a comparable examination by analyzing the enforcement of different types of machine learning algorithms. Furthermore, the performance of the model is measured with the help of performance evaluation matrices. So, from all the outputs of proposed classifier implementations, it shows that Random Forest algorithms have achieved the maximum precision of 88.2% compared to the other classifications. Finally, we discovered that the Part classifier surpasses the others.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v

CHAPTER

CHAPTER 1: INTRODUCTION 1-5

1.1 Introduction	1-2
1.2 Objectives	3
1.3 Motivation	3-4
1.4 Rational of the Study	4
1.5 Research Questions	4
1.6 Expected Outcome	4-5
1.7 Layout of the Report	5

CHAPTER 2: BACKGROUND 6-12

2.1 Introduction	6
2.2 Related Works	6-8
2.3 Background Information	9-12
2.4 Challenges	12

CHAPTER 3: RESEARCH METHODOLOGY	13-23
3.1 Introduction	14
3.2 Research Subject and Instruments	14
3.3 Workflow	14
3.4 Implementation Procedure	15-16
3.5 Training the Model	16
3.6 Data Description and Analysis	17-21
3.7 Classifier Description	20-23
CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION	24-27
4.1 Introduction	24
4.2 Performance Evaluation	24-27
4.3 Result Discussion	27
CHAPTER 5: SUMMARY AND CONCLUSION	28
5.1 Conclusion	28
5.2 Future Work	28
REFERENCES	29-30
APPENDIX	31
PLAGIARISM REPORT	32

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.4: Working procedure of disease prediction using data mining technique based on symptoms	14
Figure 3.7.1: The Random Forest method is depicted in the diagram	21
Figure 3.7.2: Support Vector Machine Diagram	22
Figure 3.7.3: Logistic Regression Transformation	22
Figure 3.7.4: K-Nearest Neighbor Algorithm Classification	23

LIST OF TABLES

TABLES	PAGE NO
Table 1: Disease prediction related works	8
Table 2: Attribute with their possible values	16-19
Table 3: Disease name and their possible values	21
Table 4: Comparison of four classifier performance	24-26
Table 5: Overall comparison of four classifiers Performance	26

CHAPTER 1

INTRODUCTION

1.1 Introduction

In today's world, we cannot think of our regular life without technology. Artificial Intelligence is one of the core parts of computer science and technology. Artificial intelligence is the process of machines, particularly computer systems, recreating human understanding. Master frameworks, native dialect handling, discourse recognition, and machine vision are examples of AI applications. Normally Artificial Intelligence is a proposal to form up a computer, a robot, or an item to know how smartly humans can think. AI normally a knowledge of how the human brain can smartly think, how humans can learn, and lastly how human select work. Artificial Intelligence mainly shows their work with the actual thinking of human thinking. Artificial Intelligence has been around since 1970. The first stage of AI is not famous at all. There have been many ups and downs in the field during this period. Artificial Intelligence is now a trending topic in modern science. Not only is the trend upward, but the rate of improvement appears to be exponential. Artificial intelligence will end up ever more imperative in arrange to help us with organizing the data.

Approximately 37% of organizations have actualized AI in their day-to-day operations and this number is set to extend as the amount of data, which has to be categorized, increments as well. AI needs a base of trained hardware and software for writing and conditioning machine learning algorithms. No other programming language is similar to AI, it has its own different techniques and programming knowledge. But some languages like including python, R, and Java, are famous. In general, AI frameworks work by absorbing large amounts of labeled preparation data, evaluating the data for relationships and designs, and then using these designs to make predictions about future states. In this approach, a chatbot that has been fed examples of content discussions can learn to deliver comparable transactions with people, or a picture recognition system can learn to differentiate and depict items in pictures by examining millions of images. Artificial intelligence can be humanity's last development. Creating a sort of AI that's so modern it can itself make AI substances with indeed more prominent intelligence might change man-made innovation until the end of time. Such substances would outperform human insights and reach superhuman accomplishments.

Another algorithm within the field of Machine Learning is Data Mining Which is a fast-growing area. Data mining is a technique for extracting and locating patterns in large data sets using methodologies that combine machine learning, measurements, and database frameworks. Data mining may be a sub-field of computer science and its goal is to make a statistic to secure the information with methods. After that using the information to make an understandable structure for further use. Data mining is considered to be one of the well-known terms of machine learning because it extricates significant data from the huge heap of datasets and is utilized for decision-making tasks. Data mining is demonstrated to be an awesome apparatus for investigating modern roads to consequently examine, visualize, and reveal designs in the information that encourage the decision-making process. In August 1989, the first Information Disclosure in Databases Workshop was held in connection with the 1989 Worldwide Joint Conference on Artificial Intelligence, and this workshop arrangement evolved into the 1995 Worldwide Conference on Information Revelation and Data Mining (KDD). Artificial intelligence and Data mining methods have been utilized in numerous spaces to illuminate classification, segmentation, affiliation, determination, and forecast issues. Data mining and Machine Learning, an area of Artificial Intelligence, deserve credit.

Nowadays, many people are not mindful of their well-being. Peoples are not concerned about their health issues and they don't know about the exact disease reaction. There are some people who are not curious about getting to the doctor and that's why people of younger ages are also getting into serious diseases at that time. There are some people who check up on their first stage of health issues but after that, they are not serious about their further treatment. They just take the old medicines and on the deeper side they fall into a big disease problem and they have no idea of it. Old people cannot go to the hospital due to their bodily weakness and that's why they always skip the routine checkup. Another side there are some people who are poor and can't manage to go to the doctor for financial issues.

After seeing such kind of problem, we got a thought to make a machine to predict human diseases by their symptoms. We have done with nine diseases and dealt with 53 symptoms under some queries. By doing this we have collected our data and also work with different types of algorithms like Random Forest, SVM, Logistic Regression KNN, etc. in our research. We calculated many performance evaluation criteria and compared the results to select the best classifier in the working situation.

1.2 Objectives

The main objective of this research is to create a system that can predict disease at an early time by using data mining and machine learning techniques. In today's healthcare, our main challenge is to get the best quality services and predict the exact diagnosis. Nowadays people are unknowing of their health issues. They cannot specifically know about which disease they are suffering. The doctor gives people so many checkups to find out their actual disease. People get messed up with so much overthinking and its impact on people's health issues. For this reason, we try to figure out a system that can easily diagnosis the exact disease of people's health. So, at first for the illness people, the system will see people's health condition, and then it will predict a disease by their symptoms. Another side Normal people also can check up their regular routine by this system. Sometimes people get worried about their sudden health issues. In that case, people for their mental satisfaction can easily check up on their health condition. This system can reply to complex questions for diagnosing disease and hence help healthcare professionals to create clever clinical choices. Then the doctor can easily identify the disease and they can make the treatment at early stages. By giving successful medications, moreover makes a difference to decrease treatment costs. So, this system disease prediction can help both sides to doctors and normal people to know health conditions.

