GLOBAL SALES DATA ANALYTICS

PREDICTION OF DISEASE AND THE APPROPRIATE HEALTHCARE USING RANDOM FOREST CLASSIFER

A PROJECT REPORT

TEAM ID: PNT2022TMID20211

SAMUEL AARON K

SRUTHI SRI S C

SUSHMITHAA S

TILAK D V

ABSTRACT

In recent years, artificial intelligence has woven itself into our daily lives in ways we may not even be aware of. Artificial Intelligence touches every aspect of our personal and professional online lives today. Machine learning, a part of AI can produce accurate results and analysis by developing efficient and fast algorithms and data-driven models for real-time processing of this data. The major area our project focuses on AI(ML). The main aim of our project is to help people find healthcare based on their problems. Health care is conventionally regarded as an important determinant in promoting the physical, mental, and social well-being of people around the world and can contribute to a significant part of a country's economy, development and industrialisation when efficient. Our project is basically an application program that can be used by everyone in their day-day life. The complete process of this project takes place in two phases. First phase of the project would be knowing the symptoms (getting inputs) from the user and predicting their disease/problem using Random Forest Classifier. On predicting the disease, this application automatically displays their respective healthcare which will be second phase our idea. The app not only displays the healthcare but also it displays further information about the healthcare which includes its area, hospitality, cost, expertise and more. This way of finding healthcare by knowing symptoms would be a great measure in the field of telemedicine.

TABLE OF CONTENTS

`

CHAPTER 1

INTRODUCTION

India has a multi-payer universal health care model that is paid for by a combination of public and private health insurance funds along with the element of almost entirely tax-funded public hospitals. The public hospital system is essentially free for all Indian residents except for small, often symbolic co-payments in some services. At the federal level, a national publicly funded health insurance program was launched in 2018 by the Government of India, called Ayushman Bharat. This aimed to cover the bottom 50% (500 million people) of the country's population working in the unorganized sector (enterprises having less than 10 employees) and offers them free treatment at both public and private hospitals. For people working in the organized sector (enterprises with more than 10 employees) and earning a monthly salary of up to ₹21,000 are covered by the social insurance scheme of Employees' State Insurance which entirely funds their healthcare (along with unemployment benefits), both in public and private hospitals. People earning more than that amount are provided health insurance coverage by their employers through either one of the four main public health insurance funds which are the National Insurance Company, The Oriental Insurance Company, United India Insurance Company and New India Assurance or a private insurance provider. All employers in India are legally mandated to provide health insurance coverage to their employees and dependents as part of Social Security in India.

Public healthcare is free for every Indian resident. The Indian public health sector encompasses 18% of total outpatient care and 44% of total inpatient care. Middle- and upper-class individuals living in India tend to use public healthcare less than those with a lower standard of living. Additionally, women and the elderly are more likely to use public

services. The public health care system was originally developed in order to provide a means to healthcare access regardless of socioeconomic status or caste. However, reliance on public and private healthcare sectors varies significantly between states. Several reasons are cited for relying on the private rather than public sector; the main reason at the national level is poor quality of care in the public sector, with more than 57% of households pointing to this as the reason for a preference for private health care. Much of the public healthcare sector caters to the rural areas, and the poor quality arises from the reluctance of experienced healthcare providers to visit the rural areas. Consequently, the majority of the public healthcare system catering to the rural and remote areas relies on inexperienced and unmotivated interns who are mandated to spend time in public healthcare clinics as part of their curricular requirement. Other major reasons are long distances between public hospitals and residential areas, long wait times, and inconvenient hours of operation.

1.1 PUBLIC AND PRIVATE HEALTHCARE

Different factors related to public healthcare are divided between the state and national government systems in terms of making decisions, as the national government addresses broadly applicable healthcare issues such as overall family welfare and prevention of major diseases, while the state governments handle aspects such as local hospitals, public health, promotion and sanitation, which differ from state to state based on the particular communities involved. Interaction between the state and national governments does occur for healthcare issues that require larger scale resources or present a concern to the country.