1.3 Motivation

A major challenge confronting healthcare organization is the arrangement of quality services at reasonable costs. Nowadays, the Maximum number of people of Bangladesh are not concerned about their health condition for the high cost of treatment. Sometimes they just visit a doctor but they don't take the checkup seriously for the high cost. So, in that case, they don't know about their actual health disease. In Bangladesh, we can also find some people who are really lazy about their daily routine checkups and for that reason, they attack with dangerous diseases without knowing it. Peoples neglect their health conditions for laziness and the high cost of treatment. After seeing such kind of condition, we got a motive to make a system which can give an easier way to find out people's diagnosis. An important additional part we found is that there are only a few works done in this research field for multiple disease prediction. And We have done our work with nine diseases and 53 symptoms below several examine. So, our work is fully new approaches to this research field and that's why we thoughts machine learning merged with data mining it's a very good topic to contribute to.

Another side Food adulteration is dangerous since it may be harmful and can influence well-being and it might deny supplements fundamental for appropriate development and improvement of a human being. The most exceedingly bad portion is a few adulterated foods indeed causes cancer, the most life undermining illness. After seeing such incidents, we raise an important question: "How can we turn data mining and machine techniques that can make a system to predict disease by people's symptoms and Can you help healthcare providers make better clinical decisions?" This is the primary motivation for this study.

1.4 Rational of the Study

There is a lot of work with data mining and machine learning going on in the medical field. Everyone is working with a single disease in their work. But we have done our work with nine diseases and given a suspect disease result by their symptoms. As a result, our work will be a fresh take on the medical profession. Using our best efforts, we have improved the existing data mining and machine learning development process. Machine learning might be a fascinating area in and of itself. Combining it with another fascinating discipline of data mining could result in a truly amazing work of energy.

1.5 Research Questions

Our work will assist data mining approaches to solve their surrounding difficult problems. A professional and useful classifier establishing is the main goal to get the higher result. To guide a research project, we must first develop a collection of expert questions, which will all be answered at the conclusion of our job. We've compiled a list of questions that we'll address through our work. They are,

1. Is it possible to use different data mining approaches for testing and several performances to measures?
2. Can data mining algorithms be combined with machine learning?
3. How can we choose the best classifier to get the best result?
4. Will the evolutionary methods produce better results than the traditional methods?

1.6 Expected Outcome

AI is present in every aspect of our digital lives. In this digital world, there is seeing extensive elevations of machine learning and data mining which is another range of AI. Nowadays, in our medical field, so many things are surrounded by machine learning and

data mining. So, after our work, we will be able to make a comfort zone for all ages of people in their medical services. Now, we can describe our outcomes in the given below:

- Different data mining approaches are used for improving speed and better accuracy.
- Optimized data mining results will drastically upgrade execution.
- Testing data mining approaches several times for choosing the best classifier output.
- The accuracy of the classifier will improve.
- The human can know their body condition.
- Get proper advice for good health.
- They can also be conscious about their body.
- People of any age can do their body check-ups.

1.5 Layout of the Report

With the use of data mining, we described how Machine Learning is becoming increasingly important. Here, it is also explained. how we have collected our database with the use of data mining and the work of Machine Learning. How we defined our inquire-about question. And how we get the arrangement within the, to begin with the chapter.

We covered the background of this platform in the second chapter. Here are some examples of earlier work in this discipline that deal with a single condition in the medical sector. And some past works on this subject are likewise ancient, but they were extremely useful in instilling knowledge into the work we are performing. We'll provide some background information to help you understand what we're doing.

The entire workflow is detailed in the research methods section of this paper. From data collecting to planning algorithms, we've got you covered. In this section, everything is explained in great depth.

In the fourth, we described the result part of our test. We have also talked approximately how we ran the test and what are the come about. Moreover, clarified results in detail.

Finally, we attach our report, which includes chapter five, which details what we did, what we watched, and what our future work entails.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

In the early days of AI, there were a lot of studies done on expanding strings. This field has recently experienced a resurgence of friction. Machine learning and data mining are two AI work fields that are gaining traction. Now, in this background portion, we'll look at past work done in this field within the related work section. In the following part, a few background information details are offered to aid in the understanding of the research effort. Finally, a few challenges were discussed.

2.2 Related Works

This type of problem has a lot of complications that change over time. Therefore, in this section, we looked at various publications to see whether there is a research gap in previous work.

Cheng-Ding Chang et al. [1] made a two-phase research approach for predicting hyperlipidemia and hypertension at the same time. They began by selecting specific risk variables for both these two diseases using six data mining methodologies and then utilized the voting principle to discover the shared risk factors. After that, they built multiple predictive models for hyperlipidemia and hypertension using the Multivariate Adaptive Regression Splines (MARS) approach. Sohyun Bang et al. [2] developed a multi-classification method based on ML to make a distinction between the gut microbiome and the six diseases listed below: juvenile idiopathic arthritis, multiple sclerosis, chronic fatigue syndrome/myalgic encephalomyelitis, acquired immune deficiency syndrome, colorectal cancer, and stroke. To create the prediction model, they used the abundance of microorganisms at five taxonomic levels as characteristics in only 696 samples obtained from various research. Four multi-class classifiers and two feature selecting approaches, including forwarding selection and backward removal, were used to create classification models. Ajinkya Kunjir et al. [3] proposed a method for efficient and advanced disease prediction based on historical training data. Analyzing and evaluating different data methods is the best strategy. For each disease algorithm training data example, the datasets chosen for implementation purposes comprise more than 20 medical relevant attributes. Heart disease, breast cancer, arthritis, and diabetes are among the medical datasets chosen for the research.

The Naive Bayes method was chosen to implement in this project after assessing the prediction accuracy and latency test results. Julian Besag et al. Using an epidemiological

dataset of COVID-19 patients from South Korea, L. J. Muhammad et al. [4] established a model for predicting COVID-19 affected patients' recoveries. To create the models, the decision tree, support vector machine, naive Bayes, logistic regression, random forest, and K-nearest neighbor algorithms were directly implemented on the dataset using the Python programming language. [6] presented and demonstrated a Geographical Analysis Machine Learning method for detecting tiny illness clusters. A secondary goal is to go over some frequent difficulties in applying clustering tests to epidemiology data. For the classification of breast cancer disease, F. M. Javed Mehedi Shamrat et al. [7] employed six supervised classification approaches. SVM, NB, KNN, RF, DT, and LR are examples of early breast cancer prediction algorithms. As a result, we used specificity, sensitivity the f1 score, and total accuracy to assess the breast cancer dataset.