Considering the goal of obtaining universal health care as part of Sustainable Development Goals, scholars request policy makers to

acknowledge the form of healthcare that many are using. Scholars state that the government has a responsibility to provide health services that are affordable, adequate, new and acceptable for its citizens.[24] Public healthcare is very necessary, especially when considering the costs incurred with private services. Many citizens rely on subsidized healthcare. The national budget, scholars argue, must allocate money to the public healthcare system to ensure the poor are not left with the stress of meeting private sector payments.

Following the 2014 election which brought Prime Minister Narendra Modi to office, the government unveiled plans for a nationwide universal health care system known as the National Health Assurance Mission, which would provide all citizens with free drugs, diagnostic treatments, and insurance for serious ailments. In 2015, implementation of a universal health care system was delayed due to budgetary concerns. In April 2018 the government announced the Aayushman Bharat scheme that aims to cover up to Rs. 5 lakh to 100,000,000 vulnerable families (approximately 500,000,000 persons – 40% of the country's population). This will cost around $1.7 billion each year. Provision would be partly through private providers.

Since 2005, most of the healthcare capacity added has been in the private sector, or in partnership with the private sector. The private sector consists of 58% of the hospitals in the country, 29% of beds in hospitals, and 81% of doctors.

According to National Family Health Survey-3, the private medical sector remains the primary source of health care for 70% of households in urban areas and 63% of households in rural areas. The study conducted by IMS Institute for Healthcare Informatics in 2013, across 12 states in over 14,000 households indicated a steady increase in the usage of private healthcare facilities over the last 25 years for both Out-Patient and In-Patient services, across rural and urban areas. In terms of healthcare quality in the private sector, a 2012 study by Sanjay Basu et al., published in *PLOS Medicine*, indicated that

health care providers in the private sector were more likely to spend a longer duration with their patients and conduct physical exams as a part of the visit compared to those working in public healthcare. However, the high out of pocket cost from the private healthcare sector has led many households to incur Catastrophic Health Expenditure, which can be defined as health expenditure that threatens a household's capacity to maintain a basic standard of living. Costs of the private sector are only increasing. One study found that over 35% of poor Indian households incur such expenditure and this reflects the detrimental state in which Indian health care system is at the moment. With government expenditure on health as a percentage of GDP falling over the years and the rise of private health care sector, the poor are left with fewer options than before to access health care services. Private insurance is available in India, as are various through government-sponsored health insurance schemes. According to the World Bank, about 25% of India's population had some form of health insurance in 2010. A 2014 Indian government study found this to be an over-estimate, and claimed that only about 17% of India's population was insured. Private healthcare providers in India typically offer high quality treatment at unreasonable costs as there is no regulatory authority or statutory neutral body to check for medical malpractices. In Rajasthan, 40% of practitioners did not have a medical degree and 20% have not completed a secondary education. On 27 May 2012, the popular show Satyamev Jayate did an episode on "Does Healthcare Need Healing?" which highlighted the high costs and other malpractices adopted by private clinics and hospitals.

1.2 ACCESS TO HEALTHCARE:

As of 2013, the number of trained medical practitioners in the country was as high as 1.4 million, including 0.7 million graduate allopath. Yet, India has failed to reach its Millennium Development Goals related to health. The definition of 'access is the ability to receive services of a certain quality at a

specific cost and convenience. The healthcare system of India is lacking in three factors related to access to healthcare: provision, utilization, and attainment. Provision, or the supply of healthcare facilities, can lead to utilization, and finally attainment of good health. However, there currently exists a huge gap between these factors, leading to a collapsed system with insufficient access to healthcare. Differential distributions of services, power, and resources have resulted in inequalities in healthcare access. Access and entry into hospitals depends on gender, socioeconomic status, education, wealth, and location of residence (urban versus rural). Furthermore, inequalities in financing healthcare and distance from healthcare facilities are barriers to access. Additionally, there is a lack of sufficient infrastructure in areas with high concentrations of poor individuals. Large numbers of tribes and ex-untouchables that live in isolated and dispersed areas often have low numbers of professionals. Finally, health services may have long wait times or consider ailments as not serious enough to treat. Those with the greatest need often do not have access to healthcare.

The Government of India, while unveiling the National Health Portal, has come out with guidelines for electronic health record standards in India. The document recommends a set of standards to be followed by different healthcare service providers in India, so that medical data becomes portable and easily transferable. India is considering to set up a National eHealth Authority (NeHA) for standardization, storage, and exchange of electronic health records of patients as part of the government's Digital India programme. The authority, to be set up by an Act of Parliament will work on the integration of multiple health IT systems in a way that ensures security, confidentiality and privacy of patient data. A centralised electronic health record repository of all citizens which is the ultimate goal of the authority will ensure that the health history and status of all patients would always be available to all health institutions. Union Health Ministry has circulated a concept note for the setting

up of **NeHa**, inviting comments from stakeholders.

1.2.1. URBAN AREA

The problem of healthcare access arises not only in huge cities but in rapidly growing small urban areas. Here, there are fewer available options for healthcare services and there are less organized governmental bodies. Thus, there is often a lack of accountability and cooperation in healthcare departments in urban areas. It is difficult to pinpoint an establishment responsible for providing urban health services, compared to in rural areas where the responsibility lies with the district administration. Additionally, health inequalities arise in urban areas due to difficulties in residence, socioeconomic status, and discrimination against unlisted slums.

To survive in this environment, urban people use non-governmental, private services which are plentiful. However, these are often understaffed, require three times the payment as a public center, and commonly have bad practice methods. To counter this, there have been efforts to join the public and private sectors in urban areas. An example of this is the Public-Private Partnerships initiative. However, studies show that in contrast to rural areas, qualified physicians tend to reside in urban areas. This can be explained by both urbanization and specialization. Private doctors tend to be specialized in a specific field so they reside in urban areas where there is a higher market and financial ability for those services.

1.3. ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is intelligence - perceiving, synthesizing, and infering information - demonstrated by machines, as opposed to intelligence displayed by animals and humans. Example tasks in which this is done include speech recognition, computer vision, translation between

(natural) languages, as well as other mappings of inputs. OED (OUP) defines artificial intelligence as:

the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

Several important sub-fields of *AI research* (as opposed to AI itself) have *used* working definitions of the intelligent agents field of study, which refers to any system that perceives its environment *using an AI-component* and takes actions (using procedural / hard-coded components) that maximize its chance of achieving its goals. While intelligent agents as systems that use artificial intelligence are an important application of AI, many AI systems do not perform any procedural (hard-coded) steps with the outputs of the AI at all, such as computer vision, speech recognition, or recommender systems (often not even decising on an output from probabilities, but outputting several).

The term "artificial intelligence" had previously been used to describe machines that mimic and display "human" cognitive skills that are associated with the human mind, such as "learning" and "problem-solving". This definition has since been rejected by major AI researchers who now describe AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated.

As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect. For instance, optical character recognition is frequently excluded from things considered to be AI,[5] having become a routine technology.

Artificial intelligence was founded as an academic discipline in 1956, and in the years since has experienced several waves of optimism, followed by disappointment and the loss of funding (known as an "AI winter"),[9][10] followed by new approaches, success and renewed funding. AI

research has tried and discarded many different approaches since its founding, including simulating the brain, modeling human problem solving, formal logic, large databases of knowledge and imitating animal behavior. In the first decades of the 21st century, highly mathematical-statistical machine learning has dominated the field, and this technique has proved highly successful, helping to solve many challenging problems throughout industry and academia.

1..3.1. BRANCHES OF AI:

Here are the major branches of Artificial Intelligence: Experts Systems, Robotics, Machine Learning, Neural Network, Fuzzy Logic, Natural Language Processing.

Experts Systems:

Expert Systems is an Artificial Intelligence (AI-based) system that learns and imitates a human being's decision-making ability. Expert Systems does not use conventional programming to solve complex problems but instead uses logical notations to achieve such an aim.It is mainly used in the medical field to operate medical facilities and detect virus infections. It is also used in the banking sector for loan and investment analysis.

Robotics:

This is a very interesting branch of Artificial Intelligence that focuses on the design and development of robots. Robotics deals with the designing, constructing, and operating of robots by incorporating both science and engineering techniques. The aim of deploying robots is to help humans with tedious and bulky tasks. These tasks involve the control of computer systems, information transformation and manufacturing of automobiles. It is used by NASA to move heavy objects in space. Robots also act as artificial intelligence agents that perform tasks in a real-world environment with the aim of

actualizing results. This branch of AI is so amazing.