The results of the breast cancer prediction performance analysis show that SVM had the best results, with a classification accuracy of 97.07 percent. NB and RF, on the other hand, have the second-highest forecast accuracy. Apurb Rajdhan et al. [8] Used data mining techniques such as Random Forest, Naive Bayes, Logistic Regression, Decision Tree the suggested work predicts the likelihood of heart disease and classifies the patient's risk level. As a result, this research conducts a comparative analysis of the performance of several machine learning methods. M. Banu Priya et al. [9] analyzed liver patient datasets in order to develop classification algorithms for predicting liver disease. The min-max normalization technique is applied to the original liver patient datasets obtained from the UCI repository in the first phase. In the second step of liver dataset prediction, a subset (data) of the liver patient dataset is obtained from the complete normalized liver patient datasets, containing only significant attributes, using PSO feature selection.

The data set is then subjected to categorization algorithms in the third phase. The accuracy will be calculated in the fourth phase using the root mean error value and root mean square value. A.K.M Sazzadur Rahman et al. [11] Used Six types of supervised classification algorithms are used in this research, there are Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Naive Bayes, and Random Forest. The execution of different classification methods was assessed on distinctive estimation procedures such as accuracy, precision, recall, f-1 score, and specificity. In a large community pediatric clinic, Terisa P. Gabrielsen et al. [14] developed controls and children who screened positive during universal autism screening. Following the screening, medical evaluations were conducted to ascertain the pattern models (autism, language delay, or typical).

Unaware of participants' diagnosis status, licensed psychologists with toddler and autism expertise assessed two 10-minute video samples of participants' autistic evaluations,

evaluating five behavioral patterns: Responsive, Conducting, Verbalizing, Play, and Responding to Name. Reviewers were asked to give their opinions on autism referrals based purely on 10- minute assessments.

TABLE 1: RELATED WORK OF DISEASE PREDICTIOND

Article	Disease Name	Functionality	Observed Features	Models / Algorithms	Results
[1]	Hypertension and Hyperlipidemia	Using a two-phase analysis approach, they forecast hyperlipidemia and hypertension at the same time	To create a multivariate predictive model for hypertension and hyperlipidemia, they employed the Multivariate Adaptive Regression Splines (MARS) approach	Logistic regression CART CHAID MARS	Accuracy rate: 93.07%
[2]	multiple sclerosis, juvenile idiopathic arthritis, myalgic encephalomyelitis/chronic fatigue syndrome, acquired immune deficiency syndrome, stroke and colorectal cancer	A few microbes' abundance has been used as a flag to forecast a variety of diseases. They anticipated in this work that utilizing a multi-classification ML technique, they could discriminate the gut microbiota from six illnesses	They used four multi-class classifiers and different feature selection strategies, using forward choice and backward removal, to create classification techniques	KNN LMT SVM Logit Boost	Accuracy rate: 83.1%
[3]	Diabetes, Breast cancer, Heart Dataset	Create a simple decision-making system that can identify and extract Previously unknown patterns, connection, and theories linked to numerous diseases from previous database files of various illnesses.	The developed scheme can answer challenging questions for recognizing a specific condition and may also help healthcare professionals make patient care decisions that existing decision support systems couldn't	Naive Base J48	Accuracy rate: Diabetes Dataset: 76.30% Breast cancer Dataset: 71.45% Heart Dataset: 83.49%

2.3 Background Information

2.3.1 Machine Learning

Machine learning is an evolving section of modeling algorithms and designed to mimic human brilliance by getting knowledge from surroundings. It has been used famously in various fields like email filtering, computer vision, predictive analytics, medical applications, etc. Machine learning also gives good results in all these fields and that's why the uses of machine learning are increasing day by day. Nowadays, in the medical section, the cancer patient-doctor recommends radiotherapy as part of their treatment. The proficiency of machine learning algorithms in generalizing learning from the current context and in invisible tasks will allow improving both the safety and effectiveness of radiotherapy practice which leads to better results. So, machine learning plays a much bigger role as a helper in the medical field. Machine learning types are showing effective results in their performance. They play a very important role in increasing classifier accuracy in their work. Machine learning types are divided into three stages. They are-

Supervised learning is the primary model of machine learning. In supervised machine learning, machine learning algorithms have to be trained like labeled data otherwise without exact labeled data the method can't work properly. Supervised learning will act extremely strong when used in the right situations. Supervised machine learning algorithms will always try to improve, finding new shapes around and making relationship like it train oneself on new data.

Unsupervised machine learning has the unique technique of being able to deal with unlabeled data. In this type, algorithms don't need to wait for pre-assigned labels for training. This means that human work is no demand here for training the dataset to be machine-readable. The program allows large data set for work by itself. Unsupervised learning can easily adjust with the data and make it possible to change the hidden structure. This algorithm offers more deploying development than supervised learning algorithms. Reinforcement learning is a type of machine learning method that conducts making a rhythm of decisions.

Reinforcement learning gets inspired by humans that how humans take learns from data in their actual lives. So, this algorithm learns every time from interacting with surrounding environments. After that, it gives a positive or negative reward based on its action.

Machine learning is a scorching trend topic in the research area. All the new techniques of machine learning always looking for development all the time. Sometimes this field's speed and complexity connect with new techniques which makes it difficult even for

experts. So, let us know about different techniques to make machine learning demystify and to offer a learning path for those who are new to the original idea. The following techniques described provide an overview and are based on what you can build based on your machine learning knowledge and skills: Tracking Patterns, Classification, Association, Outlier Detection, Clustering, Regression, Prediction. Through knowing techniques, we will be able to learn about how data mining experts rely on strategies and technologies from the junction of database management statistics and machine learning to better understand their carriers. How does machine learning work? To answer this question first we have to know the machine learning working parts. Machine learning consists of three parts: Computational algorithm is the main basis of diagnosis, Variables and properties take action to make a decision, Best knowledge makes easy to understand the system and trains the systems to learn.

Recently, the progress in machine learning and the improvement in predictive analytics technology have made things simpler. Machine learning showed observable resolves in the making of human disease prediction. Machine learning normally adds some implementations by using the data and training their algorithms for making predictions, exposing solution vision within data mining works. In the prediction section machine learning already showed so many effective works with the data and using algorithms in the data mining field. In our work, we also used many machine learning algorithms to measure our work. After labeling the data we conduct the classifiers into the data for looking at which classifier will give the best result. By measuring, we can find which classifier worked best and give a good accuracy of the result. Machine learning can easily identify trends and patterns for the newest work in their field. In their work, there is no need for the intervention of humans. Machine learning is gaining a lot of recognition in front of the world day by day due to continuous improvement. It can handle multi-dimensional and multi-variety data. Machine learning produces a huge around of wide applications. Machine learning is becoming increasingly popular, and it is showing promise in predictive analytics. The main reason behind that it's using the data flow then training its algorithms to merge with data mining projects. As a result, it has the potential to outperform other methods.