Machine Learning:

Machine Learning is a highly demanding branch of Artificial Intelligence. It is the science that enables machines and computer systems to process, analyze and interpret data with the aim of providing solutions for real-life challenges. Computer systems can learn and take actions on their own due to the level of sufficient data provided through Machine Learning. The algorithm is set up in such a way that machines can predict outcomes based on past occurrences. Machine Learning algorithms and techniques help in training a model with data presented which will then predict and adjust to future outcomes. It is the science of allowing computer systems to learn and translate data for the sake of task execution without programming. Technology discoveries such as web search, speech recognition and automatic vehicles are results of Machine Learning.

Here are three major categories under Machine Learning;

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Neural Network:

Neural Network is a branch of Artificial Intelligence associated with the use of Neurology to incorporate cognitive science in helping computer systems and machines to execute tasks. It is known as “DEEP LEARNING” because it involves making use of artificial brain neurons to solve complex problems. Neural Network helps machines process how the human brain operates. This branch of AI also involves implementing mathematical functions and statistical techniques to solve real-world problems. It is used in fields such as risk analysis, market research, fraud

detection, forecasting, and stock exchange prediction. Face verification algorithms on social media sites are a result of the implementation of Neural Network. This wonderful branch of AI is also responsible for virtual assistant apps such as "ALEXA and SIRI". You can check out **AI Certification Cost** online so as to know if you can afford to study Neural Network AI certification.

Fuzzy Logic:

This branch of AI is the technique of modifying and representing uncertain information by analysing the degree to which the hypothesis is true. Fuzzy Logic helps to offer a certain level of reasoning flexibility when faced with uncertainties. This might sound a bit complex but it is simply a case of using standard logic to determine if a concept exhibits a degree of truth. For instance, standard logic is 1.0 if a concept is TRUE and 0.0 if a concept is FALSE. However, there are cases where a concept can either be partially true or partially false. Just as humans face dilemmas in their day-to-day activities, a computer system can be made to experience such with the aim of finding a solution. Fuzzy Logic is used in automatic gearboxes and medicine for decision making.

Natural Language Processing:

Communicating with someone who doesn't understand your language can be very challenging and the same can be said of humans trying to communicate with a computer system. A computer will find it difficult to interpret words because it only understands the language of binary digits. This challenge has led to the development of Natural Language Processing in computer science.This is simply the process of making computer systems and machines to understand basic human interactions. This process involves a machine receiving human sound from interaction and converting it to text format so that it can be easily read and understood. These texts are then converted to components by the computer system that will make it understand the intention of the human.

1.3.2. MACHINE LEARNING

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, agriculture, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers, but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory, and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

Learning algorithms work on the basis that strategies, algorithms, and inferences that worked well in the past are likely to continue working well in the future. These inferences can be obvious, such as "since the sun rose every morning for the last 10,000 days, it will probably rise tomorrow morning as well". They can be nuanced, such as "X% of families have geographically separate species with colour variants, so there is a Y% chance that undiscovered black swans exist".

Machine learning programs can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced

tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers. For example, to train a system for the task of digital character recognition, the MNIST dataset of handwritten digits has often been used.

1.3.3 DECISION TREE

Decision tree learning uses a decision tree as a predictive model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining, and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels, and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to represent decisions and decision making visually and explicitly. In data mining, a decision tree describes data, but the resulting classification tree can be an input for decision-making.

1.3.4 RANDOM FOREST CLASSIFIER

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression

problems in ML. It is based on the concept of **ensemble learning,** which is a process of *COMBINING MULTIPLE CLASSIFIERS TO SOLVE A COMPLEX PROBLEM AND TO IMPROVE THE PERFORMANCE OF THE MODEL.*

As the name suggests, **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

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

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.

It can also maintain accuracy when a large proportion of data is missing.

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points.