2.3.2 Data Mining

Data mining is a set of inconsistencies in the set to predict the results and find out the correlation. Data mining could be a source of information where it's also called a knowledge discovery of databases. It's also containing the new process of discovering, useful patterns and making relationships with a large amount of data. This field getting combines apparatuses from statistics, artificial intelligence, and machine learning with the control of databases to analyze the huge number of datasets. Data mining is mostly

used in business, scientific research, and government security. Data mining techniques and tools are empowered undertakings for predicting future movement and gathering more information about the business section. Using a wide range of data mining techniques you can use data to progress profits, decrease price, decrease risks, and make better customer connections, and much more. It's a process that companies use to turn their hard data into beneficial data. The data mining process and their works are famous for having satisfying results in the field. In the next generation purpose, data mining will feel more complex leader type, in addition to other variables and their relationships designed for any model, and further refinement is possible through research that will create new methods for determining the most interesting features. Data mining is based on three interconnected scientific disciplines:

Statistics the numerical collection of data is present for purpose work. Probably working mathematically with the analysis and version of collecting numerical data using data mining accesses. Already been so many successful works done in this statistics data mining section.

Artificial Intelligence with data mining approaches there are so many effective works that already taking huge placement in our world. Mainly in artificial intelligence human-like intelligence work is mainly executed and work is presented by machine or software. So, data mining work was to prepare the actual solid data for the working process.

Machine learning with data mining processing work is booming right now. Machine learning is a process where they run their algorithms so that they can learn from data and apply it to build predictions. In the meantime, for better result machine learning depends on the data mining process where data mining labels all the data and give actual data to classifiers for testing.

Data mining explores and analyzes big data blocks to gather meaningful shapes and styles. Data mining can be used in many ways such as spam email filtering, fraud detection, disease prediction, database marketing, etc. Its works are getting viral day by day. In the AI field, it can be merged with other subfields of AI by giving good results in their prediction work. So, since we already know about their successive works, now we see how it works. The data mining process is divided into five parts and they are- At First, the system collects data and then packs it into the stores. Then managed the collected data and make it secure. After that access the data and find out how to arrange it. Then, utilization software sees the user's result and sorts the data based. Finally, the end-user gives the data in an easier share format

Combining work of data mining and machine learning is increasing rapidly. Machine learning and data mining combined works are giving a great successful reward in the

world. In this research work, we have also done our work with data mining and machine learning where data mining leads a vital role to increase our accuracy of the work. From the beginning, firstly data mining collects all the data and then stores them. After that, they labeled the data for having the actual right data to apply for testing purposes. Like in our research work we have also collected a total of 503 individual data are used here to accomplish this work. Then the data is used for the training of the classifier and the rest of the data is employed for testing purposes. So, it's given the sorted data to the classifier for having a good accuracy of work. Data mining's importance is growing in the section of market decisions, making predictions, and more. Nowadays. Nowadays Data industries are growing rapidly and demanding more data in their industry for that reason it also increases the demand for data analysts. In the business marketplace, we can analyze customers' behavior's reaction and their insights with the help of data mining. This makes a glorious success and leads the data types business one step further. It is mostly used for prediction type method where a lot of high range of work is done by the data mining process. Data mining also achieved massive success in the medical field where its use is huge. Even after learning so much about data mining, the question may come to mind that "Why is data mining important?" For clarifying this we put some point in that purpose. We can use data mining to sort through all of your information's disorganized and monotonous cacophony. Determine which is applicable, and then select the appropriate data to assess the most likely outcomes. Quicken the time of making exact decisions of informed. The impact of data mining is spreading so much day by that it will be very profitable in their future works. In that case, the results of data mining will be profitable if it will explore new niches in global business-related areas and advertisement will be target potential customers of new selections. We use these tools to discover the best airfare and phone number of a long-lost classmate or to find the best deal on lawnmowers, and we use them to find the best airfare and phone number of a long-lost classmate. Assume that intelligent agents have lost their grip on medical research data and atomic particle physics. Computers can provide fresh insights into the nature of the universe or novel remedies for diseases. It is clear from the description that data mining is gaining more and more successful in its new work through the use of new techniques.

2.4 Challenges

The primary challenge of this work was to create questions from symptoms that were verified by the doctor. We created data set by collecting data from 500 people through questions. We have faced so many challenges in this work. Collected data from everyone and then edit the data set. Our experiment has to be made so much carefully, otherwise; the result will become wrong.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Multiplied well-known investigate groups have begun creating optimized machine learning and data mining. Then they developmental consolidating them in their investigate works. It recently become very popular for combining machine learning and data mining for classification. Among typical Evolutionary Algorithms, Data mining is the foremost prevalent one to utilize with machine learning.

3.2 Research Subject and Instrumentation

The research subject is a platform that was inspected and considered for free ideas. So, in our research work, it was machine learning and data mining. We learned about our work not only for operation but also for designing models, collecting data, and training the model. The other element is high-tech instrumentation, as well as the approaches we used. We used the Windows platform, python languages with many libraries like Pandas, NumPy, seaborn, etc. In the operating system where we used Anaconda, Google Collab, and extra tools like weka, orange application for all the training and testing processes. These are the open-source sequence of python for machine learning applications.

3.3 Workflow

Primary 4 Steps:

- Data Collection
- Data Preprocessing
- Implement algorithm
- Performance Evaluation

3.4 Implementation Procedure

The purpose of this work is to achieve disease prediction. Many important characteristics, especially disease symptoms, are considered to ensure an accurate prediction. Figure 3.4 depicts the many stages we followed to execute this project. First and foremost, a 53-question survey based on disease symptoms has been produced. Then, with the help of this survey, we obtained data from a huge number of people. We next employed several preprocessing procedures to feed this data into the classifier. To label a specific question, only one variable is used. To identify all of the questions, a total of 53 variables are employed. After preprocessing, our data is separated into training and testing sets. 70 percent of the total data set was utilized for training purposes in this example. The remaining 30% of the data set was used for testing purposes only. This is a completely random division. The classifiers were then trained using the training data. We employed testing data to predict the present disease status after training the classifiers. Some of the performance evaluation measures have been computed here.

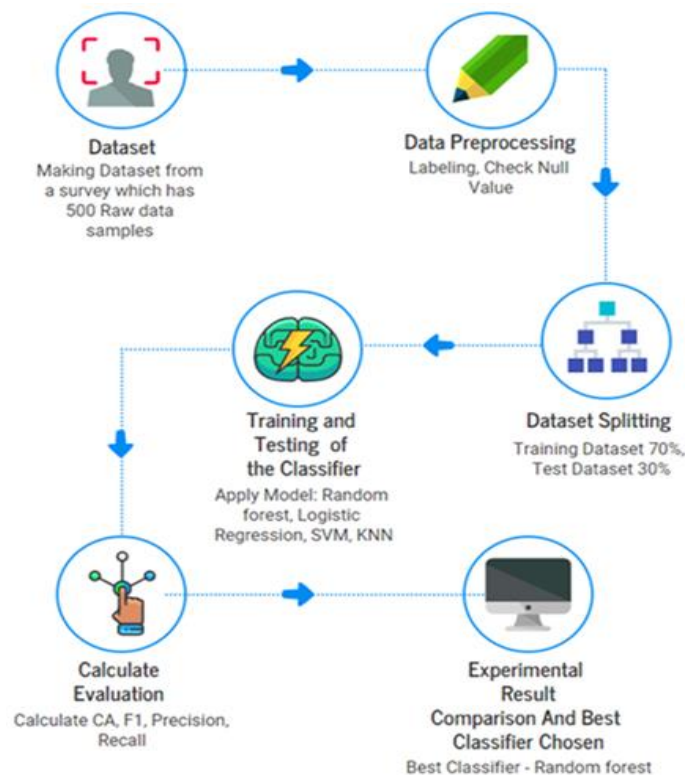


Figure 3.4: Working procedure of disease prediction using data mining technique based on symptoms

We calculated the best classifier to predict in this environment using these metrics. Several performance measures in percentage have been calculated based on the confusion matrix created by the classifier using below equations.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%$$

$$\text{Sensitivity or Recall or True Positive Rate (TPR)} = \frac{TP}{TP + FN} \times 100\%$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\%$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\%$$

3.4.1 Implementation Requirements

- Hardware:
 - Processor: Core i5
 - Ram: 16GB
 - GPU: GTX 1660
- Software:
 - Google colab, Anaconda, Weka, Orange
 - Language: Python 3.8
 - Libraries: Numpy, pandas, sklearn, label encoder.

3.5 Training the Model

Overall, for this work we divided our collected data into the training and testing sets after preprocessing. In this case, 70 percent of the complete data set was used for training purposes. The remaining 30% of the entire data set was used for testing reasons. We took those data which are appropriate for our work and can't take any garbage data like as unfinished data. And we also apply many different data mining techniques on this data set. Then we trained the collected data set and find out the best classifier.

3.6 Data Description and Analysis

Dataset used in this research is collected by a survey. The survey mainly consists of 52 questions. We have created these questions from the symptoms of the 11 diseases that we are working with. And each question taken by analyzed and made from the main

symptoms of diseases. Both personal and disease symptom factors are considered in these 52 questions. A total of 52 variables are utilized to label all of the questions. The number of independent variables in the dataset is 51, whereas the number of dependent variables is one. Table 2 lists all of the variables and their possible values.

In order to complete this task, 503 individual records were used. 70 percent of the data is utilized to train the classifier, while 30 percent is used for testing.

TABLE 2: ATTRIBUTE WITH THEIR POSSIBLE VALUES

Variable	Full-form	Variable type	Possible value
Dyhof	Occasional fever	Independent	Yes (1), No (0)
Mof	Measure of fever	Independent	98-102(1), 102-105(2), Normal (3)
Amc	Almost cough Yes (1), No (0)	Independent	Yes (1), No (0)
Toc	Types of cough	Independent	Dry (1), Blood (2), Mucus (3), Normal (4)
Ft	Feel tired	Independent	Yes (1), No (0)
Vm	Vomiting	Independent	Yes (1), No (0)
Tov	Types of vomiting	Independent	Stomach bloated and vomiting (1), Vomiting (2), Nausea (3)
Ftotp	Following types of throat problems	Independent	Throat pain (1), Voice change (2), Sore throat (3), None (4)
Rsptp	Respiratory problems	Independent	Breath weakness (1), Shortness of breath (2), None (3)
Sos	Sense of smell	Independent	No olfactory power (1), Not get smell properly (2), Olfactory power is ok (3)
Dodi	Occasional Diarrhea	Independent	Yes (1), No (0)

Hpyr	Head problems	Independent	Dizziness (1), Severe Headache (2), Normal headache (3), None (4)
Oryb	Occasional rashes	Independent	Yes (1), No (0)
Hrn	Runny Nose	Independent	Yes (1), No (0)
Flitb	Feel less in the body	Independent	Yes (1), No (0)
Sfkd	Suffer in Depression	Independent	Yes (1), No (0)
Snha	Stiff neck	Independent	Yes (1), No (0)
Dyep	Eye's pain	Independent	Yes (1), No (0)
Pobsf	Pain on both sides of forehead	Independent	Yes (1), No (0)
Ybin	Body numbs	Independent	Yes (1), No (0)
Tslt	Tolerate sound, light, touch	Independent	Yes (1), No (0)
Ftope	Following types of problems experience	Independent	Chest pain (1), Chest pressure (2), Chest discomfort (3), Chest throbbing (4), Sweating (5), Chest tightening (6), None (7)
Thbt	Type heartbeat	Independent	Decreases Rapidly (1), Increase Rapidly (2), Normal (3)
Slc	Lose consciousness	Independent	Yes (1), No (0)
Npc	Pain problems experience	Independent	Neck pain (1), Jaw pain (2), Spinal pain (3), None (4)
Atw	Ability to work	Independent	Yes (1), No (0)
Lwvr	Lose weight without reason	Independent	Yes (1), No (0)

Tpof	Types of appetite	Independent	Excessive appetite (1), Appetite Depression (2), Normal (3)
Urpm	Urination Problems	Independent	More (1), Less (2), Not at all (3), normal (4)
Obib	Observation in body	Independent	Weight gain (1), Swelling in the body (2), Swelling of eyes and face (3), None (4)
Scyh	Sleep Condition	Independent	More (1), Less (2), Normal (3), Not at all (4)
Abdp	Abdominal Pain	Independent	Complete abdominal pain (1), Lower Abdominal pain (2), Flatulence (3), None (4)
Diyh	Indigestion	Independent	Yes (1), No (0)
Bsdf	Black stools during defecation	Independent	Yes (1), No (0)
Asys	Airway sore	Independent	Yes (1), No (0)
Dwpd	Drink Water per day	Independent	1-2L (1), 3-5 (2), 5-8(3)
Dysbv	Blurred vision	Independent	Yes (1), No (0)
Wbdul	Wound body dry up late	Independent	Yes (1), No (0)
Ioynf	Infections on your nails or fingers	Independent	Yes (1), No (0)
Imhad	Irritable mood	Independent	Yes (1), No (0)
libd	Itching in body	Independent	Yes (1), No (0)
Nacht	Noticed any changes height	Independent	Increase rapidly (1), Decrease (2), Fixed (3)