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

CHAPTER 2

LITERATURE REVIEW

PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS ON DIABETES DATASET USING BIG DATA ANALYTICS

Recent years healthcare data is growing very vastly and it has different dimensions such as structured, semi-structured and unstructured. Now-a-days technology plays a vital role in providing services in the area of healthcare. Utilization of technologies such as cloud, big data, sensor data etc., as a whole internet is widespread in healthcare domain to resolve problems arising in the field of e-healthcare. There are various chronic diseases such as childhood Pneumonia, Diabetes, Thyroid etc., affected by the people across the world. Diabetes Mellitus is one of the metabolic diseases and where the blood glucose levels raise over a prolonged period. Diabetes is life threatening and long-standing disease affecting other body parts. Normally Glucose is the break down product of any carbohydrate that entered the body and this special sugar fuels the cells. In diabetes either the body fails to produce adequate insulin, a hormone or unable to utilize insulin or both. Diabetes mellitus can be categorized into three types [2]: Type 1 diabetes is also known as "Insulin-Dependent Diabetes Mellitus" (IDDM) or "juvenile diabetes" is usually occurs in children. In this type of diabetes damaged pancreas does not produce insulin. It is an auto immune condition may also be caused by a genetic predisposition. Various organs affected by Type 1 Diabetes are eyes, tiny blood vessels, kidneys, heart, and nerves. Second type of diabetes is Type 2, which occurs frequently in adults for about 95% of diabetic cases. In this condition pancreatic cells fail to produce enough insulin, later as the disease progresses lack of insulin develops

and the cells become insulin resistance. It is also called as “Noninsulin-Dependent Diabetes Mellitus” (NIDDM). It is a mild form of diabetes but it may produce high risk of health complications affecting the small blood vessels which serve the organs such as kidneys, eyes, nerves and heart. Persons may have high risk with type 2 diabetes due to overweight and less or no exercise. Type 2 diabetes is not curable however can be controlled with regular exercise, normal maintenance of weight, healthy diet and avoiding tobacco usage.

In this paper, proposed approach considers diabetes dataset in its experimentation because it is a life-threatening chronic disease. In this proposed approach different machine learning algorithms of different representations and mode of learning styles are considered in predicting diabetes type and based on the predictions we analysed accuracy of individual algorithm. As a result, we identified most accurate algorithm which predicts data accurately with compared to other algorithms.

The proposed approach in this paper has three steps in its methodology. In step-1, load the diabetes dataset into RStudio for the purpose of pre-processing. Further Data pre-processing is done on loaded dataset with cross validation method with 10 folds and this process is repeated 3 times. This is a common configuration or standard method for comparing different models. Next to that, the pre-processed data is randomly divided into two sets namely training set and test set with the ratio of 80: 20 respectively which is commonly used ratio in literature. Apply different machine learning algorithms such as RF, LDA, CART and k-NN to learn the data patterns and train the data to get predictions. Then learn about the model to test the predictions with test dataset. After this step, analysis is performed based on accuracy and kappa metrics.

PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS ON DIABETES DATASET USING BIG DATA ANALYTICS

New Technologies such as Big Data and Cloud is playing a vital role in providing solutions to Healthcare problems. Now-a-days healthcare data is growing very drastically day-by-day and it requires an efficient, effective and timely solution to reduce the mortality rate. One of the most critical chronic healthcare problems is diabetes. In Long run, this problem may leads to damage eyes, heart, kidneys and nerves of diabetes patient if improper medication is done which also leads to death. The aim of this paper is to analyze and compare different machine learning algorithms to identify a best predicting algorithm based on various metrics such as accuracy, kappa, precision, recall, sensitivity and specificity. A comprehensive study is done on diabetes dataset with Random Forest (RF), SVM, k-NN, CART and LDA algorithms. The achieved results shows that RF is giving more accurate predictions with compared to other algorithms.

MACHINE LEARNING ALGORITHM IN HEALTHCARE SYSTEM: A REVIEW

In the last decades Machine learning techniques are widely used in the field of healthcare systems due to its data processing and analysis capabilities. Machine Learning is a sub domain of artificial intelligence that collects data from various sources and in various format. Despite its major capability to handle the huge data still classification of data is still the major difficulty in the field of healthcare. Now a day, many people are facing such kind of vital diseases which need to be identified at the early phase of diseases so that treatment can be start in relevant time. After passing such stage the diseases may be uncurable. This can be possible with the help of various Machine learning technique. Many Machine learning technique are much more capable to analyse the huge complex medical data, medical reports and medical images in a very less time with accuracy. There are various cases available where many fatal diseases may not be identified by experts. Just like many other fields, in healthcare Machine learning algorithms are widely used to tackle such kind of situations. This research article focused on the various field of machine learning that

are being used for handling complex data for the purpose of decision making in healthcare system. This paper attempt to provide the brief details about various machine learning approach and review the role of these algorithms in field of healthcare system like diabetic, detection of cancer, brain tumour, bioinformatics and many more.