Bppf	Body Pain Problem	Independent	Pain to sit up (1), Muscle pain (2), pain to bend down (3), Pain in the joints (4), Tingling in the hands and feet (5), None (6)
SpCY	Speak clearly	Independent	Yes (1), No (0)
RtCOO	Respond to the call of others	Independent	Yes (1), No (0)
KaeOO	Keep an eye on others	Independent	Yes (1), No (0)
Dyew	Empathy	Independent	Yes (1), No (0)
LadyO	like aloneness	Independent	Yes (1), No (0)
GuiY	Gesture unusual	Independent	Yes (1), No (0)
Syaay	scared	Independent	Yes (1), No (0)
Oaay	Overly abusive	Independent	Yes (1), No (0)
Rtswaa	Repeat the same word again and again	Dependent	Yes (1), No (0)
Disease	All disease name		Diseases name labelling (a, b, c, d, e, f, g, h, I, j, k, x)

We have worked on 11 diseases where we have taken each disease in a variable and also using encoder label for numbering these diseases. Label encoding prescribe to the transformation of dataset labels into numbers. So that it is readable for the machine. And Machine learning algorithms can make better decisions about how to handle those labels. This is an important pre-processing step for structured datasets in supervised learning.

And we have taken the X variable for no problem. Here no problem means that diseases free body. In our research we work for detect multiple disease in human body but if a human body can't contain any symptoms of disease that means disease free body. So that we put no problem option. Table 3 shows diseases name their possible values and encoder label.

TABLE 3: DISEASES NAME & THEIR POSSIBLE VALUES

Disease's Name	Possible Values	Label Encoder
Covid-19	a	0
Normal Flue	b	1
Migraine	c	2
Heart Disease	d	3
Lung Disease	e	4
Kidney Disease	f	5
Stomach Disease	g	6
Gastric	h	7
Diabetes	i	8
Bone disease	j	9
Autism	k	10
No Problem	x	11

3.7 Classifier Description

The Random Forest classifier is a supervised learning algorithm that creates forests by categorizing groups of random trees. It can be used for classification and regression problems. It's a popular method for solving classification difficulties. It picks samples at randomly from a particular dataset. It uses data samples to generate decision trees, which are subsequently used to make predictions [18]. Then, using the voting method, chooses the appropriate solution. While developing the trees, the random forest contributes more randomness to the pattern. When dividing a node, it searches for the best trait from a specified distribution rather than the most significant characteristic. As a result, there is a great deal of variety, resulting in a better model. Because it is an ensemble learning technique, Random Forest outperforms a single decision tree. The overfitting problem is reduced by averaging the results.

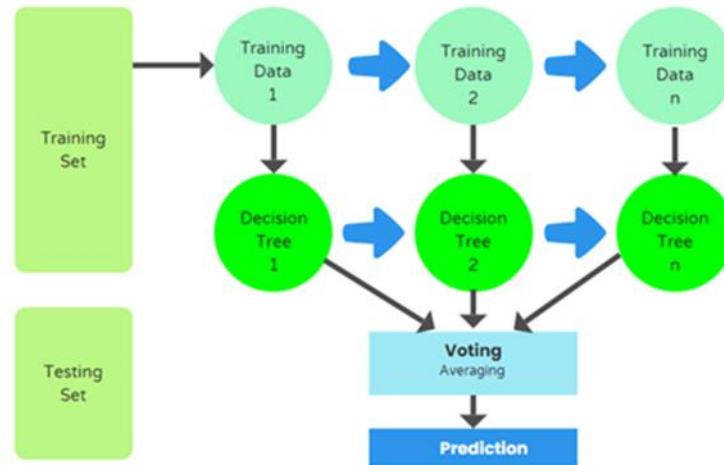


Figure 3.7.1 The Random Forest method is depicted in the diagram

The "Support Vector Machine" (SVM) is a supervised machine learning technique for classification and regression. However, it is usually used to address categorization problems [19]. Every data point is represented as a point in an n-dimensional area (where n is the number of characteristics we have), with the value of each feature being the SVM classifier's score at a specific place. Support Vector Machine is used to select the maximum number of nodes that will help form the hyperplane. The algorithm is known as a Support Vector Machine, and support vectors are also the maximum instances. Observe the picture below, that shows how a decision hyperplane is used to classify two separate groups. Next, we identify the hyper-plane which clearly distinguishes the class labels to complete identification (look at the below Fig. 3.7). Simply put, support vectors are also the positions of each accuracy assessment. The Classification algorithm is a frontier that separates the two categories (hyper-plane/line) the most effectively.

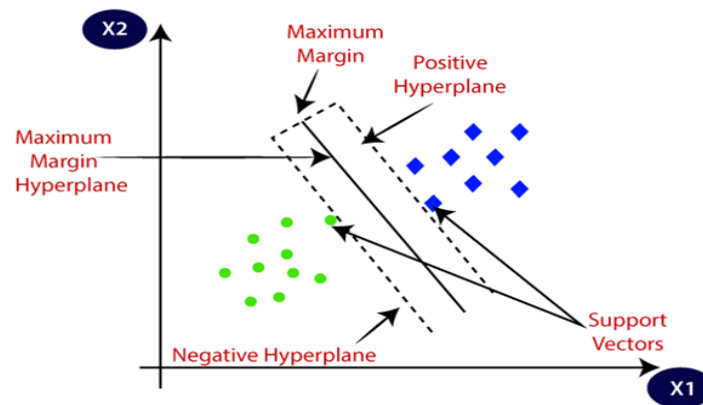


Figure 3.7.2: Support Vector Machine diagram

The supervised learning approach of logistic regression has been used to predict the categorical outcome variable using only a collection of individual variables [17]. This is an important and strong method since it can provide possibilities and identify updated information using those discrete and continuous data. The logistic (Sigmoid) model is a mathematical equation that maps anticipated outcomes to probabilities. It can convert any actual value between 0 and 1 into another. This method's major implication is that while the predicted output must be classified and that the input parameter must not be multi-collinear.

- Logit: $\ln[p/(1-p)] = a + BX$
- Logistic: $p = \frac{e^{a+BX}}{1+e^{a+BX}}$

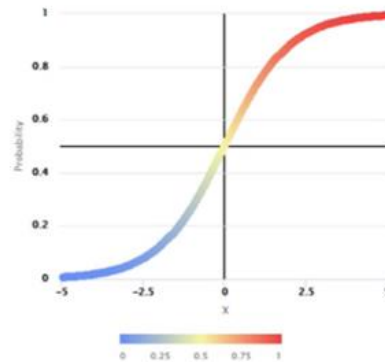


Figure 3.7.3: Logistic Regression Transformation

The K-Nearest Neighbour method is one of the most fundamental Machine Learning techniques. It is predicated on the Supervised Learning approach. The KNN approach is used to address both classification and regression problems. Feature matching is the foundation of the KNN method. K Nearest Neighbor (KNN) is a simple, easy-to-understand, and adaptable machine learning approach.

KNN has applications in handwriting recognition, finance, political science, healthcare, image recognition, and video recognition, to name a few. Credit ratings are used by financial companies to anticipate a customer's credit rating [20].

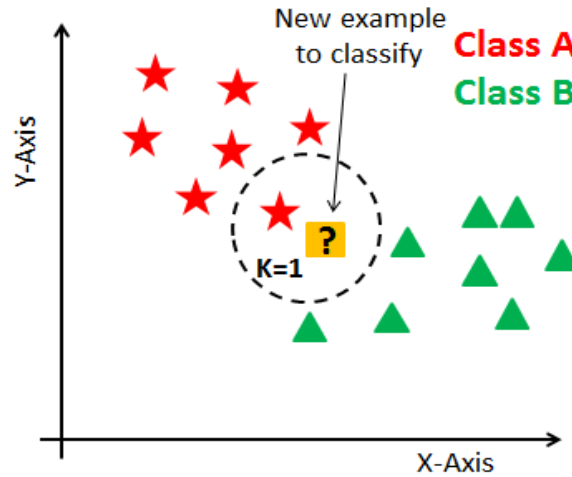


Figure 3.7.4: K-Nearest Neighbor Algorithm Classification

The KNN method implies that the current specific instance and old instances are comparable, and it assigns the new case to the category that is closest to the classifications. KNN represents the number of closest neighbors. The number of neighbors is the most crucial factor to consider. K is usually an odd number when there are two classes. When $K=1$, the process is referred to as the nearest neighbor algorithm.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

We intend to discuss the findings of the general test in this section. We'll look at how to measure execution and see what results from it at the end, using various types of estimations.

4.2 Performance Evaluation

The classifier generated a 12*12 confusion matrix because this is primarily a multiclass problem. Table 3 shows the resulting matrix for each of the classifiers. Accuracy, F1, Precision, and Recall Score are calculated from the above confusion matrix to evaluate this work. Table 3 shows the results of numerous performance evaluation metrics. In the overall examination of the results, Table 4 shows that the Part classifier outperforms the other four classifiers. The Part classifier has the highest accuracy of all the classifiers for all classes 88.2, 88.1, 88.5, and 88.2. Other Table 3 and Table 4 results also support the Part classifier. After apply SVM, Random Forest, and Logistic Regression algorithm we found best classifier algorithm as Random Forest. The Result shown below:

TABLE 4: COMPARISON OF FOUR CLASSIFIER'S PERFORMANCE

Classifier	Class Name	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Random Forest	Covid 19	99.3	93.6	96.4	88.0
	Normal Flue	97.6	92.1	90.6	69.4
	Migraine	96.4	81.4	94.6	81.6
	Heart Disease	99.8	97.1	98.4	93.7
	Lung Disease	99.5	87.5	100	81.3
	Kidney Disease	99.5	91.7	100	94.6
	Stomach Disease	99.5	93.3	99.8	91.4
	Gastric	100	100	100	100

	Diabetics	99.3	94.1	96.0	92.3
	Bone Disease	95.5	96.7	91.9	92.7
	Autism	93.5	88.7	84.1	83.8
	No Problem	92.6	88.9	81.7	79.3
SVM	Covid 19	98.8	88.9	93.4	80.0
	Normal Flue	93.8	80.3	75.7	85.5
	Migraine	96.2	81.4	94.6	71.4
	Heart Disease	100	99.0	83.3	81.4
	Lung Disease	99.5	87.5	100	81.3
	Kidney Disease	98.3	93.2	95.4	91.6
	Stomach Disease	100	99.5	100	100
	Gastric	99.0	91.7	88.0	86.7
	Diabetics	98.6	87.5	84.7	80.7
	Bone Disease	89.5	83.1	81.9	79.5
	Autism	82.7	74.8	81.8	74.7
	No Problem	89.0	79.6	84.1	84.7
	Logistic Regression	Covid 19	98.1	83.3	87.0
Normal Flue		91.1	69.9	70.5	69.4
Migraine		91.6	69.8	78.4	77.3
Heart Disease		89.8	86.3	81.5	79.6
Lung Disease		100	96.7	98.3	94.7
Kidney Disease		99.0	83.3	90.9	86.9
Stomach Disease		100	100	100	100
Gastric		100	100	100	100

	Diabetics	97.1	87.5	81.8	79.2
	Bone Disease	91.6	84.4	78.6	75.2
	Autism	81.5	70.3	72.3	80.5
	No Problem	86.3	80.1	84.7	86.6
KNN	Covid 19	82.1	77.9	76.7	73.5
	Normal Flue	83.9	73.8	75.6	72.9
	Migraine	89.4	72.9	84.4	73.8
	Heart Disease	89.4	83.5	76.1	73.4
	Lung Disease	87.6	82.3	79.4	77.3
	Kidney Disease	97.1	80.0	77.1	73.8
	Stomach Disease	93.1	83.3	91.4	85.9
	Gastric	96.4	88.2	86.7	85.3
	Diabetics	93.3	89.8	83.5	77.6
	Bone Disease	81.5	79.7	74.3	71.1
	Autism	73.5	72.1	69.9	80.5
	No Problem	80.8	79.9	76.4	73.1

We have applied four classifiers to our dataset, where we have found Accuracy, F1, Precision, and Recall scores for each classifier, which we have shown in Table 5. After that we averaged the Accuracy, F1, Precision, and Recall scores of the four classifiers based on Table 3 and compared their performance. Average comparison shown in bellow table:

TABLE 5: AVARAGE PERFORMANCE OF FOUR CLASSIFIER'S

Model	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
Random Forest	88.2	88.1	88.5	88.2
SVM	77.0	76.2	80.6	78.7

Logistic Regression	71.9	71.7	73.0	71.9
KNN	68.7	66.1	66.4	67.9

4.3 Result Discussion

The classifier generated a 12*12 confusion matrix because this is primarily a multiclass problem. Table 4 shows the resulting matrix for each of the classifiers. Accuracy, F1, Precision, and Recall Score are calculated from the above confusion matrix to evaluate this work. Table 35 shows the results of numerous performance evaluation metrics. In the overall examination of the results, Table 5 shows that the Part classifier outperforms the other four classifiers. The Part classifier has the highest accuracy of all the classifiers for all classes 88.2, 88.1, 88.5, and 88.2. We examined numerous quality assessment criteria to evaluate the effective classification algorithm. We discovered that part classifier beats all other data mining algorithms. After apply SVM, Random Forest, and Logistic Regression algorithm we found best classifier algorithm as Random Forest.