CHAPTER 3
RESEARCH METHODOLOGY

3.1. EXISTING MEDHODOLOGY

The proposed approach considers diabetes dataset in its experimentation because it's a life-threatening chronic disease. In this proposed approach different machine learning algorithms of different representations and mode of learning styles are considered in predicting diabetes type and based on the predictions we analyzed accuracy of individual algorithm. As a result, we identified most accurate algorithm which predicts data accurately with compared to other algorithms.

New Technologies such as Big Data and Cloud is playing a vital role in providing solutions to healthcare problems. Now-a-days healthcare data is growing very drastically day-by-day and it requires an efficient, effective and timely solution to reduce the mortality rate. One of the most critical chronic healthcare problems is diabetes. In Long run, this problem may lead to damage eyes, heart, kidneys and nerves of diabetes patient if improper medication is done which also leads to death. The aim of this paper is to analyze and compare different machine learning algorithms to identify a best predicting algorithm based on various metrics such as accuracy, kappa, precision, recall, sensitivity and specificity. A comprehensive study is done on diabetes dataset with Random Forest (RF), SVM, k-NN, CART and LDA algorithms. The achieved results shows that RF is giving more accurate predictions with compared to other algorithms.

The proposed approach in this paper has three steps in its methodology. In step-1, load the diabetes dataset into RStudio for the purpose of pre-processing. Further Data pre-processing is done on loaded dataset with cross validation method

with 10 folds and this process is repeated 3 times. This is a common configuration or standard method for comparing different models. Next to that, the preprocessed data is randomly divided into two sets namely training set and test set with the ratio of 80: 20 respectively which is commonly used ratio in literature. Apply different machine learning algorithms such as RF, LDA, CART and k-NN to learn the data patterns and train the data to get predictions. Then learn about the model to test the predictions with test dataset. After this step, analysis is performed based on accuracy and kappa metrics.

3.2. PROPOSED METHODOLOGY

The main aim of our project is to predict the disease and to provide the right choice of healthcare (list of hospitals) based on the symptoms provided by the user.

The project is divided into two phases. First phase is predicting the disease type based upon the symptoms given. Second phase is predicting the future patients count of a particular disease provided that the particular disease's historic patients count is given. Both are implemented using Machine Learning algorithms. For predicting the type of disease, Random Forest classifier has been used for classification and for predicting the patient's count, Holt's method (Time series analysis) has been used (Regression method).

A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Holt-Winters is a model of time series behavior. Forecasting always requires a model, and Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of

one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

First phase:

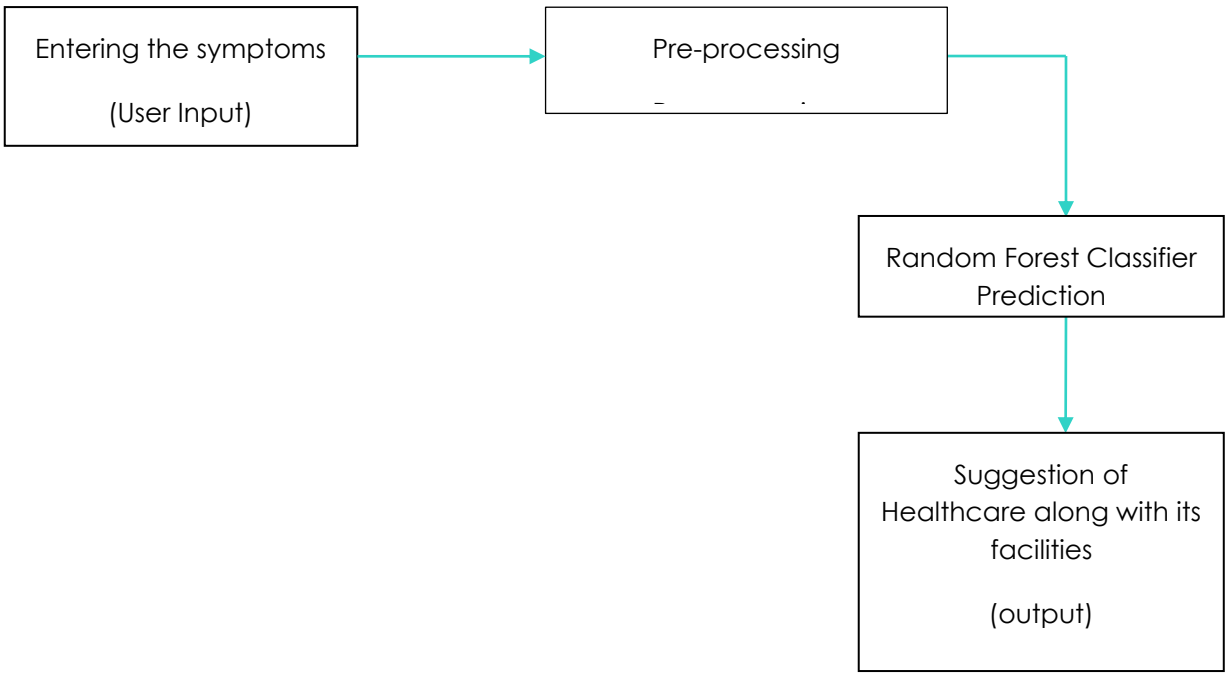
To bring awareness about their medical conditions to the people (Given that the predictions will not be very perfect) and suggesting them to the appropriate health care.

Second phase:

By knowing the tentative/approximate count, the health care sector/hospitals can plan for more/less medication and medical process. For example, if common cold is more prevalent in third and fourth quarter of a year (Winter season). More tablets related to it can be manufactured/imported. Likewise, this model predicts for all the disease based upon the data of the disease we are providing.

Data has been taken from Kaggle source for second phase and mock data has been prepared from research for first phase. *KAGGLE* is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.

3.3. BLOCK DIAGRAM:



CHAPTER 4

CONCLUSION

Today's world people are more involved in their hectic schedules by not taking care of their heath, which leads to chronic problems such as diabetics. In this paper, author tries to give a comprehensive comparative study on different machine learning algorithms. This comparative study is done based on different metrics such as Accuracy, Kappa, Precision, Recall, Sensitivity and Specificity. The achieved results show that RF algorithm is predicting the data more correctly and accurately.

There are various cases available where many fatal diseases may not be identified by experts. Just like many other fields, in healthcare Machine learning algorithms are widely used to tackle such kind of situations. This research article focused on the various field of machine learning that are being used for handling complex data for the purpose of decision making in healthcare system. This paper attempt to provide the brief details about various machine learning approach and review the role of these algorithms in field of healthcare system like diabetic, detection of cancer, brain tumor, bioinformatics and many more.

CHAPTER 5

REFERENCE

1. Jaehun Lee et al, “Emerging Technology and Business Model Innovation: The Case of Artificial Intelligence, MDPI, vol.05, no.3,July 2020.
2. Andrea L Guzman et al, ”Artificial Intelligence and Communication :A Human Machine Communication", Sage Journals, vol.22,no. 01, July 2019.
3. Yi Zhang et al, ”Ethics and privacy of artificial intelligence: Understandings from bibliometrics", Knowledge based systems, vol. 222,no. 106994, June 2021.
4.U. Varshney, Pervasive Healthcare Computing: EMR/EHR, Wireless and Health Monitoring, 2009.
5."Diabetes Fact sheet N°312". WHO. October 2013. Retrieved 25 March 2014.
6. Shoback, edited by David G. Gardner, Dolores, "Chapter 17". Greenspan's basic & clinical endocrinology (9th ed.). New York: McGraw-Hill Medical., 2011, ISBN 0-07-162243-8.
7. RSSDI textbook of diabetes mellitus. (Rev. 2nd ed.). New Delhi: Jaypee Brothers Medical Publishers. 2012. p. 235. ISBN 9789350254899.
8.Cash, Jill. Family Practice Guidelines (3rd ed.). Springer, 2014, p. 396.ISBN 9780826168757.
9.IDF. International Diabetes Federation. Retrieved 29 November2014.