CHAPTER 5

SUMMARY AND CONCLUSION

5.1 Conclusion

This task primarily consists of predicting an individual's symptoms and identifying the ailment. This is accomplished using a variety of data mining approaches. Essentially, the goal of this study is to use machine learning and data mining to produce a good output for human diseases. A total of 70% and 30% of data is needed to train and test the classifier, respectively, to complete this task. We examined numerous performance evaluation criteria to evaluate the working classifier. We discovered that part classifier beats all other data mining algorithms. And the best classifier we found that Random Forest. We can also get a good output from Other three classifier. We got least score from K-nearest Neighbor (KNN) classifier and that was Accuracy (68.7%), F1 (66.1%), Precision (66.4), and Recall (67.9) that means above 66%. So, we think that our work gives a better output for human body condition checking.

5.2 Future Work

In this work we work with 500 raw data, so that firstly we collect very large number of datasets. And apply different types of classifiers on datasets. In our work we apply four classifiers, and we work with eleven diseases. So that, in the future, we will work with more diseases with more attributes and we apply more classifier and data mining techniques. After all we can say that in future, we will continue our work to make this work as best work.

References

- [1] Chang, Cheng-Ding, Chien-Chih Wang, and Bernard C. Jiang. "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors." *Expert systems with applications* 38, no. 5 (2011): 5507-5513.
- [2] Bang, Sohyun, DongAhn Yoo, Soo-Jin Kim, Soyun Jhang, Seoae Cho, and Heebal Kim. "Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data." *Scientific reports* 9, no. 1 (2019): 1-9.
- [3] Kunjir, Ajinkya, Harshal Sawant, and Nuzhat F. Shaikh. "Data mining and visualization for prediction of multiple diseases in healthcare." In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, pp. 329-334. IEEE, 2017.
- [4] Muhammad, L. J., Md Milon Islam, Sani Sharif Usman, and Safial Islam Ayon. "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery." *SN Computer Science* 1, no. 4 (2020): 1-7.
- [5] Besag, Julian, and James Newell. "The detection of clusters in rare diseases." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 154, no. 1 (1991): 143-155.
- [6] Shamrat, FM Javed Mehedi, Md Abu Raihan, AKM Sazzadur Rahman, Imran Mahmud, and Rozina Akter. "An analysis on breast disease prediction using machine learning approaches." *International Journal of Scientific & Technology Research* 9, no. 02 (2020): 2450-2455.
- [7] Rajdhan, Apurb, Avi Agarwal, Milan Sai, Dundigalla Ravi, and Poonam Ghuli. "Heart disease prediction using machine learning." *International Journal of Research and Technology* 9, no. 04 (2020): 659-662.
- [8] Priya, M. Banu, P. Laura Juliet, and P. R. Tamilselvi. "Performance analysis of liver disease prediction using machine learning algorithms." *International Research Journal of Engineering and Technology (IRJET)* 5, no. 1 (2018): 206-211.
- [9] Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. "Comparing different supervised machine learning algorithms for disease prediction." *BMC medical informatics and decision making* 19, no. 1 (2019): 1-16.
- [10] Rahman, AKM Sazzadur, et al. "A comparative study on liver disease prediction using supervised machine learning algorithms." *International Journal of Scientific & Technology Research* 8.11 (2019): 419-422.
- [11] Swapna, G., R. Vinayakumar, and K. P. Soman. "Diabetes detection using deep learning algorithms." *ICT express* 4, no. 4 (2018): 243-246.
- [12] Candás, Juan Luis Carús, Víctor Peláez, Gloria López, Miguel Ángel Fernández, Eduardo Alvarez, and Gabriel Díaz. "An automatic data mining method to detect abnormal human behaviour using physical activity measurements." *Pervasive and Mobile Computing* 15 (2014): 228-241.
- [13] Gabrielsen, Terisa P., Megan Farley, Leslie Speer, Michele Villalobos, Courtney N. Baker, and Judith Miller. "Identifying autism in a brief observation." *Pediatrics* 135, no. 2 (2015): e330-e338..

- [14] Chaurasia, Vikas, and Saurabh Pal. "Data mining approach to detect heart diseases." *International Journal of Advanced Computer Science and Information Technology (IJACSIT)* Vol 2 (2014): 56-6.
- [15] Wang, M-Y., X-Y. Zhang, L. Xu, Y. Feng, Y-C. Xu, L. Qi, and Y-F. Zou. "Detection of bone marrow oedema in knee joints using a dual-energy CT virtual non-calcium technique." *Clinical radiology* 74, no. 10 (2019): 815-e1.
- [16] Ibrahim, Ibrahim, and Adnan Abdulazeez. "The role of machine learning algorithms for diagnosing diseases." *Journal of Applied Science and Technology Trends* 2, no. 01 (2021): 10-19.
- [17] Logistic Regression in Machine Learning: <https://www.javatpoint.com/logistic-regression-in-machine-learning>, last accessed 20/09/2021
- [18] Classification Algorithms - RandomForest: <https://www.tutorialspoint.com/machine-learning-with-python/machine-learning-with-python-classification-algorithms-random-forest.htm>. Last accessed 20/09/2021
- [19] Support Vector Machine (SVM) Algorithm. <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>. Last accessed 20/09/2021.
- [20] K-Nearest Neighbor(KNN) Algorithm for Machine Learning. <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learnin>, Last accessed 20/09/2021
- [21] Muthu, B., Sivaparthipan, C.B., Manogaran, G., Sundarasekar, R., Kadry, S., Shanthini, A. and Dasel, A., 2020. IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. *Peer-to-peer networking and applications*, 13(6), pp.2123-2134.

APPENDIX

Appendix A: Related Issues

To complete this work, we faced some complicated issues. Our main issue was collecting data from people. Because of, when we surveyed for data collection many people didn't give their proper data or didn't any data. So, that we have wasted lot of time for data collection. We have faced another issue, that was based on what we thought we would do our works of disease prediction. Then we decided that we take diseases symptoms of detecting disease for doing this work. And we also face some issues like as we need to numbering our collected data for applying algorithm, and also using label encoder for numbering for disease.

Common Human Disease Prediction Using Machine Learning Based on Survey Data

ORIGINALITY REPORT

20%	13%	13%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Proceedings of International Conference on Trends in Computational and Cognitive Engineering", Springer Science and Business Media LLC, 2021 Publication	3%
2	Submitted to Daffodil International University Student Paper	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	www.ijstr.org Internet Source	1%
5	Submitted to St. Mary's College Twickenham Student Paper	1%
6	Barnali Adhikari, Syeda Ismatara Era, Mohamed Dahir Mohamed, Jueal Mia. "Depression Level Prediction For Students Using Machine Learning In The Context of Bangladesh", 2021 12th International	1%